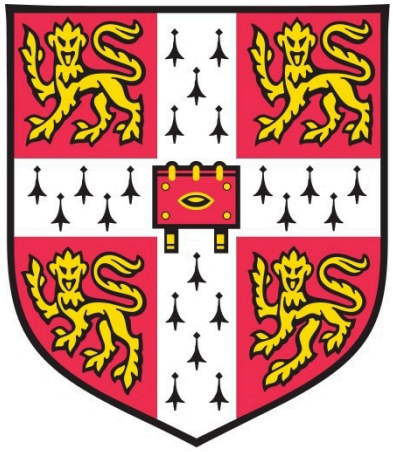
Using transcriptomic data to detect, understand, and treat injury in the context of drug toxicity and fibrotic disease

Anika Liu

St Catharine's College

September 2022



This thesis is submitted for the degree of Doctor of Philosophy.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the Physics and Chemistry Degree Committee.

Using transcriptomic data to detect, understand, and treat injury in the context of drug toxicity and fibrotic disease

Anika Liu

In drug discovery, it is crucial to understand how drugs relate to complex phenotypes. This includes understanding how a drug can help to treat a condition, but also how it can result in adverse effects so that safety risks can be mitigated earlier. How effects propagate from the molecular to the systems scale is, however, in many cases not fully clear, in particular for complex phenotypes which cannot be narrowed down to individual causes. In this thesis, transcriptomics data was used as intermediate layer alongside additional data sources to study links between compounds and phenotypes.

First, safety biomarker candidates in Drug-Induced Vascular Injury were identified based on changes in tissue expression across adverse and non-adverse treatments. Further characterization of their biological role and predictive performance thereby identified multiple secreted proteins as most promising candidates. Secondly, pathways and transcription factors involved in the pathogenesis of Drug-Induced Liver Injury, another safety-related endpoint, were identified and characterised based on time concordance in repeat-dose studies in rats. In Chapter 3, it is demonstrated how time concordance can be combined with other streams of evidence towards causal hypothesis and mechanistic biomarkers. In order to make time concordance analysis on the Open TG-GATEs liver data also accessible to researchers in an interactive manner, the R/Shiny app “DILI Cascades” is presented in Chapter 4. Instead of drug toxicity, the last chapter then focusses on efficacy and aims to prioritise repurposing candidates, direct targets and downstream effectors which may promote alveolar regeneration in Idiopathic Pulmonary Fibrosis. This demonstrates how single-cell RNA-Seq data can be leveraged for drug repurposing through better characterization of cell transitions followed by signature matching.

In summary, data-driven approaches with transcriptomics as key modality were used to derive insights on how drug perturbations are linked to adverse effects and fibrotic disease. Thereby, the presented work did not only aim at better mechanistic understanding but also provides actionable starting points for the discovery of new biomarkers and drug indications.

Acknowledgements

Firstly, I would like to thank my PhD supervisor, Dr Andreas Bender, for giving me the opportunity to pursue a PhD in his group, for providing me with opportunities to present my work and network within the field(s), and for his perspective and guidance throughout. I would also like to thank my secondary academic supervisor, Dr Namshik Han, for welcoming me into his group and the Milner Therapeutics Institute, and for enabling me to broaden my horizon by contributing to internal projects as well as external collaborations.

GlaxoSmithKline is thanked for funding my research and special thanks go to my industrial supervisor Dr Jordi Munoz-Muriedas for his support. Furthermore, I'd also like to acknowledge Drs. Deidre Dalmas, Valeriu Damian, Jim Harvey and Randall Smith at GSK for their scientific and organisational efforts with the DIVI biomarker project in Chapter 2.

Furthermore, I would like to thank Dr Joo-Hyeon Lee for her support and helpful discussions on alveolar regeneration which is at the core of the drug repurposing work presented in Chapter 5.

I would also like to thank the Bender and Han groups for the supportive work environment and great memories, including pub nights, coffee breaks, conference trips, and other fun activities. Thanks to Andrew Boardman, Arushi Gandhi, Benoit Baillif, Dr Danilo Basili, Dr Dezső Módos, Dr Hongbin Yang, Dr Ines Smit, Layla Hosseini-Gerami, Dr Maria-Anna Trapotsi, Miguel García Ortegón, Morgan Thomas, Peter Wright, and Srijit Seal, as well as Dr Gehad Youssef, Dr Georgia Tsagkogeorga, Dr Méabh MacMahon, Nicholas Katritsis, Dr Sanjay Rathee, Winnie Lei, Dr Woochang Hwang– and many others!

This PhD would not have been possible without the love, support and encouragement from my family. I'd like to thank the Behr family for welcoming me into their lives and their home. Furthermore, I owe everlasting thanks to my parents, Weixia Yan and Yu Liu, for everything they have done for me and for their continuous support. And last but not least, to my partner Jonathan Behr, thank you for being there for me and brightening my day always!

List of publications

Liu, A.; Munoz-Muriedas, J.; Bender, A.; Dalmas, D. A. Identification of potential biomarker candidates of drug-induced vascular injury (DIVI) in rats using gene expression and histopathology data. *bioRxiv* **2022**, 2022.08.24.505120. https://doi.org/10.1101/2022.08.24.505120.

Sections from this pre-print are included in Chapters 1 and 2.

Liu, A.; Han, N.; Munoz-Muriedas, J.; Bender, A. Deriving Time-Concordant Event Cascades from Gene Expression Data: A Case Study for Drug-Induced Liver Injury (DILI). *PLOS Comput. Biol.* **2022**, *18* (6), e1010148. https://doi.org/10.1371/journal.pcbi.1010148.

Sections from this publication are included in Chapters 1, 3 and 4.

Liu, A.; Lee, J.-H.; Han, N.; Bender, A. ScRNA-Seq-Based Drug Repurposing Targeting Idiopathic Pulmonary Fibrosis (IPF). *bioRxiv* **2022**, 2022.09.17.508360. https://doi.org/10.1101/2022.09.17.508360.

Sections from this pre-print are included in Chapters 1 and 5.

Liu, A.; Walter, M.; Wright, P.; Bartosik, A.; Dolciami, D.; Elbasir, A.; Yang, H.; Bender, A. Prediction and Mechanistic Analysis of Drug-Induced Liver Injury (DILI) Based on Chemical Structure. *Biol. Direct* **2021**, *16* (1), 1–15. https://doi.org/10.1186/s13062-020-00285-0.

This publication is reviewed in Chapter 1.

Han, N.; Hwang, W.; Tzelepis, K.; Schmerer, P.; Yankova, E.; MacMahon, M.; Lei, W.; Katritsis, N. M.; **Liu, A.**; Felgenhauer, U.; Schuldt, A.; Harris, R.; Chapman, K.; McCaughan, F.; Weber, F.; Kouzarides, T. Identification of SARS-CoV-2–Induced Pathways Reveals Drug Repurposing Strategies. *Sci. Adv.* **2021**, *7* (27), 2020.08.24.265496. https://doi.org/10.1126/sciadv.abh3032.

Rathee, S.; MacMahon, M.; **Liu, A.**; Katritsis, N. M.; Youssef, G.; Hwang, W.; Wollman, L.; Han, N. DILIC: An AI-Based Classifier to Search for Drug-Induced Liver Injury Literature. *Front. Genet.* **2022**, *0*, 1026. https://doi.org/10.3389/FGENE.2022.867946.

Katritsis, N. M.; **Liu, A.**; Youssef, G.; Rathee, S.; MacMahon, M.; Hwang, W.; Wollman, L.; Han, N. Dialogi: Utilising NLP with Chemical and Disease Similarities to Drive the Identification of Drug-Induced Liver Injury Literature. *Front. Genet.* **2022**, *0*, 1757. https://doi.org/10.3389/FGENE.2022.894209.

Table of Contents

1 Introduction

1.1 Pathogenesis in the context of drug discovery

The overall aim of drug discovery is to develop treatments which can improve patient health, which requires them to be efficacious and safe. However, the R&D productivity in the pharmaceutical industry has declined over recent years [1], and the absence of efficacy and safety were found to be the key reason for 52% and 24% of all clinical failures [2]. This highlights that an even better understanding of the pathological processes in diseases and adverse effects is needed so that more promising drug candidates can be identified in the pre-clinical stages.

1.1.1 Targeting disease mechanisms

From a disease biology perspective, drugs can be grouped into six nonexclusive categories based on whether and how they modulate disease [3] (Table 1.1). This includes four therapeutic categories targeting specific processes in disease pathogenesis and two which don't and consequentially can't impact the development of disease. Thereby, drugs in the first two therapeutic categories can prevent or cure a disease as they target the cause of the disease. For example, pathogens present the underlying cause of infectious diseases, and these can be cured using anti-infective medicines [4,5] (Category 0), which kill specific pathogens, or can be prevented through vaccines [6] (Category I), which train endogenous response to identify and eliminate pathogens before exposure to it. The next two categories do not target the cause but a specific disease process or manifestation, respectively, which can reduce disease progression by therapeutically restoring homeostasis. For instance, antifibrotic medicines target fibrotic signalling [7,8] (Category II) while hypertension is treated through the modulation of the Renin-Angiotensin-Aldosterone System [9] (Category III). Hence, a drug does not only need to target a disease-specific process to be therapeutic, the type of disease mechanism targeted is also directly related to the therapeutic potential. Therefore, understanding the underlying disease biology can provide directly relevant starting points for drug discovery.

Table 1.1: A classification of drug action based on therapeutic effects [3].

Bjornsson described six categories of drugs, out of which four are based on the type of disease process they act on (I-IV), namely the disease aetiology, specific disease processes, specific or non-specific disease manifestations. Two additional categories, disease prevention and non-disease-directed drug use, complete the classification system.

	Category	Targeted disease mechanism	Therapeutic effect on disease	Examples
0	Disease Prevention	Cause (before its presence)	Prevent occurrence or development	Vaccines [6]
I	Disease Aetiology	Cause	Reversal or prevention of progression	Antiinfectives [4,5]
II	Specific disease processes	Fundamental process	Reduce progression	Antifibrotics [7,8]
III	Specific disease manifestations	Specific symptom	Symptomatic effect, limited effect on progression	Antihypertensives [9]
IV	Non-specific disease manifestations	Unspecific symptom	None, only symptomatic effect	Analgesics [10]
V	Non-therapeutic drug use	None	None, only pharmacological effect	Anaesthetics [11]

To leverage interesting disease mechanisms therapeutically, these then need to be translated to a measurable assay which allows the rationalized identification of promising leads. An example for this are statins, where it was already known beforehand that increased serum cholesterol levels were a known risk factor for coronary events, and subsequent studies on cholesterol metabolism identified HMG-CoA reductase as a target in cholesterol synthesis [12]. Then, screening identified potent inhibitors of this enzyme, the first statins, which have, since their introduction to the market, contributed to a decrease in major coronary events as well as a generally reduced mortality in coronary patients [13].

The two nonexclusive main directions in this regard are target-based and phenotypic drug discovery. Target-based drug discovery relies on finding compounds targeting specific biomolecules, such as proteins or mRNA, which are known to be important for a given disease [14]. Once a target is identified, compounds with the desired interaction can be screened using affordable and fast *in vitro* assays, which can be supported by computational approaches. However, a disadvantage of target-based approaches is that these may bias R&D early efforts [15] but then do not guarantee translation to the clinic later on as interaction with a single target molecule may not be sufficient to treat complex diseases. In contrast, phenotypic drug discovery is instead focused on the empirical identification of compounds with the ability to induce a disease-relevant phenotype. It hence does not bias the discovery process towards a certain target or mechanism of action (MoA),

and, in fact, does not require any prior knowledge about the compound (although this information is of course useful). Whether or not the identified compounds are promising thereby strongly depends on how relevant the measured phenotype is for the given disease, which itself depends on the readout as well as on how closely the model system resembles the *in vivo* situation.

1.1.2 Anticipating adverse effects

Besides the desired therapeutic effect, drugs can have unwanted adverse effects ranging from a mild headache to severe implications such as liver failure. Adverse drug reactions are a major reason for compound failure in clinical trials [2,16] and a significant cause for post-marketing withdrawals. To avoid exposing patients to these risks and to allocate R&D resources to the most promising compounds, it is desired to identify adverse events earlier, more cost-efficiently, and better in the drug development process. Mechanistic understanding of adverse event pathogenesis is crucial in this regard to identify predictive events in the pathogenesis, which can be used to anticipate adverse events using predictive *in vitro* assays.

From a general molecular perspective, toxicity can be caused by either the drug itself or one of its metabolites (Table 1.2). Both can dysregulate specific cellular functions, either through the pharmacological target (on-target effects) or secondary off-target effects. Additionally, reactive metabolites can lead to the formation of adducts with intrinsic biomolecules, such as proteins or DNA, inducing cellular stress and immune response. Overall, this can then lead to acute or chronic drug-induced injury on the tissue and organ level.

Table 1.2: Mechanistic classes of drug toxicity, based on Liebler *et al.* [17]

Category	Cause of toxicity	Example
On-target	Interaction with the desired pharmacological target	Statins
Off-target	Interaction with an alternative target	Terfenadine
Biological activation	Reactive metabolites	Acetaminophen
Hypersensitivity	Activation of immune response	Penicillin
Idiosyncratic	Combination of individual and chemical factors	Isoniazid[18]

To detect toxicity early in drug development, *in vitro* screening assays are used to evaluate secondary pharmacology, which evaluates binding to off-targets, and can additionally provide insights into the general promiscuity of compounds [19]. Commonly tested targets

include G protein-coupled receptors (GPCRs), ion channels such as hERG, enzymes, transporters and nuclear receptors [20]. In this context, it should be noted that knowledge on relevant targets is largely incomplete and is often only identified after clinical adverse drug reactions are observed. While secondary pharmacology panels are most established in a preclinical setting, also additional efforts to detect toxicity should be mentioned, which include cell-based assays ranging from high-throughput screening using hypothesis-free assays such as imaging or transcriptomics to more physiologically relevant models such as Organ-on-a-chip models [21].

Overall, however, it is still largely unclear how to best conduct early toxicity profiling in line with the 3R principles [22], also because it is not fully understood how the initial interactions between the compound and biological system lead to different kinds of adverse effects on the systems level. To formalize mechanistic information in this regard, the Adverse Outcome Pathway (AOP) framework [23] was created building on previous efforts such as the mode of action framework [24,25]. AOPs describe multiscale event cascades from Molecular Initiating Events (MIE) through Key Events (KE) on different biological levels to the Adverse Event (AE). Once it is mechanistically understood which early events lead to AEs, this can then guide the development of suitable *in vitro* assays to anticipate toxicity for new compounds (Figure 1.1).

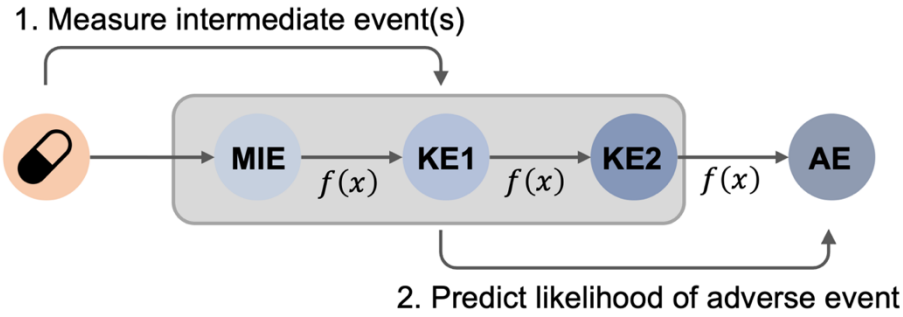


Figure 1.1: Adverse Outcome Pathways (AOP).

AOPs describe the event cascades from the first interaction of a compound with the biological system, termed Molecular Initiating Event or MIE, to Key Events (KEs) on different biological levels to the Adverse Event (AE). Practically, AOPs can help to anticipate AEs by identifying useful intermediate events, which can be measured or estimated using suitable assays, and then in turn are informative for the likelihood of the subsequent events including the AE itself.

While there are already established AOPs, e.g. available from the AOP knowledgebase (`https://aopkb.oecd.org`) or the AOP Wiki (`https://aopwiki.org`), the current

understanding of toxicity is yet largely incomplete. This is particularly true for complex phenotypes such as organ injury which can usually be caused by a wide range of compounds perturbing the biological system at different points mediated through multiple biological scales and entities [26,27]. Hence, a better mechanistic understanding of adverse events is needed in order to facilitate better and earlier detection of toxicity.

1.2 The pathogenesis of injury and fibrosis

In this work, pathogenic processes are studied in the context of drug-induced injury and fibrotic disease. Irrespective of the fact that these conditions are induced by distinct and partially idiosyncratic causes, there are some common biological mechanisms in the cellular and tissue-level response, such as ways in which injury can be detected, managed and countered. These mechanisms will be the focus of the following sections.

1.2.1 Cellular and tissue-level stress response

On the cellular level, there are different sensing mechanisms, which can identify different types of stress indicating cellular damage or malfunction [28]. These lead to the activation of stress signalling pathways which induce cellular programs to eliminate the stressor or to adapt to the new conditions. Representative pathways, including key sensors and transcription factors (TFs), are summarised in Table 1.3. Here, it should be noted that these do not act independently but that there is significant cross-talk resulting in an integrated stress response [29,30].

Besides intracellular signalling, paracrine and endocrine signalling contribute to the stress response [31], e.g. through induced chemokines or cytokines, which can warn neighbouring cells which are potentially exposed to the same stress and can initiate a tissue-level stress response (Figure 1.2). In the case of hypoxia, for instance, the metabolism is shifted to anaerobic glycolysis on a cell-intrinsic level, but also angiogenesis is induced via the vascular endothelial growth factor (VEGF) to increase the supply of oxygen on the tissue level [32]. If a cell is not able to adapt or eliminate the stress, either because of prolonged or severe exposure, senescence [33] or regulated cell death [34] can be initiated to prevent harmful effects on the surrounding tissue, or accidental cell death can take place [35].

Table 1.3: Representative cellular stress pathways including key sensors and transcription factors (TFs).

Adapted from Chovatiya *et al.* [30] and Simmons *et al.* [36]

Pathway	Sensor	TF	Cellular response
Oxidative stress [37]	Keap1	Nrf2	Scavenging of Reactive Oxygen Species (ROS)
Heat shock response [38]	Hsp90	HSF-1	Degradation of unfolded proteins, stabilization of misfolded proteins
DNA damage [39]	MDM2	p53	DNA repair, cell cycle arrest
Hypoxia [40]	VHL	HIF-1	Angiogenesis, shift to anaerobic glycolysis
ER stress [41]	BiP	XBP-1, ATF6, ATF4	Degradation of unfolded proteins, stabilization of misfolded proteins, suppression of nonessential translation, ER expansion
Osmotic stress [42]	None	NFAT5	Induction of aquaporins and electrolyte transporters

On the tissue level, homeostasis is maintained through the joint action of dynamically interacting parenchymal and accessory cells [43]. There are specialized sensory cells, namely tissue-resident macrophages, mast cells, and sensory neurons, which are involved in the defence response through direct recognition of stressor features [44], e.g. pathogen-associated molecular patterns (PAMPs), as well as in tissue-level stress response (para-inflammation) which can be activated by recognition of damage-associated molecular patterns (DAMPs) indicating cellular injury (Figure 1.2). These attempt to eliminate the cause of stress and to adapt to the new conditions, and if this is not feasible, they can induce inflammation via cytokines which recruit additional immune cells from the blood circulation, in particular phagocytes which can help to remove dead cells and tissue debris.

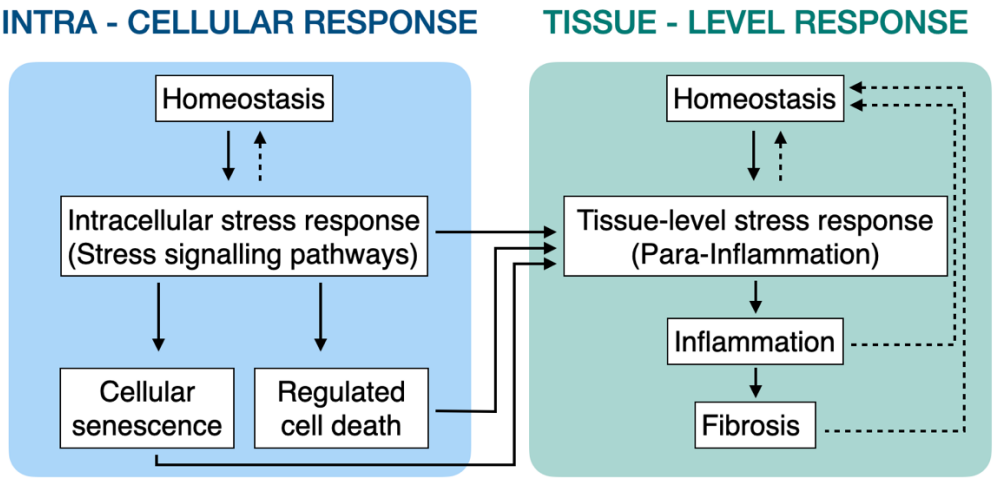


Figure 1.2: Mechanisms of the cellular and tissue-level stress response.

All steps involved in the biological response to stress aim at stress elimination and adaptation. If this is successful, homeostasis can be restored (dotted arrows). However, if it is not the response is escalated to the next level (continuous arrows) and can eventually result in fibrosis.

Inflammatory monocytes and tissue-resident macrophages are additionally important regulators of tissue repair and injury [45]. In tissue regeneration, damaged cells are replaced with new ones through differentiation and proliferation of cells from the same or also another cell type [46]. Additionally, myofibroblasts are formed from multiple cell populations including resident fibroblasts [47]. These are involved in wound repair, e.g. by producing and maintaining the extracellular matrix (ECM), as well as producing various chemokines and cytokines [48].

In the case of chronic inflammation, however, this can lead to fibrosis, which entails excessive ECM production and leads to scar formation [49]. By disrupting the overall tissue architecture, fibrosis can hence lead to functional decline in various organs [47] and is also a risk factor for cancer [50–52]. As it was long thought to be irreversible, current therapeutics primarily focus on slowing down disease progression. However, recent research efforts demonstrate that a reversal of fibrosis can be feasible [53,54], which will be further discussed below in the context of IPF (1.3.3).

1.2.2 Morphological changes and histological analysis

Histopathology refers to the examination of tissue biopsies using optical microscopy [55] and is the gold standard for the definition of toxicological effects [56] as well as for the diagnosis of many diseases [57]. Thereby, different stains can be used to highlight particular cell or tissue features. The most commonly used staining in histopathology is thereby H&E, which is a combination of haematoxylin, which stains the nuclei dark-purple, and eosin, which stains proteins in the extracellular matrix and the cytosol [58]. Jointly, this provides a broad overview of morphological cellular features, such as changes in cell size, number, and appearance, as well as cell composition and the tissue-level context, e.g. where certain changes are located, as e.g. demonstrated in Figure 1.3.

Frequent cellular morphologies in injury include hyperplasia (an increase in cell number), hypertrophy (an increase in cell size), and atrophy (a decrease in cell size). Furthermore, also apoptotic and necrotic cells can be identified, e.g. through chromatin condensation and cellular swelling, respectively [59]. Overall, the advantage of histology is hence that not only individual cells can be inspected, but that the spatial tissue context can be analysed more broadly.

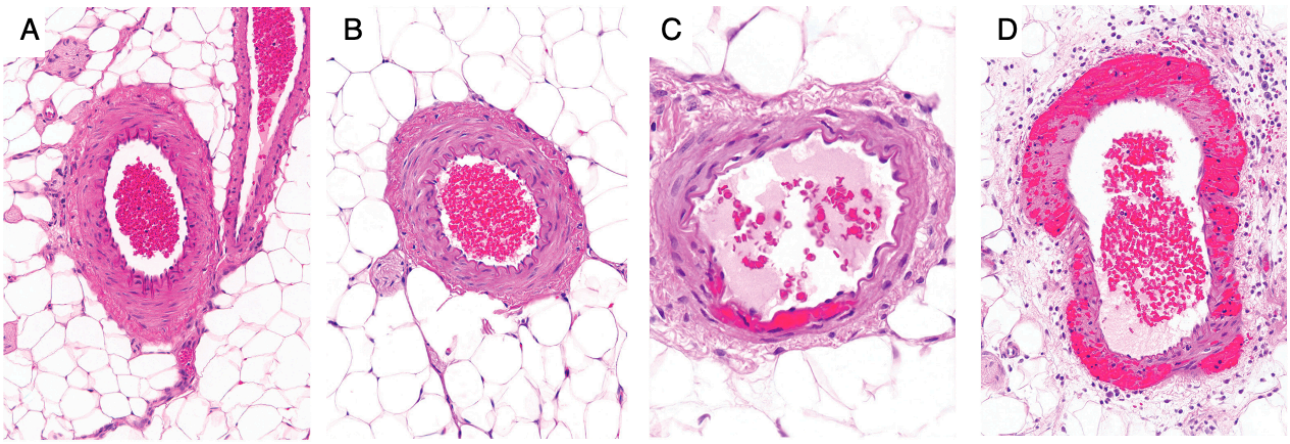


Figure 1.3: Representative haematoxylin and eosin (H&E)–stained sections of mesenteric arteries from rats (40× objective).

Histological samples were taken from the mesenteric arteries of rats treated with vehicle (0.9% sodium chloride) or subcutaneous injection of 100 mg/kg fenoldopam. No microscopic evidence of vascular injury was found in control rats (A) or fenoldopam-treated rats after 4 hours (B). C) At the 12-hour time point after treatment, focal to segmental areas with loss of medial smooth muscle cells were observed, which were replaced by red blood cells. D) At the 24-hour time point, segmental to circumferential loss of medial smooth muscle cells, replacement by an accumulation of red blood cells and perivascular acute inflammatory cell infiltrates were observed. Figure reprinted from Dalmas *et al.* [60] with permission.

While histopathology can give very detailed insights into injury on the tissue level, its disadvantage is that it is invasive, costly, and time-consuming, which is why alternative strategies are desired using which injury can be anticipated *in vitro* or monitored non-invasively. To develop these, histopathology needs to be integrated with other data sources which can provide information on associated changes on the molecular or cellular level. In this thesis, histopathology data is hence integrated with transcriptomics data to identify safety biomarkers (Chapter 2) and to derive mechanistic insights into pathogenesis (Chapter 3-4).

1.3 Pathogenic processes studied in this thesis

In this thesis, multiple pathogenic processes are studied. While all of them centre on injury and hence share some of the responses outlined above, they are caused by different stressors, are found in different tissues (and organs), and are relevant to drug discovery for different reasons. For the respective pathologies, an introduction to the biological background and their relevance for drug discovery is provided.

1.3.1 Drug-Induced Vascular Injury (DIVI)

1.3.1.1 Relevance for drug discovery

Drug-induced vascular injury (DIVI) can be induced within hours of drug administration and is phenotypically identified as morphological changes in the vascular wall including but not limited to medial arterial necrosis (MAN), one of the main hallmark lesions [61]. In nonclinical species, DIVI manifests across a wide range of structurally and pharmacologically diverse compound classes [62]. However, the clinical relevance of findings in animals is unclear and several compounds inducing DIVI in model species, such as minoxidil [63], the phosphodiesterase IV inhibitor apremilast [64] or caffeine [65], have not been reported to date, to result in DIVI in humans.

This poses a hurdle for assessing drug safety, as preclinical evidence can result in delays in drug development and termination [61], although it is often unknown whether observations may or may not be relevant to man. One reason for this is the lack of non-invasive, sensitive, and specific methods to detect the presence or predict the development of drug-induced vascular lesions, leaving histopathological examination the only option. Unlike cardiovascular effects, for instance, DIVI is often found to show local vasoactivity which cannot be monitored via heart rate or blood pressure. Furthermore, the location or even occurrence of DIVI in humans is yet unclear [66]. While DIVI is predominantly found in the coronary arteries of dogs, the mesenteric arteries are rather affected in rats. Caution in this regard arises from a limited understanding of the longer-term consequences of DIVI which may include chronic vascular injury, or cardiovascular morbidity [67]. To support informed

decision-making, current research efforts hence focus on identifying non-invasive safety biomarkers, a better mechanistic understanding of DIVI as well as better model systems.

1.3.1.2 Vascular physiology and mechanisms of DIVI

The vascular wall consists of three layers, also known as *tunica intima*, *tunica media* and *tunica externa* (Figure 1.4). The innermost layer, *tunica intima*, entails one layer of endothelial cells in direct contact with the bloodstream. This endothelial layer takes in key functions in vascular response to stress and injury. First, it shows properties which are typical for immune cells [68,69], e.g. they are phagocytic [70], recognise damage- and pathogen-associated patterns [71] and can present antigens [72]. Moreover, they modulate immune response via a range of cytokines and chemokines and also recruit leukocytes directly through various adhesion molecules [69,73], play a key role in angiogenesis [74], and can undergo endothelial to mesenchymal transition (EndMT), which has been observed across various cardiovascular diseases [75]. Other physiological functions include the regulation of metabolic homeostasis and vascular permeability [76], as well as the regulation of vascular tone together with the vascular smooth muscle cells in the *tunica media*.

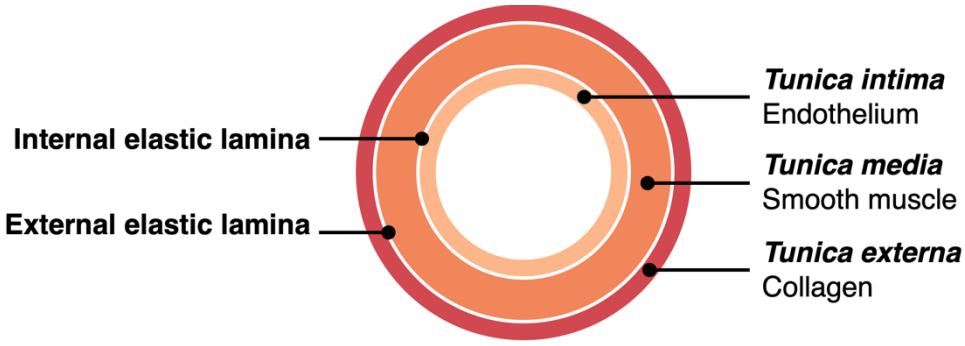


Figure 1.4: Structure of the arterial wall.

The arterial wall consists of three layers which are separated by two elastic laminas. The innermost and middle layers contain endothelial and smooth muscle cells, respectively, while the outermost layer is characterised by extracellular matrix, in particular collagen.

The middle layer, *tunica media*, is separated from the inner layer through an elastic lamina, and can adjust blood flow through contraction of the vascular smooth muscle cells, which adjusts the diameter of the blood vessel. In response to injury, they can switch to a less differentiated state with functions in tissue repair, as well as higher proliferation and migration [77]. The outermost perivascular layer, the *tunica externa*, is largely formed by

collagen, stabilizes the blood vessel and is connected to the *tunica media* through a second external elastic lamina.

While the biological mechanisms of DIVI are not yet completely understood, multiple modes of injury have been proposed. These include hemodynamic effects, such as shear or hoop stress, which can result in biomechanical injury [78], direct toxicity to or dysregulation of vascular cells, either vascular smooth muscle cells or endothelial cells, or injury mediated by inflammation and immune response [61].

1.3.2 Drug-Induced Liver Injury (DILI)

1.3.2.1 Relevance for drug discovery

Drug-Induced Liver Injury (DILI) is a major concern in drug discovery as it is a main cause for drug failure in clinical stages, for issuing warnings and modifications of use, and was also the most common adverse drug reaction leading to post-marketing-withdrawal, attributing to 81 out of 462 withdrawals between 1953 and 2013 [79]. Furthermore, it is one of the leading causes of acute liver failure and liver transplantation, highlighting also its relevance for patients and healthcare systems [80].

DILI is observed across nearly all compound classes, with varying incidence and severity [81]. It can manifest as hepatitis, cholestasis or a mixture of both, which can be characterised based on serum levels of alanine aminotransferase (ALT) and alkaline phosphatase (ALP) [82], and it can additionally be classified as intrinsic or idiosyncratic toxicity. In case of intrinsic toxicity, adverse effects are dose-dependent and consequentially predictable, such as for acetaminophen (paracetamol), which is the cause of 75% of acute liver failure cases attributed to DILI [83]. In contrast, effects are generally less frequent for idiosyncratic toxicity with risk estimates ranging from 1 in 10,000 to 1 in 100,000 for individual medications [84]. These depend on host-specific factors, which e.g. influence the formation of reactive metabolites as well as downstream immune response and explain an individual's susceptibility to DILI [85] (Figure 1.5).

To reduce risks to patients and avoid drug failure in the late stages, it is key to anticipate risks for DILI earlier in the drug discovery process [86]. This includes the development of better

assays and safety biomarkers, as well as better computational predictive models. In our previous work, we built models based on the compound annotations from DILIRank [87] and SIDER [88] which classify compounds with and without risk of DILI based on their chemical structure [89]. However, deriving accurate and meaningful DILI labels for compounds is a complex process given the uncertainty of causality assessment and the difficulty in incorporating relevant factors such as drug administration and patient populations [90]. Furthermore, phenomena such as idiosyncratic DILI which cannot usually be detected even in preclinical studies make the task of accurate DILI labelling and prediction from chemical structure even harder. While associated targets and pathways in DILI were successfully inferred from our chemical structure-based models by incorporating protein target prediction with PIDGIN [91], this also highlights that a better understanding on intermediate levels using biological data is needed.

1.3.2.2 Liver physiology and mechanisms of DILI

The liver is a key metabolic organ, and multiple drug and host properties can contribute to the development of DILI (Figure 1.5). It which takes in a crucial role in energy and nutrient homeostasis, e.g. through the metabolism of lipids [92], glucose [93], and amino acids [94]. Hepatocytes also produce bile acids via cholesterol catabolism which are then stored in the gallbladder from where they can be released into the intestinal tract after each meal entering enterohepatic circulation [95]. Furthermore, the liver is also involved in the metabolism of drugs, which is separated into phase I reactions, where functional groups are added e.g. by members of the cytochrome P450 (CYP450) superfamily, and phase II reactions in which the compound or metabolite is conjugated with an endogenous substance such as glucuronic acid, acetate, or glutathione [96].

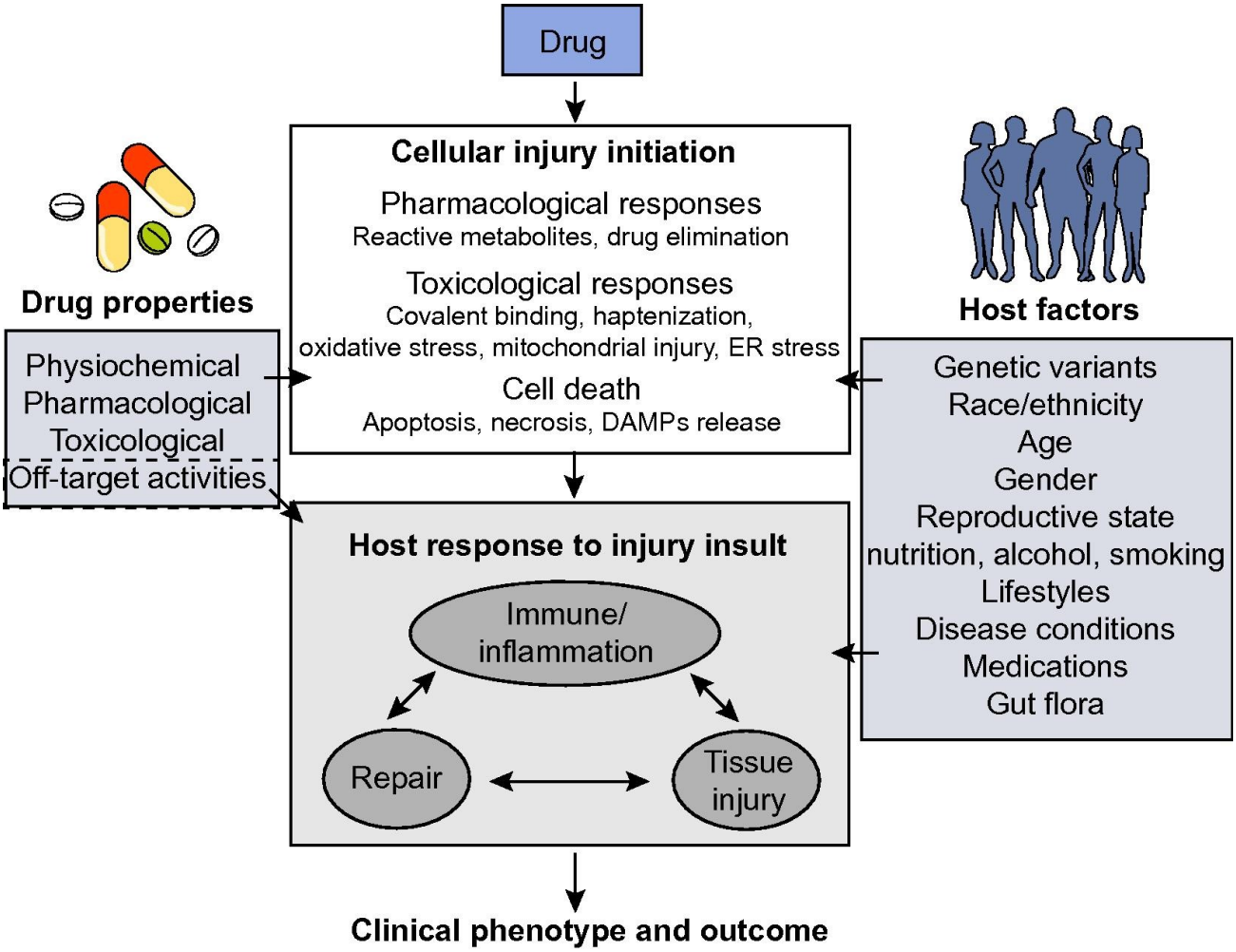


Figure 1.5: Drug properties and host factors contributing to drug-induced liver injury (DILI).

Drugs or their reactive metabolites can initiate cellular injury resulting in cell stress and death. This results in a host response on the tissue level which aims to resolve injury. If injury is not resolved, this can lead to clinical manifestation of DILI which impacts liver function. Thereby, individual drug properties as well as host factors can contribute to injury initiation and response. Figure reprinted from Chen *et al.* [97] with permission.

As biotransformation and energy metabolism are key liver functions, it is not surprising that these can be affected in DILI or can also be its direct causes (Figure 1.5). For instance, one way how DILI can be induced is through the formation of reactive metabolites [98], which then form adducts with endogenous macromolecules resulting in direct toxicity or immune-mediated reactions [99]. Thereby, some of the substructures linked to reactive metabolites are known and can guide lead optimisation to avoid toxicity [100] Furthermore, mitochondria, which are central for liver metabolism [101], are known to be affected in DILI pathogenesis, e.g.

through impairment of mitochondrial fatty acid oxidation or respiration [102], and a retrospective analysis found that half of all drugs with black-box warnings for hepatotoxicity are linked to mitochondrial toxicity. Lastly, accumulation of bile acids in the liver can result in cholestatic DILI, and multiple transporters involved in biliary efflux have been identified as off-targets linked to DILI, such as the bile salt export pump (BSEP) [103]. Because of this, bile acid levels in systemic blood have also been proposed as metabolic biomarkers [104].

At the same time, the liver is outstandingly equipped to deal with injury as it is the only solid organ with the ability to completely regenerate, even after partial hepatectomy with a loss of ~70% of liver cell mass [105,106]. Consequentially, a clinical phenotype is only developed if the injury is not resolved by the intrinsic cell- or tissue-level stress response. In this context, previous studies have highlighted adaptive stress response pathways, especially oxidative stress and ER stress, on the cellular level, and immune-mediated DILI as a tissue-level phenomenon [86].

1.3.3 Idiopathic Pulmonary Fibrosis (IPF)

1.3.3.1 An unmet medical need

Idiopathic pulmonary fibrosis (IPF) is a form of chronic interstitial lung disease (ILD) which affects ~3 million people worldwide and shows increasing incidence rates, also because it primarily affects elderly adults. It is characterized by progressive lung scarring, thickening of the interstitium and continuous decline in lung function, which results in shortness of breath, also known as dyspnoea, and cough. This severely affects the patient's quality of life and also is associated with a poor prognosis with a median survival rate of 3-5 years if untreated [107].

Currently, two treatments are approved for IPF: Nintedanib [108], which inhibits tyrosine kinases involved in proangiogenic and pro-fibrotic pathways, and pirfenidone [8], which has anti-fibrotic, anti-inflammatory, and antioxidant activity. Both small molecules effectively reduce disease progression and the decline of pulmonary function but are not able to improve lung function or quality of life [109]. While severe side effects are rare, drug discontinuation due to adverse events was still found in 20.9% of subjects on pirfenidone and 26.3% on nintedanib [110,111]. This leaves lung transplantation as the only cure and only

alternative intervention, but this is only possible for a minority of patients due to the limited organ availability as well as the comorbidities and age of the patients [112]. Overall, this combination of a large and growing patient population, severe and deadly disease burden, and limited availability of treatment options makes IPF a societal unmet medical need [113,114].

1.3.3.2 Risk factors and pathogenesis

Multiple host-related risk factors have been identified that increase susceptibility to IPF. This includes genetic variants, with multiple ones pointing to telomere maintenance, epigenetic reprogramming which is associated with ageing [107], as well as environmental factors, such as smoking and occupational exposures to which the lung may be particularly susceptible due to its direct exposure [115].

In a predisposed individual, IPF is thought to be induced through recurrent micro-injuries to the alveoli, which are air-filled sacs where the gas exchange in the lung occurs. In general, there are two types of epithelial cells in the alveolus referred to as alveolar type 1 (AT1) and alveolar type 2 (AT2) cells. Under physiological conditions, AT1 cells are primarily involved in gas exchange, while AT2 cells produce surfactants. Furthermore, AT2 cells act as alveolar epithelial stem cells with the ability to self-renew and differentiate into AT1 cells. While largely quiescent during homeostasis, subpopulations of AT2 cells were found to rapidly expand to regenerate the alveolar epithelial cell population upon injury [116–118].

Recurrent injury to the alveolar epithelium, especially AT2 cells, was identified as a cause of pulmonary fibrosis, and, although not fully understood mechanistically, results in the recruitment and proliferation of fibroblasts and myofibroblasts [119,120]. Histologically, this is characterized by high spatial variability and fibrogenesis is thought to take place in so-called fibroblastic foci, the extent of which is predictive of survival [121]. Furthermore, IPF is characterised by a reduced AT1 population which restricts the overall gas exchange, while AT2 cells show a hyperplastic phenotype [122]. Ultimately, this leads to an accumulation of extracellular matrix, continuous scarring and loss of lung function.

1.3.3.3 Alveolar regeneration as an emerging therapeutic strategy

Previous research focussed on targeting inflammation [123] and myofibroblasts [124] as early events linked to the pathogenesis of IPF, however, has not been successful in finding cures.

Recent studies found that promoting alveolar regeneration is sufficient to resolve fibrosis and restore pulmonary function in animal models [125,126]. Promoting endogenous alveolar repair has hence emerged as a promising therapeutic target, which may not only slow down or stop disease progression but can potentially even cure IPF by regenerating lung function [127].

Recently, an intermediate progenitor cell state was identified in AT2 to AT1 differentiation which expresses epithelial and mesenchymal markers, as well as markers of senescence such as TGFβ which indicates a pro-fibrotic role [128–130] (Figure 1.6). This transitional cell state was also found to persist in fibrotic regions of IPF lungs, and explicitly around foci of high collagen expression [128,129]. This suggests that the terminal differentiation from this transitional cell state to functional AT1 cells is inhibited and that it contributes to fibrogenesis.

In the murine bleomycin lung injury model, which is one of the most extensively used and best-characterized preclinical models for acute and chronic lung injury [131], the AT2 to AT1 trajectory was further characterized using scRNA-Seq identifying an intermediate stem cell state expressing similar markers, including pro-fibrogenic factors and distinct cell-cell communication with mesenchyme and macrophages [132,133]. Additionally, chronic inflammation by sustained IL-1b treatment was found to inhibit the terminal differentiation to functional AT1 cells, also resulting in an accumulation of intermediate progenitors [133].

Promoting the intermediate progenitor to AT1 transition is hence therapeutically interesting from two angles: To restore the depleted AT1 population which may help to restore lung function, and to reduce the level of pro-fibrotic intermediate progenitors.

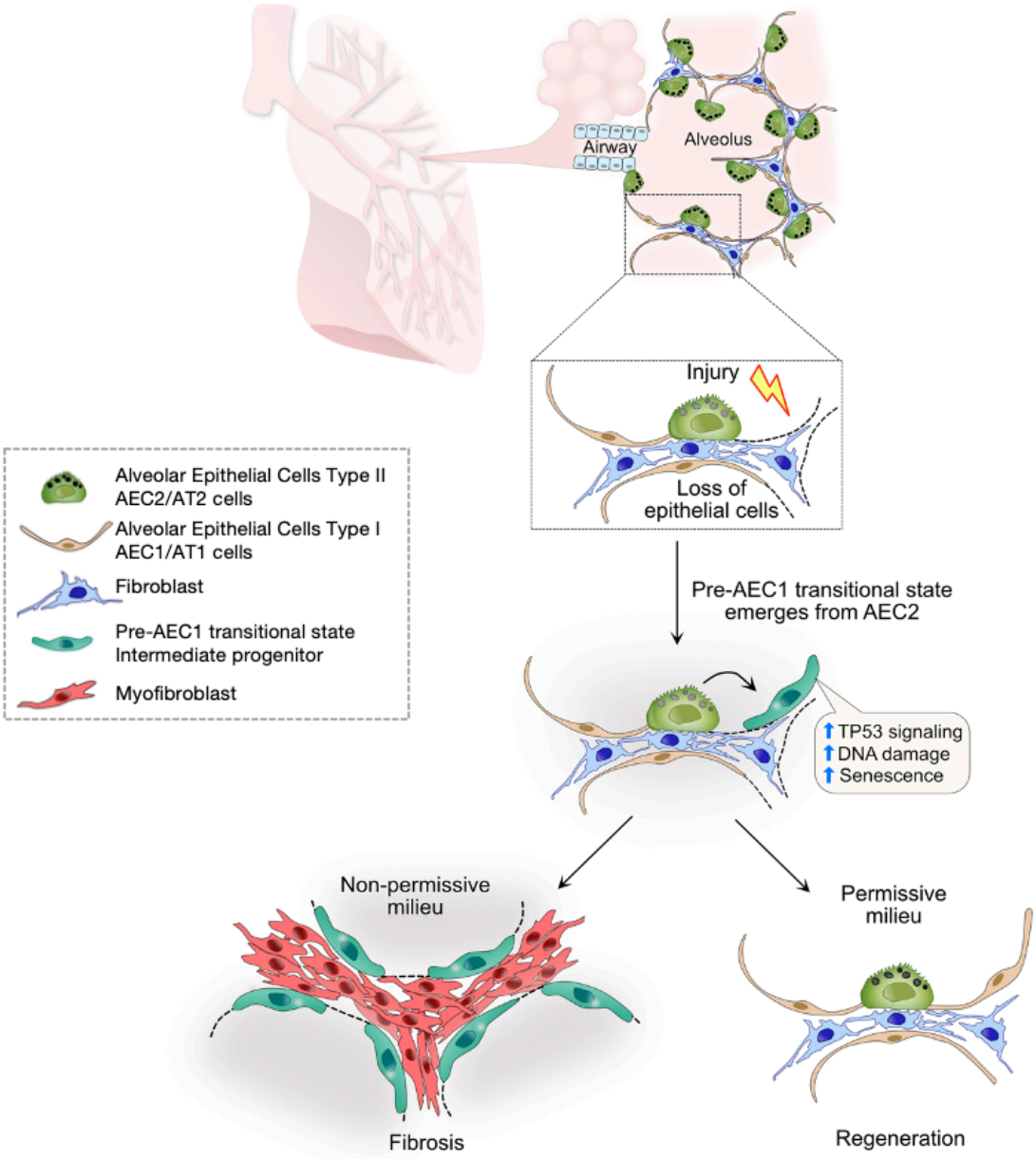


Figure 1.6: Emergence of an intermediate progenitor cell state in alveolar regeneration.

AT2 cells replicate in response to damage and generate intermediate progenitors which normally mature into functional AT1 cells but persist in fibrotic lungs. The intermediate progenitor is regulated by TP53 signalling, vulnerable to DNA damage and undergoes a transient senescent state. Figure adapted from Kobayashi *et al.* [128] with permission.

1.4 Introduction to transcriptomics data

1.4.1 Omics data and cellular regulation of gene expression

In a multicellular organism, the DNA, and hence the genetic information encoded in its nucleotide sequence, is largely identical between its cells apart from errors introduced in the replication process. Instead, the structural and functional differences between cell types or individual cells can be explained by how the genes are expressed resulting in different sets of proteins. These execute the majority of cellular functions, e.g. by selectively catalysing chemical reactions as enzymes. The central dogma of molecular biology describes this flow of genetic information in a biological system and entails 1) the replication of DNA, 2) the transcription of DNA to RNA, and 3) the translation of a subset of these RNAs, the so-called messenger RNA or mRNA, to protein (Figure 1.7).

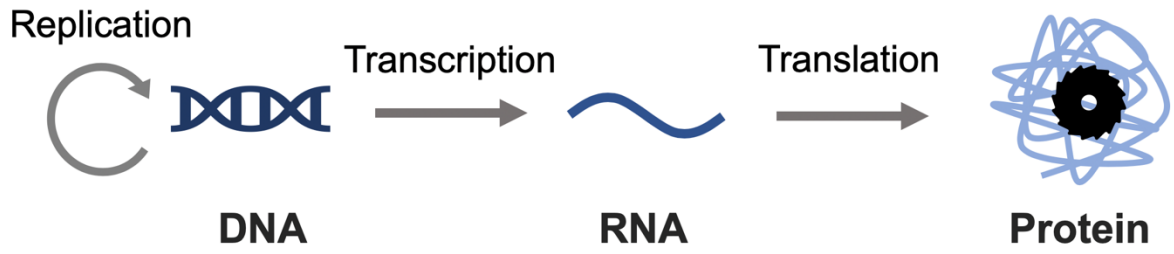


Figure 1.7: Central dogma of molecular biology.

The central dogma of molecular biology describes the transfer of information from DNA to RNA to protein. DNA can be copied to DNA (replication) or RNA (transcription). Subsequently, proteins can be produced using the RNA template (translation).

Each of these steps is tightly regulated while additional regulation can occur on the epigenetic, post-transcriptional and post-translational levels. This regulation toolbox enables fine-tuning of cellular processes on different temporal and spatial scales, e.g. as a response to signals from surrounding cells or other environmental changes. Studying these changes across different tissues, conditions or timepoints can then provide useful insights into a given phenotype, such as the molecular characteristics of a disease or the MoA of a drug.

To obtain a snapshot of the cellular state, a wide range of technologies has been developed which are continuously improving in terms of coverage, scalability and resolution [134]. In this

context, “omics” refers to approaches which measure the abundance of different types of biological entities, such as genes, RNA transcripts or proteins, on a large scale [135]. These can give a broad overview of the underlying biology, while each type of biological entity provides a different angle given that they are affected by different levels of regulation. For instance, transcriptomics measures the abundance of RNA transcripts within a cell providing insights into which genes are transcribed. In contrast, proteomics measures the abundance of proteins, which is influenced by both transcription and translation. Numerous additional types of “omics” exist and these can jointly give a detailed view on each step of the underlying cellular regulation, which has become the focus of multi-omics studies [136,137]. In this thesis, the focus will be on the analysis of transcriptomics data which gained early popularity in the 1990s with the development of DNA microarrays [138] and has since been transformative for our biological understanding [139].

1.4.2 Transcriptomics technologies

1.4.2.1 DNA Microarrays

From the technical side, DNA microarrays are chips on which various short oligonucleomers, also known as probes, are fixated which are designed to match certain transcripts of interest. The sample-derived RNA can be first amplified and is then used directly or indirectly to generate nucleic acids which are labelled with reporter probes. Thereby, the RNA can either be labelled directly or can be transcribed to cDNA or cRNA incorporating nucleotides which are either labelled with a fluorophore or biotin, so that they can subsequently be stained with labelled streptavidin[140]. When the labelled nucleic acid is then washed over the microarray, it hybridizes to the complementary probes, and the amount of reporter signal for each probe then indicates the abundance of this particular mRNA species within the sample.

Due to unspecific binding to the probes, there can be background noise, and due to the limited number of probes with the same sequence, there is also an intrinsic upper limit for quantification. As the number of potential probes is limited, these need to be carefully designed requiring reliable knowledge about the genome of interest which has led to a wide range of commercial microarray technologies, some focussing more on high throughput while others instead prioritize high coverage [141]. The two platforms from which data was used in this thesis will be described further below.

Affymetrix GeneChip arrays have been widely employed as they are able to cover up to 6.5 million features on a single 1.28 cm^2 array [142]. These high-density DNA arrays are generated using a photolithographic manufacturing process which allows the parallel synthesis of the 25-base long oligonucleotide probes [141]. One platform of particular interest in toxicology is the GeneChip™ Rat Genome 230 2.0 Array (GEO ID: GPL1355), which comprehensively covers the rat transcriptome with 28,000 genes. It is also the platform which was used to generate the transcriptomics data analysed in Chapters 2, 3 and 4.

The L1000 platform achieves a scale-up of more than 1,000-fold in comparison to Affymetrix microarrays by only measuring the expression of 978 landmark genes [143]. These were selected to be non-redundant so that the expression of 9,196 genes can be predicted with high fidelity. As part of the LINCS (Library of Integrated Cellular Signatures) project, the L1000 platform was used to characterize 19,811 compounds making it the most comprehensive resource for uniformly generated transcriptomic data on chemical perturbations [143].

1.4.2.2 Bulk and single-cell RNA Sequencing

As for microarray studies, the RNA is first transcribed to cDNA in RNA Sequencing (RNA-Seq). But then, the sequence of the cDNA is directly determined either from one end (single-end sequencing) or both (pair-end sequencing) [144]. As a consequence, in theory, any kind of RNA can be detected with the same experimental setup which enables higher flexibility, e.g. if genomic information is not available yet for the given organism, and also each sequence is identified with single-base resolution, which means that these sequence-based methods can give more detailed information e.g. about splicing variants or mutations.

In contrast to hybridization-based platforms, there is no background signal because sequences can be clearly mapped to the respective genomic regions, and there is no upper limit for quantification. Instead, the quality of the generated data largely depends on the sequencing depth, so the number of transcripts sequenced, which impacts whether rare transcripts are detected and how accurately the relative abundances can be determined.

Since the development of bulk RNA-Seq, which measures all transcripts within one sample, new technologies have been developed with the ability to map transcripts to individual cells

(single-cell RNA-Seq) or to spatial regions within a tissue (spatial RNA-Seq) which is particularly interesting to study cell-cell communication and effects on the tissue-level. Both are fundamentally possible by tagging transcripts with so-called barcodes, short oligonucleotide sequences, through which each transcript can be mapped back although all transcripts within a sample are sequenced together [145]. While the scientific implications of single-cell RNA-Seq (scRNA-Seq) will be discussed in Chapter 5 where scRNA-Seq data is utilised, the technology itself will be discussed here.

There are multiple droplet-based scRNA-Seq platforms, including inDrop [146], Drop-Seq [147] and 10X Genomics Chromium [148], which follow similar experimental protocols. First, individual cells are encapsulated in droplets using microfluidic devices, and these droplets also contain barcoded beads with oligonucleotide primers. Each barcode contains an oligo(dT) sequence through which mRNAs with complementary poly A tails are captured once the cells are lysed. Furthermore, the primer contains two additional components which eventually characterize the bound transcript: The cell barcode, which is identical across all primers on a bead and hence identifies each droplet, and the unique molecular identifier (UMI), which is variable across all primers and therefore identifies each individual transcript. While mRNA capture needs to take place within the droplets, the subsequent steps can then occur within the bead or after demulsification [149]: The captured RNA is first reverse transcribed to cDNA, which is then amplified and, after demulsification, sequenced.

1.4.3 Small-molecule transcriptomics datasets

Transcriptomics data can be used to evaluate the biological effects induced by treatments in any biological system and has been widely used to study perturbation effects across compounds, doses and timepoints. The largest publicly available datasets are summarised in Table 1.4. In this context, the Open TG-GATEs (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System) [150] and DrugMatrix [151] and databases should be highlighted which contain transcriptomics, as well as clinical chemistry, haematology and histopathology data from multiple organs in rats across 170 or 627 compounds, respectively (Table 1.4). These are invaluable resources in the field of toxicogenomics [152,153] as they contain multiple types of information within the same animal through which it is possible to derive mechanistic links between gene expression and phenotype *in vivo*.

Table 1.4: Public transcriptomic datasets.

Dataset	Technology	Model system	Number of compounds	Number of replicates	Number of doses*	Timepoint of sampling
Open TG-GATEs [150]	Microarray	Rats	170	3	3 doses	3h, 6h, 9h, 12h, 24h, 4d, 8d, 15d, 29d
DrugMatrix [151]	Microarray	Rats	627	3	Mostly 1 or 2 doses	Mostly 1,3,5 and 0.25 days
DrugSeq [154]	Targeted RNA-seq	U2OS cells	433	3	8 doses	12h
LINCS [143]	Targeted Microarray (L1000)	71 cell lines, mostly VCAP, MCF7, PC3, A549	19,811	Variable	Mostly 10 μM and 5 μM	Mostly 24h and 6h
CMap [155]	Microarray	Mostly MCF7, also PC3, HL60, SKMEL5, ssMCF7	1,309	Mostly 1-2	Mostly 10 μM	Mostly 6h, also 12h
Sci-Plex [156]	scRNA-Seq	A549, K562, MCF7	188	2 (~ 100 – 200 cells each)	4 doses	24 h

*Excluding vehicle control

From a drug discovery perspective, transcriptomics has not only been interesting as a means to better understand the biological effects of compounds but also to phenotypically characterize compounds in the early stages of drug development as complementary information to compound structure [157]. To profile large amounts of compounds in a cost-efficient manner, cancer cell lines are most frequently used as model systems as well as technologies with a stronger focus on high throughput. In this regard, the connectivity map project by the Broad Institute should be highlighted which includes the first CMap version [155] ("CMap" in Table 1.4) which contains microarray-based perturbation signatures and its successor, which contains signatures derived using the L1000 platform and is part of the Library of Integrated Network-Based Cellular Signatures (LINCS) program [143].

Given a particular transcriptional signature, these can then be queried to identify potentially related small molecule perturbations, e.g. based on the non-parametric Kolmogorov-Smirnov statistic which determines the enrichment of a signed list of genes with respect to a rank-ordered reference signature as introduced by Lamb *et al.* [155]. Furthermore, previous studies have explored perturbation signatures as means to group mechanistically similar compounds [158] or to predict other *in vitro* [159] or *in vivo* [160] properties. Overall, transcriptomics can hence contribute to a better biological characterization of compounds in general and to various drug discovery tasks in practice.

1.5 Computational analysis of transcriptomics data

While the pre-processing of transcriptomics data is platform-specific and will hence be discussed within the respective Methods sections, there are common steps in the analysis of the derived gene expression matrix which contains expression levels across genes and samples. These are relevant across all subsequent chapters and their underlying basis will therefore be introduced in this section.

1.5.1 Differential expression

It is often not the expression within one sample that is of interest but rather changes in expression across samples and especially comparisons between two categorical groups are of interest, such as between healthy and disease or between treatment and control. This is also referred to as differential expression analysis, and the relative difference in magnitude is generally quantified as fold change (FC) or logFC. However, a large fold change may be caused by fluctuations in expression which can arise from technical variation and could mask the real biological signal. Therefore, an additional statistical test should be used which evaluates whether there is a difference in average expression level across both groups.

Depending on the type of gene expression data, different statistical approaches are appropriate as e.g. continuous microarray data is generally modelled as a normal distribution while Poisson or negative binomial distributions are utilized in case of count data from RNA-Seq. For instance, a well-established approach for microarray data analysis is the moderated t-statistic implemented in the R package `limma` (LInear Models for MicroArrays) [162]. Here, fold changes and standard errors are first fitted using linear models per gene and the standard errors are then smoothed across genes using an empirical Bayesian model eventually resulting in the moderated t-statistic. This was found to increase power and provide more stable results, particularly in the case of a low sample size [163]. Similarly, information on the variance across genes can also be leveraged in the analysis of count data. For example, empirical Bayes is used to moderate the dispersion in the negative binomial models in both `edgeR` [164] and `DESeq2` [165]. Combining the given test statistic and the logFC, it is then possible to identify the most strongly and significantly up- and down-regulated genes which may hence play important roles in the phenotype at hand.

1.5.2 Biological interpretation using gene sets and signatures

The interpretation of individual genes can be limited as single genes or proteins can take in diverse functions depending on different interaction partners. To interpret differential expression signatures, it is therefore highly effective to regard gene expression in a functional context, i.e. by taking advantage of prior knowledge on cellular processes.

One well-established approach to interpret gene expression is to consider functional groups of genes as so-called gene sets, and then to look into the dysregulation of this group of genes as a whole in order to gain more interpretable insights on a higher biological level. Knowledge-driven gene sets can for instance be derived based on annotations of the gene product, e.g. from Gene Ontology (GO) [166], and can include information on molecular function, role in biological processes or also subcellular location. Other resources which have been widely used in the analysis of transcriptomics data include Reactome [167], KEGG [168] and Wikipathways [169] which focus more on biological pathways.

A different type of knowledge-driven genesets are regulons which describe groups of genes regulated by a transcription factor (TF). This information can be inferred and derived from different sources of information, such as known TF binding motifs on promoters, peaks from chromatin Immunoprecipitation DNA-sequencing (ChIP-Seq) or literature-curated resources, and can then be used to infer the upstream activity of TFs based on the expression of genes which it regulates [170].

Besides knowledge-driven approaches, gene sets can also be derived in a data-driven way from historical transcriptomics data, or more precisely historical differential expression signatures. This can then uncover similarities between the cellular response at hand and previous experiments, which may provide more detailed mechanistic insights, e.g. if it is known which gene or protein was perturbed in the previous experiment [171], and can also identify shared cellular phenotypes, e.g. when comparing to a disease signature [155]. Furthermore, also text-mining-based gene sets can be employed which describe genes mentioned within the same publication and may detect more recently published biological findings than curated databases [172].

A wide range of methods has been developed to contextualise gene expression data using gene sets, including gene set analysis [173] (GSA) which tests whether more genes of that group are found to be dysregulated than at random, gene set enrichment analysis [174] (GSEA) which tests whether genes of that group are found to be more strongly dysregulated than at random and gene set variational analysis [175] (GSVA) which describes variation of gene set activity across samples based on the variation of the group members. The concrete test statistic will be introduced within the chapters.

1.5.3 Multiple hypothesis testing and correction

In statistical testing, the aim is to identify results based on samples which are not observed by chance but can be attributed to a specific cause. To do so, it is estimated how likely the observed data would have occurred by chance under the premise of the Null hypothesis. A result from an individual test is then considered statistically significant if the likelihood of retrieving the observed or a more extreme result by chance, the p-value p, is below the pre-defined significance level α, the tolerated false positive rate. Here, it should be noted that neither statistical significance nor a specific p-value implies plausibility, presence, truth, or importance [176].

However, if multiple tests are performed simultaneously as part of a statistical analysis, such as in case of high-dimensional biological data, the overall probability of false positives is expected to increase. One strategy to reduce the probability of false positives is to reduce the number of hypotheses, if it is possible to pre-define those which are of interest [177]. Furthermore, statistical strategies exist to adjust p-values, out of which two commonly used ones will be introduced. The Bonferroni procedure [178] regards tests as significant if the family-wise error rate (FWER), the probability of identifying one or more false positives in multiple testing, is below the significance level ($FWER < \alpha$). From a practical perspective, this means that instead of rejecting individual hypotheses at $p < \alpha$, they are then being rejected at $p < \alpha/m$. The Bonferroni procedure is generally conservative, which means that the number of false positives (type I error) is kept low at the cost of potentially wrongly rejecting hypotheses (type II error). A less conservative approach is the Benjamini-Hochberg procedure [179] which instead requires the false discovery rate (FDR), the fraction of false positives among all discoveries, to fall below the significance level ($FDR < \alpha$). Practically,

this means $p_k < k * \alpha/m$ with k corresponding to the rank of p_k among all p observed within the test family ordered from smallest to largest. In this thesis, "p-value" always refers to unadjusted p-values, while "FDR" is used when the Benjamini-Hochberg procedure was applied.

1.6 Applications of transcriptomics in drug discovery

1.6.1 Identification of safety biomarker candidates

According to the FDA/National Institute of Health (NIH)'s BEST (Biomarkers, Endpoints, and other Tools Resource) guide, a biomarker is "a defined characteristic that is measured as an indicator of normal biologic processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions" [180.] Thereby, it can be measured using any technology, including transcriptomics, and includes molecular, histologic, radiographic, and physiologic readouts.

In the context of drug safety, biomarkers can be used to detect and monitor adverse effects *in vivo*, e.g. in preclinical or clinical trials [181]. These ideally indicate toxicity before a severe clinical phenotype occurs, which allows more time- and cost-efficient identification of risks in nonclinical and clinical trials, as well as better personalization of the treatment regimen post-approval. As biomarkers should be measurable non-invasively and rapidly, in particular, circulatory proteins have been of interest as molecular biomarkers.

In contrast, tissue-derived transcriptomics is rather used to generate biomarker candidates and mechanistic hypotheses due to its broad coverage. Thereby, how to prioritise biomarker candidates is not straightforward as the desired properties cannot be summarised into a single metric for ranking, but instead multiple features contribute to a good biomarker which may also vary depending on the desired context of use [181]. Generally, biomarkers should distinguish adverse and non-adverse conditions, and anticipate injury or severe injury before it manifests. Furthermore, they need to be detectable in a robust manner, ideally in a non-invasive or minimally invasive way, and if biomarkers are intended for monitoring, they should reflect the severity of injury. In this context, it should be noted that biomarkers can be purely descriptive, however, mechanistic ones can be more informative [182].

To use biomarkers for regulatory decision-making, these need to be qualified which is defined by BEST as "a conclusion, based on a formal regulatory process, that within the stated context of use (COU), a medical product development tool can be relied upon to have a specific interpretation and application in medical product development and regulatory review" [180]. Pharmaceutical companies, regulatory bodies, and academic groups are working together towards this common goal, e.g. through the Predictive Safety Testing Consortium (PSTC) [183] or the Innovative Medicines Initiative (IMI) Consortium TransBioLine [184].

It is clear that the high confidence needed for regulatory acceptance cannot be provided by data-driven analysis alone. However, it can be used as one source of evidence and can prioritize biomarker candidates which are more likely to be successful.

1.6.2 Identification of adverse event mechanisms

When studying the pathogenesis of adverse events, the goal is to identify not only statistical associations but causal cascades of events which describe how specific molecular interactions result in a systems-level phenotype, as this can then help to anticipate or even prevent adverse effects through earlier or simpler surrogate readouts. In the context of AOP development, previously introduced in 1.1.2, the Organization for Economic Co-operation and Development (OECD) [185] published three criteria to evaluate the causality of key event relationships based on the original Bradford Hill considerations in the context of epidemiological studies [186] and previous work on the related mode of action concept [25] (Table 1.5): Biological plausibility, essentiality of key events and empirical support for the key event relationship (KER).

Table 1.5: Tailored Bradford-Hill considerations for AOPs.

Multiple sources of evidence can contribute to establishing a causal relationship between E_{cause} and $E_{consequence}$. Adapted from the OECD's users' handbook supplement to the guidance document for developing and assessing AOPs [185].

Criteria		Description
Biological plausibility		Known structural/functional relationship between E_{cause} and $E_{consequence}$
Essentiality of key events		E_{cause} is necessary or even sufficient to induce $E_{consequence}$
Empirical evidence	Time concordance	E_{cause} is observed before $E_{consequence}$
	Dose concordance	E_{cause} is observed at a lower dose than $E_{consequence}$
	Incidence concordance	E_{cause} affects a larger population than $E_{consequence}$

Each source of evidence can thereby be further weighed depending on the certainty and consistency of evidence, e.g. across biological systems or compounds [187]. Besides the strength of evidence, additional quantitative information, e.g. on the predictive performance of a KER, can be useful to better understand how the KER can guide decision-making. While in most cases, KERs are so far qualitative, annotating such information is the aim of quantitative AOPs (qAOPs) [188].

Computational approaches can support these predominantly expert- and knowledge-driven efforts. For instance, computationally predicted AOPs (cpAOPs) prioritize events and KERs as starting points for expert-driven AOP development based on biological plausibility by integrating functional and statistical associations between biological entities on different levels [189–191]. However, it should be noted that these links across biological scales are in many cases not causal and incomplete. Hence, these generally do not present causal evidence, but rather support hypothesis generation in expert-driven AOP development by depicting the biological understanding at a given time, e.g. as knowledge graphs [190,191].

Furthermore, data on dose-response or time-course behaviour can contribute empirical evidence if the study design enables concordance analysis. While so far efforts in this direction have focussed on evaluating defined sets of KERs in-depth for few or single stressors using targeted readouts (mechanistic qAOPs) [188,192,193], transcriptomics has the advantage that it broadly captures potential MIE and KE, so early or intermediate events on the cell- and tissue-level. Toxicogenomics can hence characterise specific events of interest but also to prioritise potential new events [152].

1.6.3 Drug repurposing

Drug repurposing refers to the identification of new medical indications for approved or investigational drugs. Since these drugs have already passed early steps of the drug development process, including preclinical testing as well as potentially clinical trials, the compound's safety is already more established, and drug development efforts can focus more on efficacy with an overall reduced risk of failure and an accelerated drug development process requiring less investment [194,195].

For potential repurposing candidates, structure and bioactivity are often known and, due to the therapeutic opportunities, additional data has been generated that broadly characterises the biological effects of compounds. In this context, the Drug Repurposing Hub should be highlighted which is a repurposing library broadly covering compounds of diverse chemotypes which have reached clinical development [196]. These assay-ready plates can be used to perform screening on any type of disease model and are also aligned with efforts at the Broad Institute to characterize the general biological response to compound perturbation e.g. using transcriptomic signatures from the L1000 platform [143], previously introduced in 1.4.3, but also image-based readouts derived through Cell Painting [197] or proteomic signatures [198]. Using these readouts, it is e.g. possible to identify compounds with similar effects to already known treatments [199], or to identify compounds with disease-relevant effects [200].

Transcriptomics has the advantage, that it can be both used to profile the effects of compounds on biological systems *in vitro* as well as the pathological changes observed *in vivo*. This is also leveraged in this thesis using the concept of signature matching or connectivity mapping (1.4.3), which prioritizes compounds for repurposing with the ability to reverse the transcriptional changes linked to the disease and hence, potentially, also the phenotype itself [201]. A benefit compared to the above-mentioned strategies is that it directly links diseases to compounds without requiring knowledge about individual targets as an intermediate step (although this can clearly provide additional evidence) and is hence solely data-driven. It has been applied successfully across various disease areas ranging from inflammatory bowel disease to cancer [202–204].

As for all drug repurposing strategies, there are limitations to signature matching which depend on how relevant the information used is for the given disease and drug. Firstly, the approach can only be successful if the disease signature characterizes cellular programs which are causal instead of symptomatic for the disease so that they can be used to reverse or prevent it (Table 1.1). Also, gene expression itself may be symptomatic, e.g. for epigenetic effects, and reversing the transcriptional profiles may hence not reverse the cellular state. Secondly, the expression patterns observed *in vitro* may not sufficiently recapitulate the *in vivo* effect given that they are derived from simpler model systems as discussed in 1.1.1. Finally, in bulk transcriptomics, changes in expression may be attributed to changes in cell

composition, such as immune cell infiltration, which cannot be reversed by modulating gene expression within the present cells [190–192]. scRNA-Seq provides the necessary resolution to not only distinguish between compositional and transcriptional alterations, but also to characterize specific cellular transitions *in vivo*. Hence scRNA-Seq data may enable even more successful applications of signature matching which is explored in Chapter 5.

1.7 Thesis aims

In this thesis, the aim was to use transcriptomics to better detect, understand and treat injury across three organs or tissues (vasculature, liver and lung), with the ultimate goal to contribute to the discovery of safe and effective drugs. In the Chapters 2-4, safety-related endpoints are studied using histopathology and transcriptomics data with the ultimate aim to contribute to a better identification of safety concerns in the drug development process. In contrast, Chapter 5 aims to contribute to efficient drug discovery from an alternative angle, namely by prioritizing repurposing candidates which are already better characterized, e.g. with respect to safety, than completely new molecular entities. While a summary of the chapters is shown in Figure 1.8 the individual goals are further discussed below.

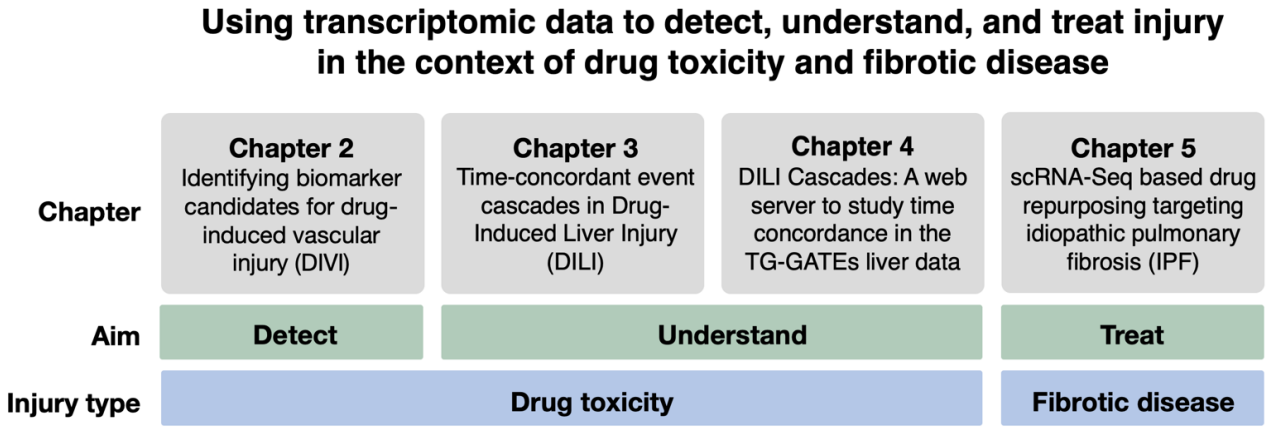


Figure 1.8: Relationship between thesis title and chapters.

Detecting DIVI: Prioritization of safety biomarker candidates

To prioritize biomarker candidates with the ability to detect and anticipate DIVI, a filtering pipeline was implemented to identify genes which show consistent, specific and dose-responsive dysregulation in the vascular smooth muscle or endothelium across multiple

treatments with structurally diverse compounds inducing medial arterial necrosis (MAN). These were further characterized with respect to the degree to which these reflect disease progression, their strength of dysregulation, and known biological functions. Additionally, an interactive dashboard was developed through which the derived biomarker properties can be explored. Overall, this provides informed data-driven starting points for biomarker development and qualification.

Understanding DILI: Inference of time-concordant mechanisms

In Chapters 3 and 4, the aim was to better understand mechanisms in DILI from time-resolved gene expression and histopathology from the TG-GATEs database. To do so, an automatable framework was developed to quantify and characterize time concordance across a large set of time-series. The approach is applied to infer time-concordant cellular events, which precede adverse histopathology. This was able to recover known events involved in the pathogenesis of DILI, and to prioritize potentially novel pathways and TFs which precede adverse histopathology. Furthermore, time concordance is combined with prior knowledge on plausible interactions between TFs to derive causal hypotheses on the TFs' mode of regulation and interaction partners. Overall, this aims to contribute to the development of Adverse Outcome Pathways for DILI.

Treating IPF: Transcriptional repurposing of drugs as regenerative medicine

In Chapter 5, the aim was then to contribute to the discovery of safe drugs from another angle, namely by prioritizing drug repurposing candidates for IPF. In this work, a cell transition in the differentiation of AT2 cells into mature AT1 cells is targeted which is inhibited in IPF and contributes to the regeneration of the alveolar epithelium. It is hypothesized that inducing this transition promotes lung regeneration and can ameliorate the disease. To this end, the intermediate population to AT1 cell transition signature is characterised using multiple recently generated single-cell RNA-Seq datasets on murine bleomycin injury and IPF patients. These signatures are then matched to drug perturbation signatures retrieved from the LINCS database to identify the most suitable candidates for drug repurposing. In combination with bioactivity data, these findings were further interpreted by identifying multiple potentially involved targets.

2 Identifying biomarker candidates for drug-induced vascular injury (DIVI)

This work was published as a pre-print on bioRXiv [205]. The data analysed in this work was provided by GSK as part of a collaboration with Dr Jordi Munoz-Muriedas and Dr Deidre Dalmas. Furthermore, Drs. Valeriu Damian, Jim Harvey and Randall Smith are thanked for their administrative support at GSK.

2.1 Introduction

To support informed decision-making, current research on DIVI focuses on identifying non-invasive biomarkers with the ability to detect vascular injury in patients and/or preclinical animal models early and reliably (1.3.1). Two key consortia with subgroups focusing on research in this area are the Predictive Safety Testing Consortium (PSTC), as part of which the Vascular Injury Working Group (VIWG) studies DIVI in animal models as well as translation across species including humans (`https://c-path.org/programs/pstc`), and the Innovative Medicines Initiative (IMI) Consortium TransBioLine focusing on DIVI in humans (`https://transbioline.com`). Both aim to produce evidence towards biomarker identification, development, and qualification and have curated panels of potential circulating protein-based biomarkers which may reflect mechanisms leading to DIVI pathogenesis based on their association with known histopathological features shared with vascular diseases and/or across species (Table 2.1).

Table 2.1: Circulating DIVI biomarker candidates previously prioritized for further qualification by SAFE-T and VIWG and their hypothesized role in pathogenesis. [206,207]

Predominant specificity	Gene/target name	Symbol
Endothelial cells	Endothelin 1	*Edn1*
	E-selectin (rat)	*Sele* (rat)
	Angiopoietin-2	*Angpt2*
	Thrombospondin 1	*Thbs1*
	Vascular endothelial growth factor	*Vegfa*
Inflammation	Calponin	*Cnn1*
Smooth muscle cells	A1 acid glycoprotein 1	*Agp1*
	C-X-C Motif Chemokine Ligand 1	*Cxcl1*
	Lipocalin 2	*Lcn2*
	Alpha-1-Acid Glycoprotein 1	*Orm1*
	TIMP Metallopeptidase Inhibitor 1	*Timp1*
	Total nitric oxide	NO

This includes adaptation in vascular function due to hemodynamic changes, endothelial cell activation and inflammatory cell recruitment indicated by endothelial adhesion molecules and pro-inflammatory cytokines, smooth muscle damage which may lead to leakage of SM-specific proteins into circulation upon necrosis, and vascular remodelling which includes fibrosis and neovascularization.

Besides efforts to provide additional evidence for the qualification of biomarkers already supported by literature and expert knowledge [61,208,209], there have also been investigative efforts to identify potential novel candidate biomarkers which are ideally sensitive and specific, mechanistically linked to the pathogenesis and additionally found to precede injury and reflect lesion severity [209,210]. In this regard, prior research by Dalmas *et al.* [66] identified candidate tissue biomarkers based on Affymetrix GeneChip data from samples of mesenteric arteries of rats. These were treated with multiple vascular toxicants and comparator vasoactive but not vasotoxic compounds, and toxicologically relevant genes were identified based on the concordance of change with dose-responsive degrees of injury.

One advantage of this study was that it covered multiple structurally diverse compounds while most studies focus on few or single stressors [60,210–215]. Furthermore, data was derived from the endothelium and smooth-muscle enriched samples by laser capture microdissection and hence can capture changes in the vascular tissue which is not diluted by the neighbouring adventitia, connective tissue or lymph nodes. Besides identifying sensitive, specific and dose-responsive potential genomic biomarkers from gene expression data, a subset of the genes was confirmed by quantitative RT PCR (TaqMan) analysis in rats treated for 1 or 4 days with dopamine or fenoldopam where medial arterial necrosis (MAN) was present. For these genes, the absence of changes was further confirmed in rats treated with yohimbine, a vasoactive and non-vasotoxic compound which did not show any evidence of vascular injury in the mesentery [60,66]. Furthermore, it was found that some of the potential candidate biomarkers were regulated in mesenteric artery tissue scrape samples from rats treated with fenoldopam approximately 8 hours before histological detection of MAN providing further evidence that the genes are good potential candidates for detection of MAN in rats (Figure 1.3) [60,66]. However, as previously reported, it is not feasible nor practical to perform this extensive validation by TaqMan across a large set of genes or

compounds, and also the biomarker filtering itself might overlook valuable genes by focusing only on a small subset with stringent cut-off criteria [66].

In this study, a subset of the previously collected microarray gene expression and histopathology data [66] for selected compounds was further analysed as part of a collaboration between GSK and the University of Cambridge using an adapted bioinformatic approach to derive an extended set of potential genomic candidate biomarkers which were shown to detect and predict MAN in rats to provide additional starting points for further DIVI biomarker discovery and development. First, genes were identified with a stronger focus on consistency across treatments at the cost of lower effect size, while still requiring specific, and dose-dependent dysregulation. Further evidence was gained for these genes by investigating the behaviour across lesion severity. This approach enabled not only the characterization of genes which correlated with the histological presence of MAN, and injury progression but also the identification of gene expression changes in treatments where MAN is anticipated but not yet microscopically identified. Furthermore, candidate biomarkers were found to encode multiple secreted proteins which may translate to circulatory biomarkers. In addition, an interactive dashboard was developed using R/Shiny, in which results from this study for genes of interest beyond the 33 genes identified to be most promising candidates for potential biomarker development can be explored interactively. This can be accessed via `https://anikaliu.shinyapps.io/divi`.

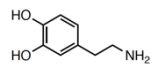
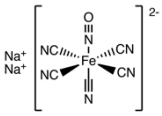
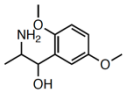
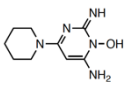
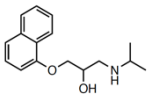
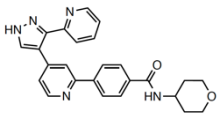
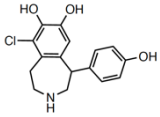
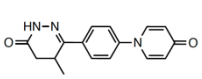
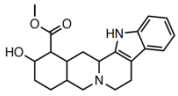
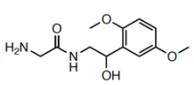
2.2 Methods

2.2.1 Study design

The data used in this work was provided by GlaxoSmithKline from selected compounds from previous experiments [66], in which male Crl:(CD)SD rats were given various known vasotoxic DIVI-inducing and vasoactive, non-vasotoxic compounds, at three different doses as well as corresponding and vehicle (see Table 2.2). Approximately 24 hours following the final dose of each treatment, rats were killed by CO_2 asphyxiation/exsanguination and necropsied as previously described.

Table 2.2: Summary of treatment conditions administered to rats analysed in this work.

The DIVI class separates DIVI-inducing compounds which show mesenteric medial arterial necrosis ("DIVI/MAN") from those which showed other histological changes in the mesenteric artery, such as perivascular and/or fibrinoid necrosis, perivascular fibrosis, EC hypertrophy and/or inflammatory cell infiltration ("Other VI"), and compounds which did not present any of the mentioned histological changes ("No DIVI"). The information was adapted from Dalmas *et al.* [66]

Compound	Structure	Vehicle	Doses (mg/kg/day)	Duration (Days)	Route of administration	Target	DIVI class
Dopamine		Sterile Water	1,30, 300	1, 4	oral	Dopamine receptor agonist	DIVI/ MAN
Sodium Nitroprusside		1% Methyl-cellulose	0.5,3,20	4	oral	cGMP agonist	Other VI
Hydralazine		1% Methyl-cellulose	1,10,50	4	oral	Unknown	Other VI
Methoxamine		1% Methyl-cellulose	0.1,1,10	4	oral	α1-Adrenoreceptor agonist	DIVI/ MAN
Minoxidil		1% Methyl-cellulose	1,50, 300	4	oral	K+ channel opener	Other VI
S-Propranolol		1% Methyl-cellulose	1,10,20	4	oral	β-Adrenoreceptor antagonist	No DIVI
GW788388		1% Methyl-cellulose	3,100, 1000	4	oral	ALK5 kinase inhibitor	Other VI
SKF-82526 (Fenoldopam)		0.9% NaCl	1,10, 100	1,4	subcutaneous	Dopamine receptor D1 selective agonist	DIVI/ MAN
SKF-95654		DMSO	0.5,5,50	4	subcutaneous	PDE3 inhibitor	DIVI/ MAN
Yohimbine		Sterile Water	0.5,5,20	4	oral	α2-Adrenoreceptor antagonist	No DIVI
Amphetamine		1% Methyl-cellulose	1,10,50	4	oral	α- and β-Adrenoreceptor agonist	No DIVI
Midodrine		1% Methyl-cellulose	1,10,50	4	oral	α1-Adrenoreceptor agonist	DIVI/ MAN

The in-life portion of the prior study from which data was obtained was conducted at Charles River Laboratories, Discovery and Development Services (CR-DDS), Argus Division, Horsham, PA, USA. All prior studies were conducted after review by the Charles River Laboratory (Discovery and Development Services, Argus Division, Horsham, PA, USA) Institutional Animal Care and Use Committee (IACUC) in accordance with the GSK Policy on the Care, Welfare and Treatment of Laboratory Animals and were in accordance with the Guide for the Care and Use of Laboratory Animals (NIH Publication, 25, No. 28, 16 August 1996). Endothelium or smooth muscle enriched samples were derived from cryosections of mesenteric arteries through laser capture microdissection, and gene expression levels were measured on Affymetrix Rat Genome 230 2.0 Arrays.

For histopathological analysis, mesentery from each animal was collected, processed, and histopathology was assessed using light microscopy by a single board-certified pathologist, and peer-reviewed by another board-certified pathologist as previously described [66]. Thereby, histopathological findings were documented using severity scores ranging from 0 (absence) to 4 (high severity). Histopathological evidence of MAN (severity score > 0), was identified for 5 out of 12 compounds, including fenoldopam, dopamine, midodrine, methoxamine, and SKF-95654 (see Table A.1). In contrast, no evidence of vascular injury was observed for yohimbine, S-propranolol, and amphetamine (i.e., no evidence in controls or test-article treated animals of MAN or perivascular and/or fibrinoid necrosis, perivascular fibrosis, endothelial cell hypertrophy or inflammatory cell infiltration).

2.2.2 Gene expression pre-processing

The raw gene expression levels for the provided samples (Table 2.2) were background corrected, $\log_2$ transformed, and quantile normalized with *RMA* (Robust Multi-array Average), accessed through the `affy` package, per study and cell type [216]. Array quality metrics describing distances between arrays, array intensity distributions and variance mean dependence were computed with the `ArrayQualityMetrics` package [217] and outliers detected based on these statistics were excluded from the downstream analysis. This resulted in overall 304 samples on smooth muscle gene expression, and 300 samples of endothelial gene expression across all treatments. The number of animals in each experimental condition is shown in Table A.1.

Batch effects were found between studies (each including treatment at multiple doses and its respective vehicle control), which are considered to be mainly due to technical variance. Technical variation may have resulted due to multiple experimental factors, such as the age of the rats at the time of treatment, the time of the day at which rats were dosed or necropsied, and differences in the vehicle or route of administration between compounds. While these batch effects are irrelevant in the case of within-study analysis, which was the sole focus in the prior work by Dalmas *et al.* [66], also trends across batches were analysed in this study. Therefore, a batch correction was performed using the *ComBat* function of the `sva` package [218] regarding the individual studies as batch covariates and using the implemented parametric empirical Bayes framework. The platform information for the Affymetrix Rat Genome 230 2.0 Array was derived from Gene Expression Omnibus [219] using GEO accession GPL1355, and used to aggregate probe IDs to rat gene symbols using the median for all probes uniquely mapping to one gene symbol. Where a probe matched multiple genes, the probe itself was kept avoiding dilution or duplication of the contained information, respectively.

2.2.3 Filtering of genes

To prioritize genes as potential transcriptomic biomarkers, an adapted approach to Dalmas *et al.* [66] was used, prioritizing consistency and specificity over effect size in DIVI conditions. To do so, DIVI conditions were first defined as those conditions in which MAN is observed for more than 20% of the animals as a trade-off between including early evidence of DIVI and excluding rare ones (Table A.1). The conditions in which MAN was observed are depicted in Figure 2.1 showing consistently increasing MAN frequency with increasing dose and MAN across all animals for SKF-95654 and midodrine at the highest dose.

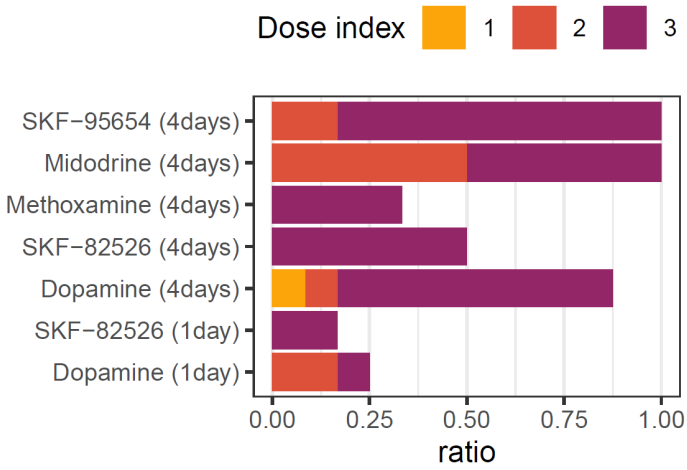


Figure 2.1: Frequency of Medial Arterial Necrosis (MAN) across conditions.

For all conditions in which MAN was found, the relative frequency is shown for each treatment, time and dose. Across all treatments, an increasing relative frequency of MAN is found with increasing dose.

The differential expression for each compound and dose group in comparison to the respective experiment-matched vehicle control was computed using gene-wise linear modelling and empirical Bayes moderated t-statistics provided by the `limma` R package [162]. Genes which significantly (p-value < 0.05) changed in the same direction across all DIVI conditions (further defined below) with a |logFC| > 0.7 were then identified as conserved genes. While a higher logFC threshold, e.g. |logFC| =1, is more common and was also used in our prior work [66], this lower logFC cut-off was chosen due to the fact that a gene would be filtered out if a logFC below this threshold was found in any DIVI condition. An alternative approach to balance potential outliers was previously used in our past work [66] requiring consistency across 8 of the 12 treatments with a |logFC| =1 (Table 2.3). In both pipelines, unadjusted p-values were used in the filtering pipeline, although p-values are commonly adjusted for differential expression analysis [220]. Here, it should be noted that only genes which are differentially expressed across all 7 DIVI conditions genes are considered as "discoveries" in this study which pass on to the downstream analysis, and that the likelihood for false positives in this case is not 0.05 but instead $(0.05)^7 < 10^{-9}$. Considering that statistical testing is only needed for the 48 (endothelium) and 75 (smooth muscle) genes, which pass the other logFC and directionality criteria, false positives are not expected and multiple testing correction was not implemented.

Next, genes were removed for which a significant change (p-value < 0.05) in the same direction was also found for any compound at any dose which did not show MAN or other evidence of vascular damage. This includes amphetamine, S-propranolol or yohimbine (Table A.1). As potential biomarker genes should reflect the dose-dependent increase in MAN frequency observed across all DIVI compounds, the Spearman rank correlation between gene expression levels of individual animals and the given compound doses, including the vehicle control as a dose of 0 mg/kg, was then computed and genes with an absolute Spearman correlation below 0.3 in any of the DIVI conditions were omitted.

Table 2.3: Comparison between filtering implemented by Dalmas *et al.* [66] and this study.

Vascular changes are defined as mesenteric medial arterial necrosis, perivascular and/or fibrinoid necrosis, perivascular fibrosis, EC hypertrophy and/or inflammatory cell infiltration.

Step		Dalmas *et al.* (2011)	This study
Pre-processing	Normalization	GCRMA	RMA
	Outlier removal	Affymetrix Statistical Algorithms Array Quality Metrics	ArrayQualityMetrics
	Batch correction for experimental batches of the same treatment	No batch correction due to analysis only within experiments	Batch correction using ComBat due to analysis across experiments
Definition of conditions	DIVI conditions	Compounds with MAN observed at any dose irrespective of other vascular changes	Compound-dose combinations with ≥ 20% MAN observed irrespective of other vascular changes
	Negative controls	No vascular changes at any dose	No vascular changes at any dose
Consistency	Fraction of DIVI treatments required	Significant in at least 8 of 12 compounds with histological evidence of MAN	Significant in all compound-dose combinations with histological evidence of MAN
	Criteria for differential expression	One-way ANOVA on dose followed by post-hoc contrast for linear trend ($p < 0.01$, $logFC \geq 1$)	Limma/eBayes per compound-dose combination ($p < 0.05$, $logFC \geq 0.7$)
Specificity	Fraction of negative controls required	Not significant in any compound at the highest dose	Not significant in any compound-dose combination
	Criteria for differential expression	One-way ANOVA on dose followed by post-hoc contrast for linear trend ($p < 0.01$, $logFC \geq 1$)	Limma/eBayes per compound-dose combination ($p < 0.05$)
Dose-responsive	Fraction of DIVI treatments required	Sign-consistent dose-response in at least 8 out of 12 compounds	Sign-consistent dose-response in all
	Criteria for dose-responsiveness	Fold change ≥ 1.7 in the two highest doses	Spearman correlation between expression level and dose with a magnitude of at least 0.3

To evaluate whether and how many genes are expected to pass the complete filtering procedure at random, a null distribution was generated using 1000 permutations of the compound labels. Only in 0.8% and 1% of permutations, any gene was identified in the endothelium or smooth muscle, respectively, and never were as many genes identified as for the true data indicating that genes are unlikely to pass the filtering procedure by chance (Figure A.1).

2.2.4 Development of an interactive dashboard

The results using the adapted analyses performed as part of this study were used to develop an interactive R/Shiny dashboard (`https://anikaliu.shinyapps.io/divi`) in R 4.1.2 [221] using the `shiny` and `tidyverse` [222,223] frameworks to enable exploration of the results beyond the 33 genes identified as most promising biomarker candidates. It should be noted, that results on a few genes currently being analysed internally or as part of other initiatives at GSK have been excluded as these data points are currently being analysed and are intended to be the subject of future publications. This includes but is not limited to data corresponding to the majority of the prioritized candidate biomarkers of the PSTC VIWG which are being analysed in conjunction with other datasets generated by the consortia as potential supporting evidence for a DIVI non-clinical rat biomarker qualification package.

2.2.5 Candidate biomarker predictivity

For each gene suggested as a candidate biomarker based on this analysis or in previous literature, the ability to predict the presence or absence of MAN was evaluated through the area under the receiver operating characteristic curve (AUC) using the `yardstick` [224] R package. Given that it was in most cases unclear whether an up- or downregulation is expected for literature markers, the higher AUC among both was reported expecting all biomarker candidates to perform better than random.

2.2.6 Biological annotation

Functional protein-protein associations between conserved genes in both tissues were derived from the STRING [172] database including all interactions with a combined score > 0.4, and the network was visualized in Cytoscape. For gene set analysis, gene sets from the Rat

Genome database [225], were combined with the canonical pathways (C2 CP) and hallmarks (H) gene set collections from MSigDB [226], which were mapped from human HGNC symbols to rat gene symbols with `biomaRt` [227]. In cases where the corresponding rat gene symbol was not represented as an individual gene, it was matched to shared probes if possible. Over-representation of identified genes in these pathway maps was analysed with the `clusterProfiler` [228] R package using the hypergeometric test statistic and all measured genes as background.

2.3 Results and Discussion

2.3.1 Identification of transcriptomic biomarker candidates for DIVI

To identify genes with the highest potential for biomarker discovery and development at the cost of losing many other promising genes, an adapted stringent filtering procedure was developed and applied to prioritize genes which show a consistent, specific and dose-dependent response (differences from the previous filtering pipeline by Dalmas *et al.* [66] are summarized in Table 2.3). The number of genes identified at each step is shown in Figure 2.2A**.**

Overall, 33 potential candidate genes with consistent, specific and dose-dependent changes were identified that correlate with the presence of DIVI; 27 in the smooth muscle- and 10 in the endothelium-enriched samples. Expression changes for each potential candidate gene across all compounds are shown in Figure A.2 and Figure A.3. For many of the genes, significant changes in expression were frequently identified at lower doses than those for which histopathological evidence of MAN was observed, thus indicating that these changes may precede and predict the occurrence of MAN. This is further supported by the fact that many of the genes identified following 4 daily doses of fenoldopam (SKF-82526), one of the conditions in which MAN was observed, were also observed to be regulated 24 hours following a single dose prior to histopathological evidence of MAN.

While all genes passing the filtering criteria show consistent changes across DIVI conditions, the magnitude of expression change varies and is summarized in Figure 2.2B**.** Overall, the strongest up-regulated gene, as noted by median logFC, in rats with histologic evidence of MAN (i.e. DIVI conditions) in both tissues was found for Tenascin C (*Tnc)*, a gene encoding

the glycoprotein Tnc known to be involved in blood vessel injury [229] while the strongest down-regulation in gene expression was observed for *Chchd10*, encoding mitochondrial Coiled-coil-helix-coiled-coil-helix domain-containing protein 10 (Figure 2.2B) highlighting that these genes might be easily detectable. For all potential candidate genes identified, consistent directionality of dysregulation in both tissues was observed.

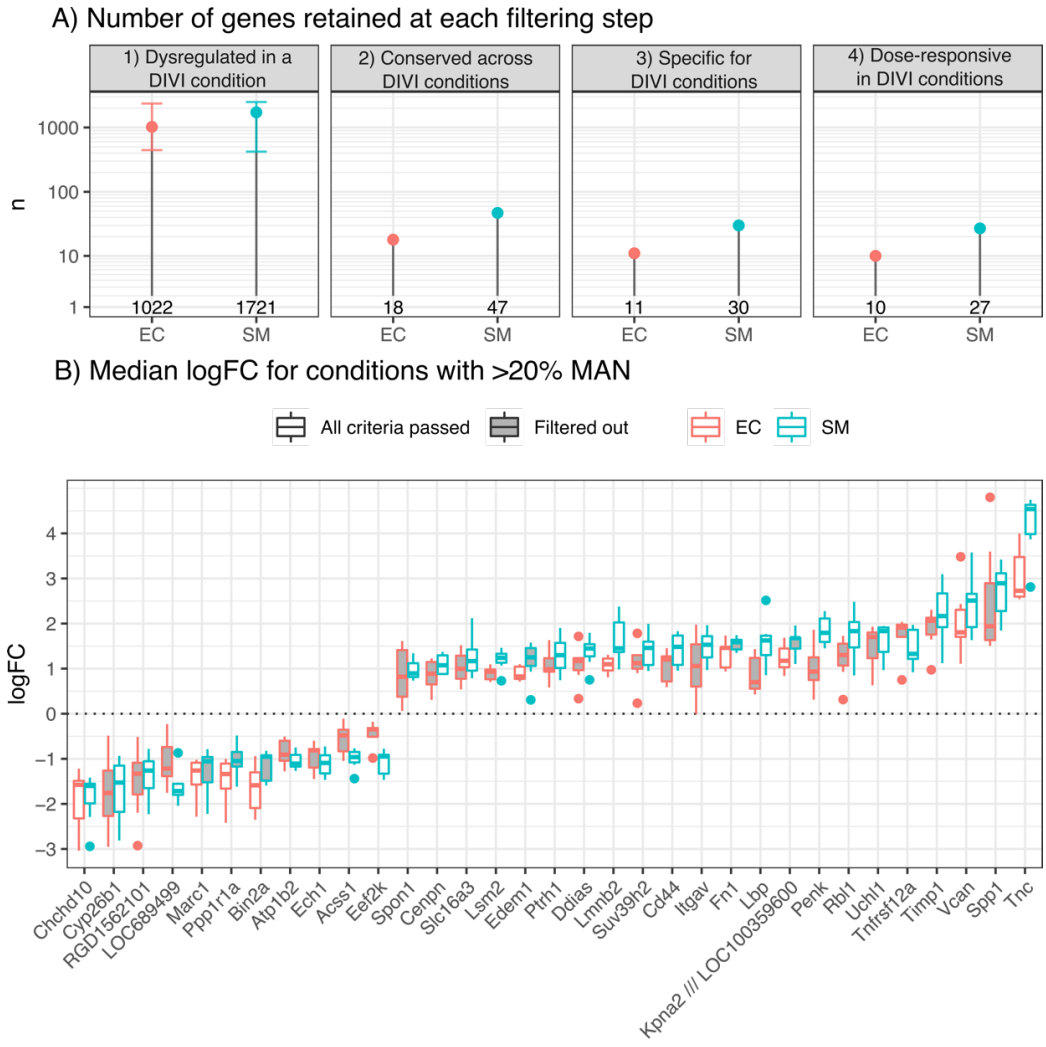


Figure 2.2: Potential biomarker candidates for MAN identified through filtering criteria.

A) Potential transcriptomic biomarker candidates for medial arterial necrosis (MAN) in endothelial cells (EC) and smooth muscle cells (SM) were identified through multiple filtering criteria, and the number of genes is shown for each tissue at each stage. First, differentially expressed genes were identified for each condition with >20% MAN. In subsequent steps, only genes with conserved significant differential expression across all DIVI conditions in the same direction, without significant differential expression in any negative control condition and dose-dependent expression changes (|Spearman correlation| ≥ 0.3) were kept. B) Distribution of logFCs for differentially expressed potential candidate genes, identified by the filtering procedure in conditions with >20% MAN.

The list of candidate genes identified in this study was then compared to the ones from the previous work [66], and 6 out of 33 biomarker candidates were found to overlap with the previously proposed 57 genes (Table 2.4). Out of the six genes that were noted to be similarly regulated in a dose-responsive manner across the analysis pipelines, tissue inhibitor of metallopeptidase 1 (*Timp1)*, fibronectin 1 (*Fn1*)*,* karyopherin subunit α2 (*Kpna2*) and versican (*Vcan*) have been previously confirmed to be regulated using quantitative RT-PCR (TaqMan), as further outlined in Table 2.4.

Gene Symbol	Gene name	Tissue	Literature	TaqMan
Up-regulated genes				
Cd44	Cd44 molecule	SM	-	-
Cenpn	centromere protein N	SM	-	-
Ddias	DNA damage-induced apoptosis suppressor	SM	-	-
Edem1	ER degradation enhancer, mannosidase alpha-like 1	EC	-	-
Fn1	fibronectin 1	EC	Dalmas (SM|EC)	Up
Itgav	integrin αV	SM	-	-
Kpna2 (LOC100359600)	karyopherin α2 /// karyopherin α2-like	EC	Dalmas (SM), VIWG	Up
Lbp	lipopolysaccharide binding protein	SM	-	-
Lmnb2	lamin B2	SM, EC	-	-
Lsm2	LSM2 homolog, U6 small nuclear RNA associated (S. cerevisiae)	SM	-	-
Penk	proenkephalin	SM	-	-
Ptrh1	peptidyl-tRNA hydrolase 1 homolog (S. cerevisiae)	SM	Dalmas (SM)	Ct values were undetermined for all samples
Rbl1	retinoblastoma-like 1 (p107)	SM	-	-
Slc16a3	solute carrier family 16 (monocarboxylate transporter), member 3	SM	-	-
Spon1	spondin 1, extracellular matrix protein	SM	-	-
Spp1	secreted phosphoprotein 1	SM	-	-
Suv39h2	suppressor of variegation 3-9 homolog 2 (Drosophila)	SM	-	-
Timp1	TIMP metallopeptidase inhibitor 1	SM	Dalmas (SM|EC), VIWG	Up
Tnc	tenascin C	SM, EC	Dalmas (SM)	Partial (Ct values were undetermined for fenoldopam)
Tnfrsf12a	tumor necrosis factor receptor superfamily, member 12a	SM	-	-
Uchl1	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)	SM	-	-
Vcan	versican	SM, EC	Dalmas (EC)	Up

Down-regulated genes				
Acss1	acyl-CoA synthetase short-chain family member 1	SM	-	-
Atp1b2	ATPase, Na+/K+ transporting, beta 2 polypeptide	SM	-	-
Chchd10	coiled-coil-helix-coiled-coil-helix domain containing 10	SM, EC	-	-
Cyp26b1	cytochrome P450, family 26, subfamily b, polypeptide 1	SM	-	-
Ech1	enoyl CoA hydratase 1, peroxisomal	SM	-	-
Marc1	mitochondrial amidoxime reducing component 1	EC	-	-
Ppp1r1a	protein phosphatase 1, regulatory (inhibitor) subunit 1A	EC	-	-
RGD1562101	similar to very large G-protein coupled receptor 1	SM	-	-
Eef2k	eukaryotic elongation factor-2 kinase	SM	-	-
Bin2a	beta-galactosidase-like protein	EC	-	-
LOC689499	similar to Y97E10AL.1	SM	-	-

Table 2.4: Genes with conserved and specific dysregulation in DIVI.

The identified genes in smooth muscle (SM) and endothelial cells (EC) are shown, including information on whether this was also by Dalmas *et al.* [66] or prioritized by VIWG [207]. Furthermore, previous quantitative RT-PCR (TaqMan) results are indicated in which expression in enriched individual tissue types and/or tissue scrapes from the mesentery of rats treated for 4 days with dopamine (300 mg/kg/day) and fenoldopam (100 mg/kg/day), as well as of rats treated with yohimbine (20 mg/kg/day) as a negative control, was measured.

While none of the results was disproved, results for tenascin C (*Tnc*) were undetermined for fenoldopam and undetermined across all treatments for peptidyl-tRNA hydrolase 1 homolog (*Ptrh1*). Furthermore, three of these genes were already prioritized by the PSTC VIWG (Table 2.4), namely *Timp1*, which was first suggested as a potential vascular injury biomarker by Dagues *et al.* [230], *Fn1* and *Tnc*. Thus, the adapted analysis workflow is able to recover previously confirmed results, as well as able to identify and prioritizes new genes as potential candidate biomarkers.

The differences in the number of genes identified between the current study and prior work by Dalmas *et al.* [66] can partially be explained by differences in filtering criteria (Table 2.3), as well as the fact that analyses in this study were performed on a subset of compounds. On a more general level, both filtering pipelines aimed at prioritizing a short list of most promising genes and do not claim that other genes are not unsuitable as biomarker candidates. If new biomarker candidates are proposed in the future, it may be of interest to understand across

which DIVI conditions expression changes, indicative of MAN, are observed. Hence, a R/Shiny dashboard was developed, which will be further introduced in 2.3.4, using which the behaviour of a gene of interest can be inspected.

2.3.2 Characterization of biomarker candidate gene properties

To evaluate the ability of the potential biomarker candidate expression levels to separate animals with and without MAN, the area under the ROC curve (AUC) was computed for each candidate biomarker prioritized in this study and previous work by Dalmas *et al.* [66]. This further ranks the individual genes with respect to their predictivity observed in the current dataset and overall identifies correlated performance in both tissues indicating that the derived performance should be representable for the vascular tissue (Figure 2.3).

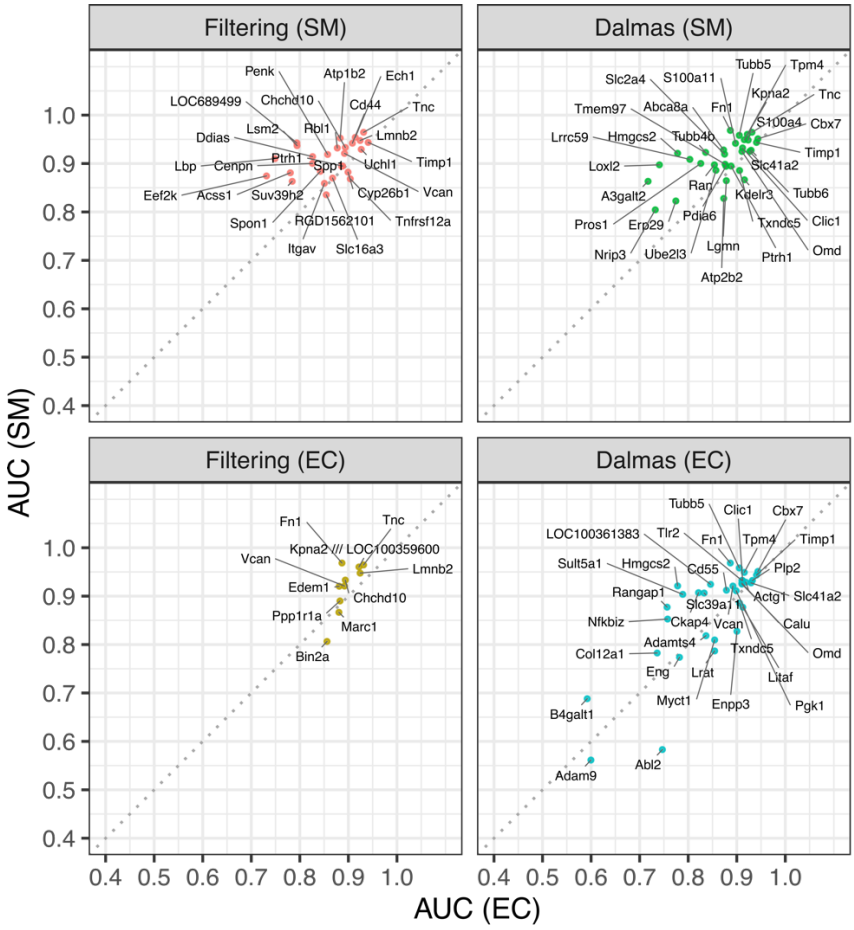


Figure 2.3: Ability of biomarker candidate to separate animals with and without MAN.

For each potential biomarker candidate gene, the ability to separate animals with and without evidence of medial arterial necrosis (MAN) based on expression in the endothelium (EC) or smooth muscle (SM) is shown as AUC. Overall, genes identified in this study and prior work by Dalmas *et al.* [66] were found to achieve high AUCs.

Among the genes identified as potential biomarker candidates for the prediction of MAN in rats in this study, the highest AUCs are observed for *Fn1* in the smooth muscle (AUC of 0.97) and *Timp1* in the endothelium (0.94), while the lowest AUCs are observed for *Bin2a* (0.81) and *Eef2k* (0.73) in smooth muscle and endothelium, respectively. As expected, there was an overall high predictive performance observed in both endothelium (AUC between 0.59-0.94) and smooth muscle (AUC between 0.56-0.97) also for the genes previously identified in the prior study [66], In this regard, *Cbx7* should be highlighted which reaches an even higher performance in the endothelium (0.94) than genes prioritized in this study as well as high performance in the smooth muscle (0.95). However, it did not show sufficiently strong dysregulation across DIVI conditions and was hence not included. In contrast, the lowest performance was found for ADAM metallopeptidase domain 9 (Adam9) with an AUC of 0.60 in the endothelium and 0.56 in the smooth muscle.

As next step, the association between expression change and lesion severity was analysed to see which markers reflect disease progression and which ones are already detectable during early pathogenesis which are desired properties for safety biomarkers [208,209] (Figure 2.4). Therefore, groups of animals were defined based on the observed severity score for MAN and the expression levels in these samples were compared to those from animals only treated with corresponding vehicle control. Animals without evidence of MAN (severity score of 0) were additionally separated into animals treated with corresponding negative control which should not show changes in marker expression if these genes are predictive of MAN in rats, and animals in DIVI conditions which may show changes in gene expression despite the absence of morphological changes, as well as other conditions.

The results of this analysis are shown in Figure 2.4 and strong and largely significant changes were identified across lesions of all severities as well as animals in DIVI conditions without MAN, while significant changes are not observed for negative control or other DIVI unrelated treatments. This indicates that changes on the transcriptomic level for selected genes might not only correlate with the presence of injury but can potentially also be predictive for impending vascular damage. While some potential candidate genes remain largely constant across all lesion severities and animals in DIVI conditions without MAN, such as lamin B2 (*Lmnb2)* or beta-galactosidase *(Bin2a)*, others show an increasing change with increasing severity scores potentially reflecting vascular injury progression, e.g., ER degradation-

enhancing alpha-mannosidase-like protein 1 (*Edem1)*, *Timp1*, TNF receptor superfamily member 12A (*Tnfrsf12a)* or cytochrome P450 family 26 subfamily B member 1 (*Cyp26b1).* However, comparisons across lesion severity should be treated with caution due to the low number of animals found with severe MAN (severity score of 3 or higher), and more generally because this analysis was specific for MAN and other vascular changes were not included (Table A.2).

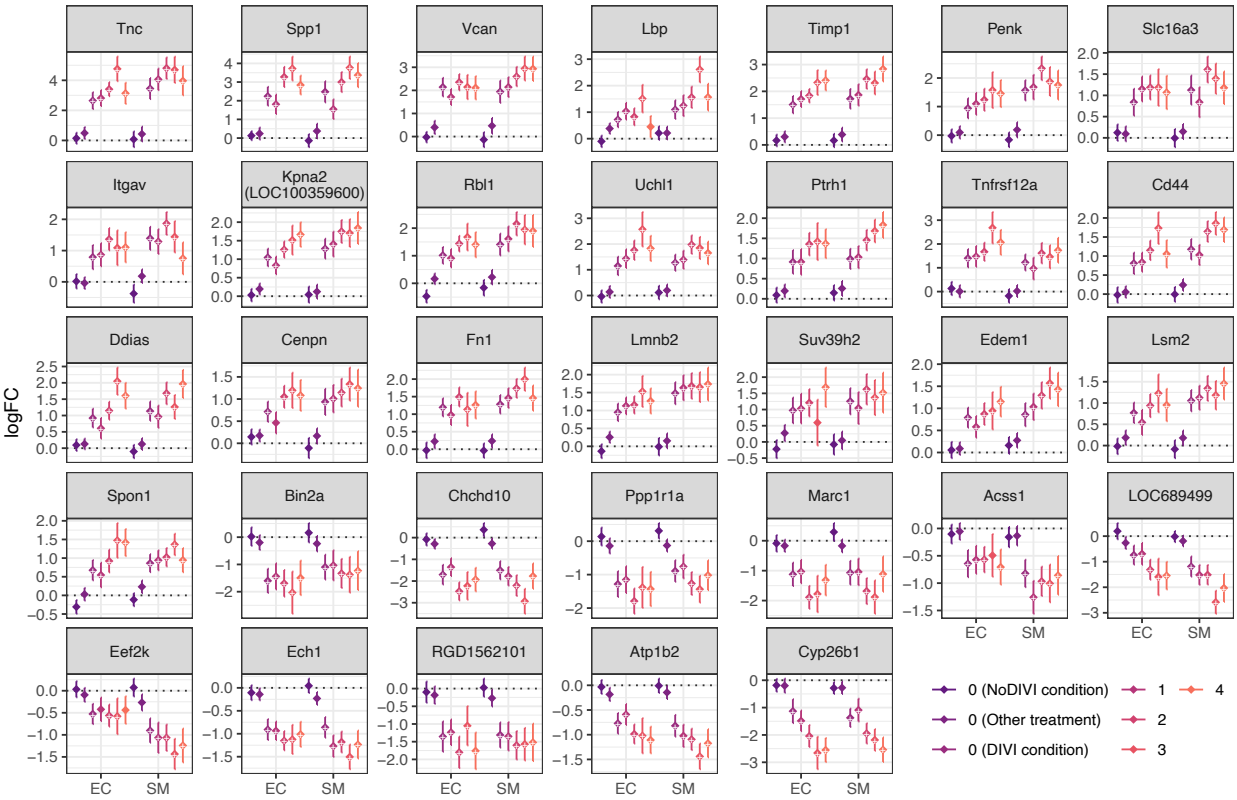


Figure 2.4: Change in marker gene expression in Medial Arterial Necrosis (MAN) severity groups in comparison to vehicle controls.

Animals which received treatment were separated into groups based on the observed histopathology. Animals with MAN were grouped by the observed severity, and animals without MAN were divided based on the respective experimental condition: Negative control treatments ("NoDIVI condition"), conditions with >20% MAN ("DIVI condition") and others which showed morphological changes in the vasculature but not MAN ("Other treatments"). For each group, significant differential expression in comparison to vehicle control (FDR < 0.05) is indicated by "+" and the logFC incl. 95% confidence interval is shown. This identifies significant dysregulation across all biomarker candidates already for animals in DIVI conditions which, however, don't show any signs of MAN yet. Furthermore, severity-dependent differential expression magnitude is identified for some genes, including *Timp1* and *Cyp26b.*

2.3.3 Biological context of biomarker candidate genes

To better understand the potential biological role of the identified genes, protein-protein interactions between proteins encoded by genes with conserved dysregulation across DIVI conditions in both tissues were derived next. Overall, more interactions than expected at random (PPI enrichment p-value [172] $< 10^{-16}$) were identified, as well as two clusters of functionally associated genes (Figure 2.5 and Table A.3). The bigger cluster includes largely ECM-related proteins including extracellular proteins Fn1, Timp1, Tnc and Vcan which are detected in both tissues and central nodes in the cluster. Cellular receptors in the cluster linked to those proteins include the integrins αV (Itgav) and α5 (Itga5), which interact with Fn1 and are involved in fibronectin matrix formation and cardiovascular development [231].

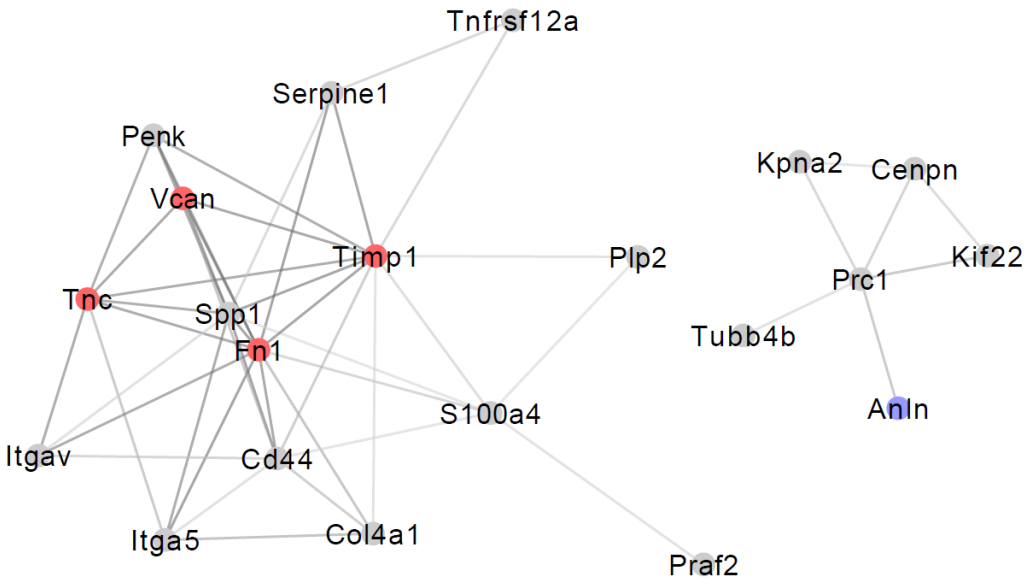


Figure 2.5: Functional protein-protein associations between conserved genes.

Interactions from the STRING database [172] with medium confidence (confidence score>0.4) between conserved genes were derived. More associations are found than expected at random (13 interactions expected, PPI enrichment p-value [172] $< 10^{-16}$). Genes identified in both tissues (labelled in red) are found at the core of the bigger cluster. Only one of the genes identified only in the endothelium, *Anln* labelled in blue, showed functional associations, while the majority of genes and interactions are found in the smooth muscle (grey).

Furthermore, secreted phosphoprotein 1 (Spp1) was identified which is increased in multiple vascular diseases [232] and is known to interact with both reported integrins as well as the hyaluronic acid receptor Cd44. The second cluster contains tubulin beta-4B chain (Tubb4b) and the kinesin-like protein (Kif22), which are related to microtubule-based transport, anillin (Anln) which is involved in cytokinesis, the centromere protein N (Cenpn) and Kpna2. All of these are linked to the protein regulator of cytokinesis 1 (Prc1), the central node in this cluster, pointing to vascular hyperplasia. Hence, plausible associations between proteins encoded by candidate genes were identified shedding further light on their potential mechanistic role.

An over-representation analysis was then performed for each tissue using conserved genes or only biomarker candidate genes, respectively. From this analysis, three pathways in the endothelium and twenty-two in the smooth muscle were identified, which can be explained by the higher number of genes observed in the smooth muscle (Figure 2.6). In both tissues, epithelial-mesenchymal transition (EMT) is over-represented which is a key process in tissue repair and fibrosis during which epithelial cells switch to a mesenchymal phenotype which is, amongst others, characterized by the expression of *Fn1* and shows strongly modulated interactions with the ECM [233]. While epithelial cells are not a key component in the vasculature, EMT is closely related to endothelial-mesenchymal transition (EndMT) which has been previously observed in various cardiovascular diseases [75].

A subset of the genes involved in EMT is additionally linked to ECM proteoglycans which are the second gene set overrepresented in both tissues. These are generally known to be upregulated in early vascular lesions and take in multiple key roles in vascular injury, such as mechanotransduction, regulation of leukocyte invasion and inflammation, control over blot clotting, ECM organization as well as vascular calcification [234]. The third shared over-represented pathway is related to the ECM and is termed “Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)”. However, to note, the identified extracellular proteins in this pathway are not directly related to IGF but phosphorylated by the same kinase as multiple IGFBPs, namely the extracellular serine/threonine protein kinase FAM20C, which is responsible for the majority of extracellular phosphorylations [235].

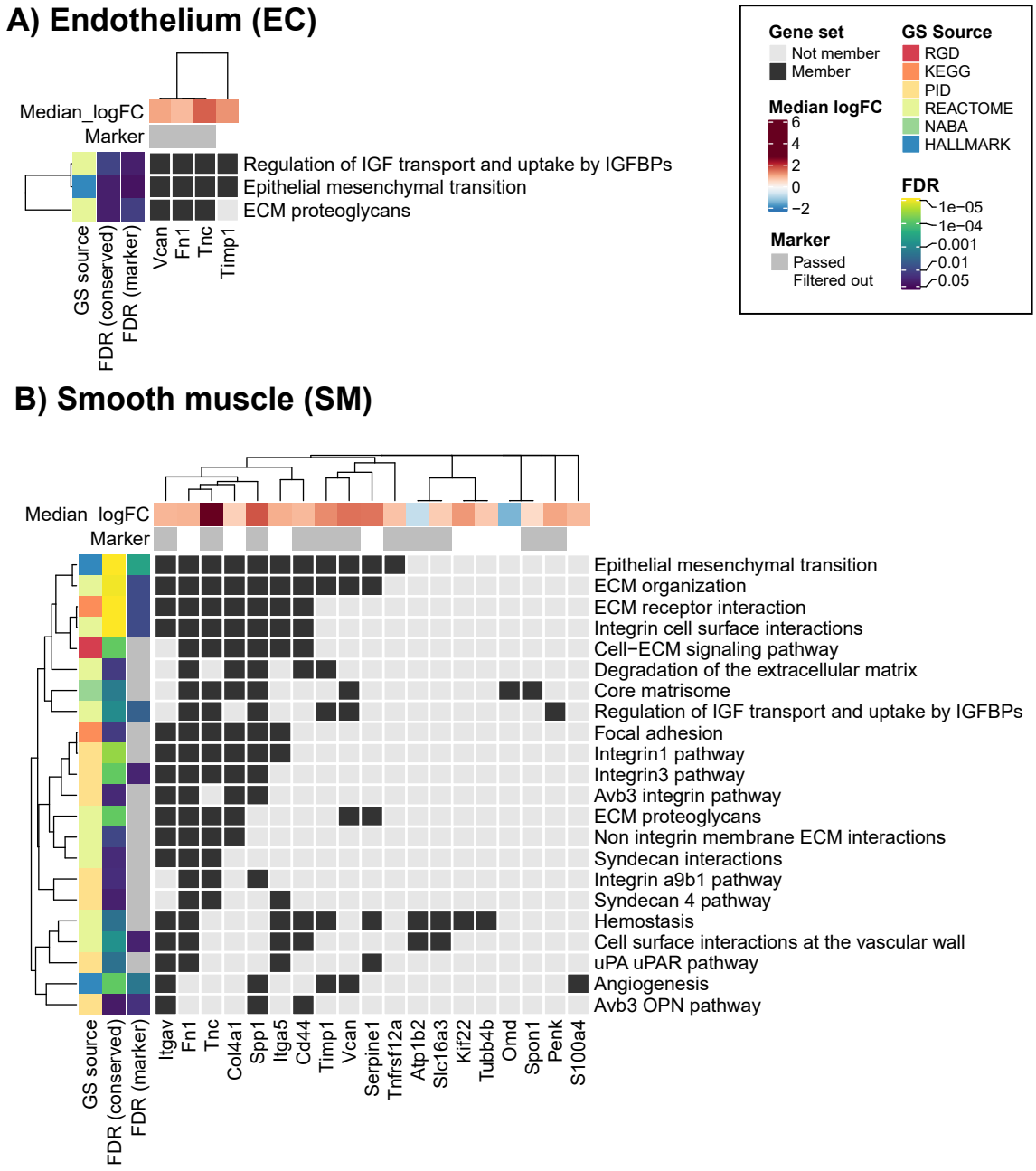


Figure 2.6: Enriched pathways in conserved and potential marker genes.

The gene set membership is shown for conserved genes and significantly enriched pathways (FDR<0.05) in endothelium (A) and smooth muscle (B), respectively. The median logFC across conditions with >20% MAN is shown for all genes and marker genes are additionally highlighted. For the genesets, FDR among conserved and marker genes is shown to identify genesets only enriched among conserved genes. Moreover, the geneset source is annotated with all genesets except the ones from Rat Genome Database being derived through MSigDB.

In the smooth muscle, multiple integrin-related pathways, such as focal adhesions, were identified which are related to the two RGD-motif binding integrins, integrin αV (Itgav) and integrin α5 (Itga5), as well as the extracellular proteins Col4a1 (arresten), Fn1, Tnc and Spp1 which interact with integrins on the membrane surface. Additionally, Cd44 is included in the "Integrin cell surface interactions" gene set, potentially due to the known crosstalk with ostepontin (Spp1)-induced signalling described in the Avb3 OPN pathway. Additional extracellular proteins classified as core matrisome by Naba *et al.* [236] are spondin 1 (Spon1) and Osteomodulin (Omd). Also, cell surface interactions at the vascular wall are identified, which indicate leukocyte extravasation, and are linked to the two integrins and their interaction partners Fn1 and Cd44 as well as the surface proteins Sodium/potassium-transporting ATPase subunit beta-2 (Atp1b2) and the monocarboxylate transporter 4 (Slc16a3) which both interact with the extracellular matrix metalloproteinase inducer Basigin. The over-represented parent process haemostasis additionally includes Tubb4b, Kif22 and the plasminogen activator inhibitor 1 (Serpine1), which is involved in the controlled degradation of blood clots via the Urokinase-type plasminogen activator (uPA) and uPAR-mediated signalling. This analysis hence provides further insight into the potential mechanistic interplay between the identified conserved and potential candidate marker genes and indicates potential biological pathways leading to the development and progression of MAN. In particular, changes on the cell surface and interaction with the ECM are highlighted, which are known processes involved in vascular injury and remodelling [237].

2.3.4 DIVI gene expression dashboard

As part of this work, a publicly available R/Shiny dashboard was developed (`https://anikaliu.shinyapps.io/divi`) using which the gene-level results derived in this study can be visualized and explored in four tabs. Since *Timp1* was identified as one of the most promising biomarker candidates, a potential follow-up question may be whether other tissue inhibitors of metalloproteinases show similar trends given their shared evolutionary history [238,239] although they did not pass the stringent filtering pipeline. In this section, the relevance of tissue inhibitors of metalloproteinases *Timp1*, *Timp2 Timp3* and *Timp4* for DIVI is hence investigated to demonstrate the functionality of the dashboard. In the dashboard, genes of interest can be selected from a dropdown list, and all figures shown in this section have been generated in the dashboard.

The first tab gives an overview of the number of DIVI conditions (conditions with ≥ 20% MAN observed) for which differential expression is found and the median logFC observed across these conditions (Figure 2.7). Significant up-regulation of *Timp1* was found in all seven DIVI conditions, and down-regulation of *Timp4* in six DIVI conditions in the endothelium and smooth muscle. In contrast, differential expression of *Timp2* and *Timp3* was found in fewer conditions and with a lower magnitude. It is also possible to select areas of interest in the plots to identify genes of interest in a hypothesis-free manner. Differential expression statistics for these genes are then summarised in a data table below.

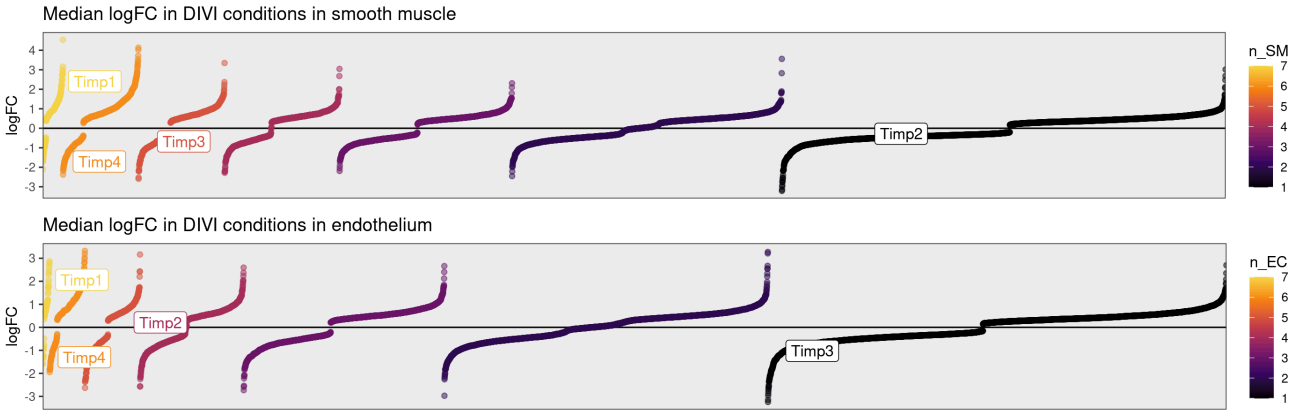


Figure 2.7: Summary of differential expression across DIVI conditions.

Genes are summarised based on the number of DIVI conditions with significant differential expression, also indicated as colour, and the median logFC observed in these conditions. User-selected genes are automatically labelled.

In the second tab, the differential expression results across all treatments, doses, and tissue layers can be visualized for selected genes providing a more detailed view of expression changes in DIVI conditions, negative control conditions, and other treatments (Group definitions are described in Table 2.3). The differential expression plot for the smooth muscle is shown in Figure 2.8, and only *Timp2* showed significant differential expression indicating that expression changes across the other genes were specific for MAN. For *Timp1*, expression changes were found in all DIVI conditions while the only DIVI condition without changes for *Timp4* was a single-dose treatment of dopamine. Furthermore, *Timp1*, *Timp3* and *Timp4* showed significant expression changes in SKF-82526 (fenoldopam) already after single-dose treatment although only repeat-dosing for 4 days was considered adverse. It should be noted, that unadjusted p-values are shown as definition for statistical significance (p-value < 0.05) here to align with the criteria implemented in the biomarker filtering pipeline.

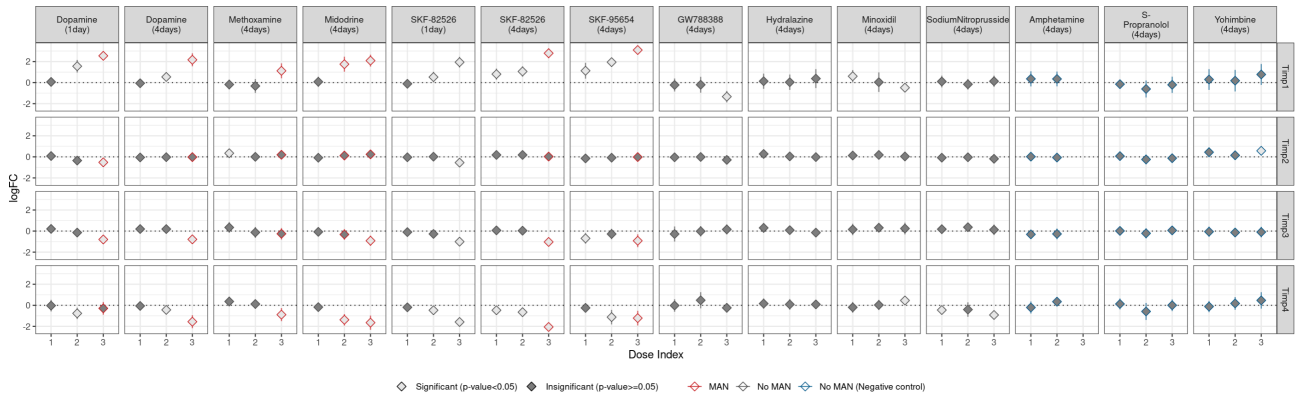


Figure 2.8: Differential expression by treatment condition in the smooth muscle.

For each user-selected gene and each compound-time combination, the differential expression across the tested doses is shown.

In the third tab, the differential expression results across histopathology groups are shown (Figure 2.9) using the same group definitions as in Figure 2.4. This shows that significant differential expression in comparison to animals treated with vehicle control was found for all groups with MAN for *Timp1*, *Timp3,* and *Timp4*. Furthermore, samples from animals in DIVI conditions but without MAN show differential expression *Timp1* and *Timp4* suggesting that these may indicate injury before it manifests.

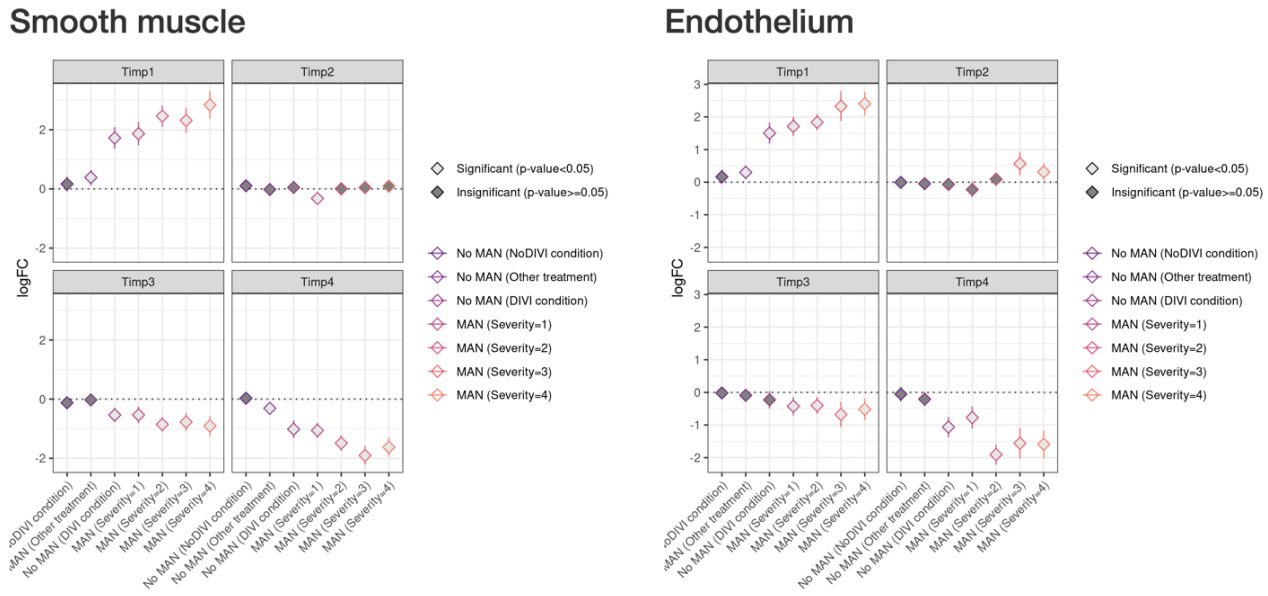


Figure 2.9: Differential expression by histopathological class.

For each user-selected gene and histopathological class, the differential expression compared to samples from animals tested with vehicle control is shown.

In the final tab, the predictive performance across all genes is shown as previously described in Figure 2.3. AUCs can be summarised as a data table for genes in interactively selected areas of interest, and selected genes of interest are automatically labelled and highlighted in the plot (Figure 2.10). This again highlights that the best predictive performance is found for *Timp1* followed by *Timp4*, while a lower predictive performance is found for *Timp3* and the lowest one for *Timp2*.

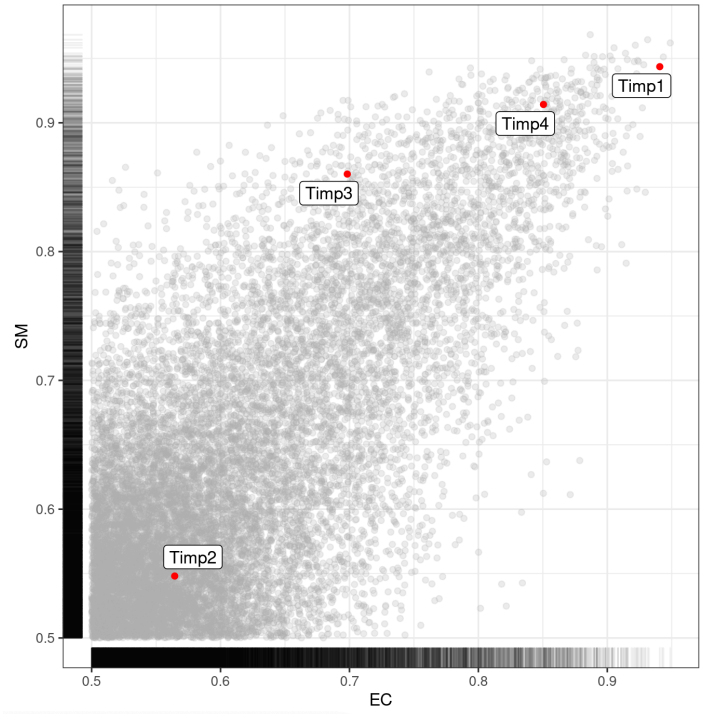


Figure 2.10: AUC distribution across measured genes across endothelium (EC) and smooth muscle (SM).

Overall, the four tabs hence provide different views on the gene expression data analysed in this study. While the first and last tabs enable a global comparison across all measured genes, the second and third tabs provide more details on when differential expression is observed. As expected, the best biomarker properties are observed for *Timp1*, which was also identified as a promising biomarker candidate in the previous analysis. However, interestingly *Timp4* showed good biomarker properties with dysregulation in opposite directionality. This demonstrates that the developed R/Shiny dashboard presents a valuable addition to this work using which also other genes with good biomarker properties can be identified, which may be supported by other sources of evidence.

2.4 Conclusion

In this study, potential genomic biomarker candidates for MAN, a histological indicator of DIVI, in rats were identified using Affymetrix GeneChip data obtained from smooth muscle- and endothelium-enriched mesenteric artery samples and corresponding histopathology data from rats treated with selected compounds evaluated in previous work, i.e. 12 compounds including vasotoxic compounds known to elicit MAN and vasoactive non-vasotoxic comparator compounds at multiple doses including vehicle controls [66]. In comparison to previous work by Dalmas *et al.* [66], an adapted bioinformatic gene filtering pipeline was used, prioritizing consistency and specificity over a larger effect size. As before, the biomarker candidate's expression was required to be dose-responsive as a proxy on whether the marker reflects increasing injury. In this study, however, also gene expression changes across lesion severity were characterised as a surrogate for injury progression and revealed that animals in DIVI conditions which do not yet show histopathological evidence of MAN show changes in expression suggesting that the identified biomarker candidates might predict the occurrence of DIVI, specifically MAN.

From a biological perspective, genes encoding proteins involved in ECM interactions were identified to be significantly enriched among the shortlisted candidate genes, and also *Tnc*, *Timp1, Spp1* and *Fn1* encoding secreted proteins were found to show the highest up-regulation and predictivity. Additionally, it was identified that these genes, as well as the integrins *Itgav* and *Itga5,* which are interaction partners on the cell surface, are highly connected through protein-protein associations further indicating a joint mechanistic role. It should also be highlighted that genes which encode secreted proteins have higher chances of translating to circulating biomarkers [61] and could hence potentially be detected directly from plasma or serum in a non-invasive manner.

Overall, hence not only an extended list of promising candidate biomarkers for DIVI based on sensitivity, specificity and dose-response is provided but also additional supporting evidence derived by analysing the genes' ability to reflect lesion severity and their potential mechanistic role in pathogenesis. Given the continued unmet need for DIVI biomarkers, the potential genomic biomarker candidates for MAN and DIVI identified in this work, and in particular, those encoding secreted proteins, provide valuable data-driven starting points

for biomarker discovery[66]. Although additional follow-up work is needed, including confirmation of gene expression changes in tissue from mesentery samples of rats with MAN as well as further investigation of injury initiation and progression using time course studies, the results in this study, offer potential new avenues to investigate potential translatable biomarkers of DIVI, mainly MAN in rats.

3 Time-concordant event cascades in Drug-Induced Liver Injury (DILI)

This work was previously published as a research article in PLOS Computational Biology [240]. I'd also like to acknowledge Dr Jürgen Pahle and Prof. Julio Saez-Rodriguez, as my previous research experiences with them in time-series modelling and causal reasoning have inspired this work.

3.1 Introduction

Adverse drug reactions are a major reason for compound failure in clinical trials [2,16] and a significant cause for post-marketing withdrawals. To counter exposing patients to these risks, it is desired to identify adverse events earlier in the individual patient but also in the drug development process (1.1.2). Mechanistic understanding of how adverse event pathogenesis is crucial in this regard, i.e. to derive early safety biomarkers or *in vitro* assays. However, current understanding of toxicity is largely incomplete, in particular for complex phenotypes such as organ injury which can usually be caused by a wide range of compounds perturbing the biological system at different points mediated through multiple biological scales and entities [26,27].

Biological readouts such as transcriptomics are particularly suited to study such intermediate key events as they provide broad insights into cellular changes, e.g. in contrast to target profiling, which can then lead to the identification of predictive signatures and mechanistically relevant insights. This is for example true in the context of DILI [241–245], which is a major cause for attrition in drug development and accounts for around half of the cases of acute liver failure in the US and European countries [246,247]. In this regard, in particular time-series data is interesting as it is able to trace the dynamic effects throughout pathogenesis. Previous studies focussed on the time (and dose) dependence of gene expression-derived events in the context of adverse findings [248,249], so the changes of individual events across changes in time (and dose), and also aimed to predict later adverse findings from fixed early timepoints [241,242]. From a mechanistic perspective, however, neither activation at a certain timepoint nor a certain progression over time is mandatory, but only time concordance, so activation of the key event before the downstream adverse effects.

In this study, time concordance across gene expression-derived cellular events and adverse events based on histopathology was hence quantified across a wide range of compounds. To do so, the concept of “first activation” is introduced for mechanistic analysis, which focuses only on the earliest timepoint an event can be reliably detected and then orders events within a time-series by their timepoint of first activation (Figure 3.1A).

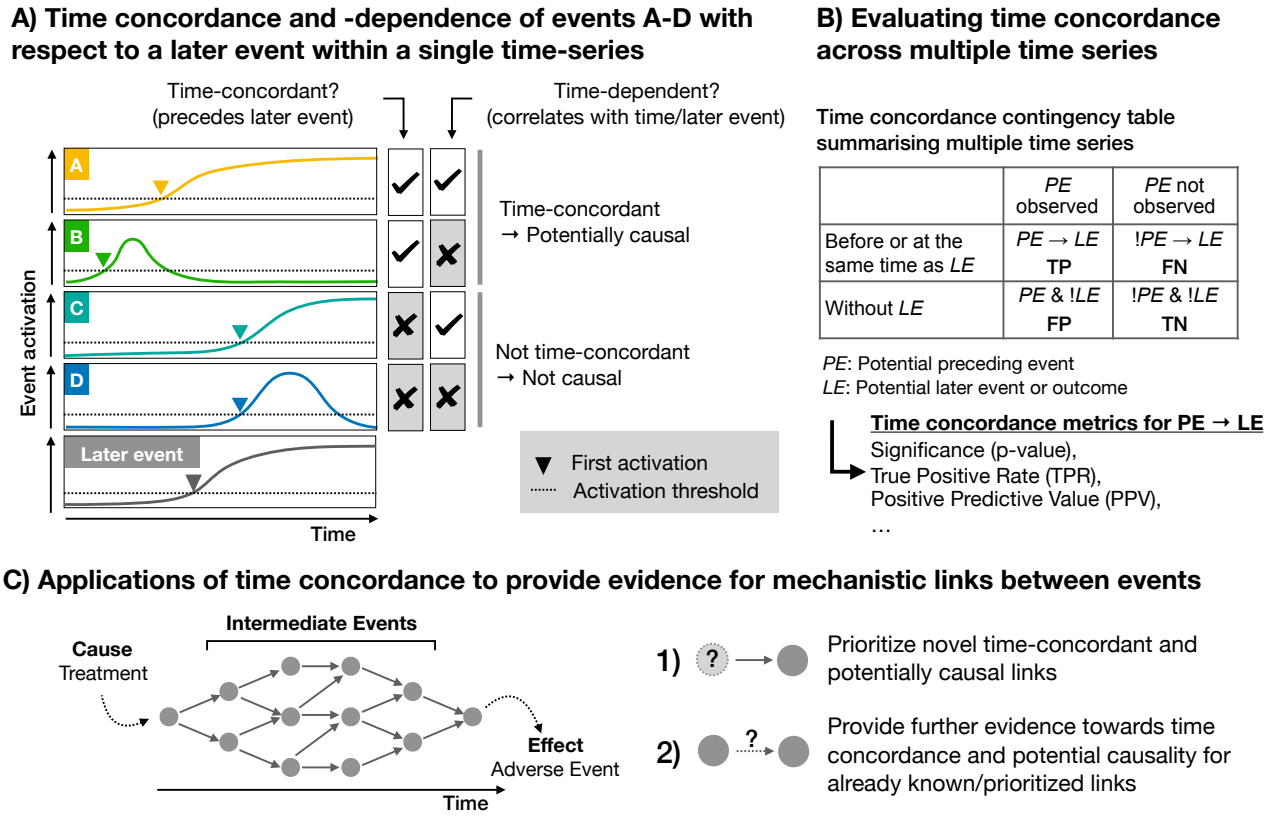


Figure 3.1: Quantifying time concordance based on first activation.

(A) The event activation of the events A-D and the later event is shown over time, as well as their timepoint of first activation, at which the event first passes the defined activation criteria. If an event takes place before a defined later event, which in our study is adverse histopathology, it is time-concordant. Time concordance indicates that there is potentially a causal relation between both events, and this is distinct from time-dependence which is defined based on the correlation to the later event or time. (B) Based on the frequency of an event before or at the same time as the later event and its frequency in background time-series without the later event, a confusion matrix and different time concordance metrics can be derived. (C) Time concordance can both prioritize potentially novel links and provide further evidence on potential mechanistic links between events. Events are indicated as nodes and mechanistic links between them as edges.

In contrast to previous time concordance analyses in AOPs which addressed a defined set of KER and known KE [188,192,193], this analysis derives statistical evidence for temporal concordance across time-series and can do so for any combination of events based on gene expression or histopathology. Although the confidence of these temporal orders per time-series is limited by the noisiness of gene expression data and the low time resolution, statistical significance can be evaluated across time-series and relations are only considered to be time-concordant if the preceding event is significantly more frequently found before the later event than in time-series without (Figure 3.1B). Furthermore, this also allows us to separate out events which depict general perturbation responses but are unspecific, as well as rare events, which are predictive but only observed for a small subset of compounds.

The utility of this concept is demonstrated in this work using liver gene expression and histopathology data from repeat-dose studies in rats provided by the Open TG-GATEs database (Figure 3.2). This allows us to take advantage of previous data curation and work on the dataset itself, in particular by Sutherland *et al.* [242] who provide an adverse classification of each compound-dose combination and toxscores summarising histopathological findings in each condition. Furthermore, DILI is well understood in comparison to other organ-level toxicities and hence processes which are expected to precede injury are already known, including cell death, inflammation and other adaptive stress responses [246]. The concept, however, is generally applicable beyond the toxicity area and transcriptomics data and can be used to derive mechanistic event cascades from time-series data of any kind as long as the first activation of events within the time-series can be defined.

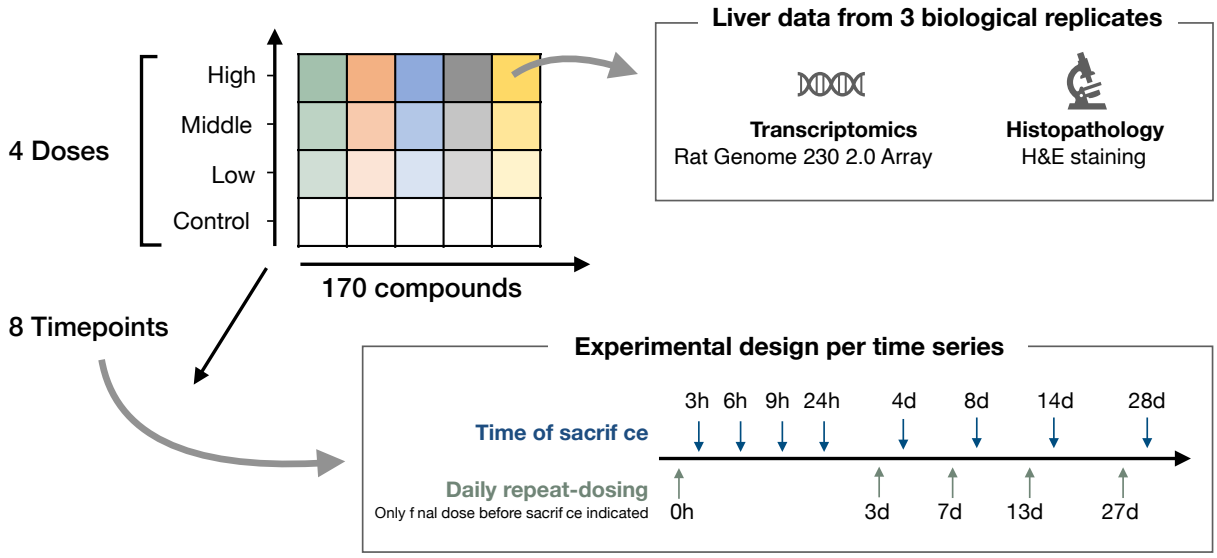


Figure 3.2: Open TG-GATEs study design.

6-week-old male Crl:CD Sprague-Dawley (SD) rats were treated with a range of compounds using daily repeat-dosing. For each compound, four doses were used including a vehicle control, and samples were taken at 8 timepoints. For each combination of compound, timepoint and dose, histopathology was annotated and gene expression measured for 3 replicates.

The time concordance is first described for known processes, similar to mechanistic qAOPs, and then used to prioritize predictive, time-concordant KE providing a strong data-driven, automatable starting point for AOP development, aligning with the objective of cpAOPs (Table 3.1). Data-driven time concordance and prior knowledge on event relations between TFs and gene expression were then combined to generate hypotheses for causal gene-regulatory mechanisms in DILI pathogenesis and to generally show how time concordance can stratify and support other streams of causal evidence. Overall, this work shows that time-resolved gene expression and histopathology data can be used to quantify time concordance across a large set of compounds and events, which allows us to characterize known mechanistic links and to prioritize potentially new ones (Figure 3.1C).

Table 3.1: Comparison of quantitative and computational Adverse Outcome Pathway models.

Comparison of the first activation concept, computationally predicted AOPs (cpAOPs) and quantitative AOP (qAOP) models with respect to their potential roles in AOP development.

		cpAOP	**Probabilistic qAOP**	**Mechanistic qAOP**	**This study**
Output KERs	Coverage of events	All events with known associations	Limited to AOP scaffold	Limited to AOP scaffold	All events that can be inferred from the available data
	Evidence for causality	No	No	Yes Experimental	Yes (Time concordance)
	Evidence for predictivity	No	Yes	No	Yes (PPV is one of the metrics)
Purpose	Support AOP development	Yes Potentially new KER	No (Uses existing AOP)	No (Uses existing AOP)	Yes (Potentially new KER)
	Advance existing AOP	No (Does not add info)	Yes (Enables prediction)	Yes (Better understanding)	Yes (Better understanding)
Approach	Automatable?	Yes	Yes	No (New experiments needed)	Yes
	Large number of chemicals	Yes (All compounds with known associations)	Yes (All compounds with relevant assay data)	No	Yes (All compounds with time-series data)
	Based on *in vivo* data	No (Potentially indirectly)	No	Yes	Yes
	Transferrable to other AE?	Yes	Yes (if AOP and suitable data are available)	Partially (New experiments needed)	Partially (if suitable time-series data is available)

3.2 Methods

3.2.1 Open TG-GATEs data processing

The Open TG-GATEs gene expression data from studies in 6-week-old male Crl:CD Sprague-Dawley (SD) rats with daily repeat-dosing (Figure 3.2) was downloaded from the Life Science Data Archive (DOI: `10.18908/lsdba.nbdc00954-01-000`).

The raw liver gene expression levels were background corrected, $\log_2$ transformed, and quantile normalized with the *rma* function of the `affy` package per treatment across all doses and timepoints [216]. Quality control was then performed using the `ArrayQualityMetrics` package [217] and detected outliers with high distance to other experiments or unusual signal distribution were removed (List of removed outliers summarised in Table B.1). The platform information for the Affymetrix Rat Genome 230 2.0 Array was derived from Gene Expression Omnibus [219] (GEO accession: GPL1355) and was then used to summarise probe IDs to rat gene symbols by median for all probes mapping uniquely to one gene symbol. Only the 360 compound-dose combinations with at least 6 measured timepoints after quality control were included. Out of these all eight timepoints were measured in most time-series, while only six timepoints were measured in two time-series, and only seven timepoints in seven time-series.

3.2.2 Definition of adverse histopathology

To characterize the extent of histological findings, the toxscores by Sutherland *et al.* [242] were used in order to consider both severity and frequency of events in a single numerical output measure. These are based on the lesion severity per animal which was first converted to a numerical scale (normal = 0, minimal = 1, slight = 2, moderate = 3, marked or severe = 4) and then averaged across all biological replicates as an aggregate measure for lesion frequency and severity. One characteristic of this measure is that the overall distributions varied between different findings, e.g. inflammation was more frequently annotated with low than with high toxscores while a more balanced distribution of scores was observed for hepatocellular single cell necrosis (Figure B.1).

Table 3.2: Compounds classified as adverse based on histopathology and concordance with previous annotations.

For the annotations by Sutherland *et al.* [242], who classified each compound at each measured dose as adverse or non-adverse at day 4 and day 29, the adverse doses for each compound are listed. Furthermore, the binary classification as adverse (1) and non-adverse (0) from DILIst [250] are included as well as the vDILIConcern and Severity Class classifications from DILIRank [87] which describe evidence for liver side effects observed in humans derived from post-marketing data.

Compound Name	DILIst Class	vDILIConcern	Severity Class	29 days	4 days
Acetamidofluorene				Middle, High	
Acetaminophen	1	vMost-DILI-Concern	5		
Aspirin	1	vLess-DILI-Concern	0		
Bendazac	1	vMost-DILI-Concern	8	High	High
Bromobenzene	1				
Captopril	1	vLess-DILI-Concern	7		
Clofibrate	1	vLess-DILI-Concern	3	Middle, High	
Clomipramine	1	vMost-DILI-Concern	8	High	
Colchicine	0	Ambiguous DILI-concern	6		
Coumarin				High	
Danazol	1	vMost-DILI-Concern	8		
Dantrolene	1	vMost-DILI-Concern	8		
Diclofenac	1	vMost-DILI-Concern	8		
Ethambutol	1	vMost-DILI-Concern	8	Middle, High	
Ethinylestradiol	1				
Ethionamide	1	vLess-DILI-Concern	3	High	High
Ethionine					
Fenofibrate	1	vLess-DILI-Concern	3	Low, Middle, High	Middle, High
Gemfibrozil	1	vMost-DILI-Concern	4	Low, Middle, High	
Ibuprofen	1	vLess-DILI-Concern	3		
Ketoconazole	1	vMost-DILI-Concern	8		
Lomustine	1	vLess-DILI-Concern	3	High	
Methapyrilene	1			Middle, High	High
Methyltestosterone	1	vLess-DILI-Concern	2		
Monocrotaline				Middle, High	
Nitrosodiethylamine				Middle, High	High
Phalloidin				High	High
Phenacetin	1				
Simvastatin	1	vLess-DILI-Concern	3		
Theophylline	0	vNo-DILI-Concern	0	High	
Thioacetamide				Low, Middle, High	High
Wy-14643				Low, Middle, High	Low, Middle, High

To study which histological findings were enriched in adverse conditions, binary histopathology labels were defined which describe the presence of histological findings with different extents in each time-series. Based on the toxscore ranges used by Sutherland *et al.* [242], three toxscore cut-offs are implemented to describe each histopathological finding "Null" (toxscore > 0), "low" (toxscore > 0.67) and "high" (toxscore > 1.34). It was then studied which labels were over-represented in adverse time-series. These were defined using the annotation of Sutherland *et al.* [242], where pathologists classified compound-dose combinations in the Open TG-GATEs database as adverse or non-adverse after 4 and 29 days of treatment. The 29 days classification was used to define 40 adverse time-series and only regarded time-series as non-adverse for compounds which were not classified as adverse at any dose in the negative control, in order to account for the fact that some of the cellular changes of interest might already take place at lower doses, although the resulting phenotype is not considered adverse yet.

Findings were defined as adverse histopathology if they are observed in at least 5 out of 40 adverse time-series to remove rare histopathological findings, and additionally require that at least 50% of findings are in adverse conditions to remove findings which are unspecific. All labels which were identified with these criteria are significantly enriched among time-series labelled as adverse by Sutherland *et al.* [242] in comparison to those that were considered non-adverse using a one-sided Fisher's Exact test (p-value < 0.0001), performed using the *fisher.test* function of the `stats` R package [221]. However, this combination of additional criteria was chosen to exclude findings which are rare or weakly associated.

While not all compounds in the Open TG-GATEs database are drugs and some mechanisms of toxicity may not translate to humans, out of the 38 compounds represented in adverse time-series, 22 have additionally been classified as hepatotoxic in DILIst [250] and 18 in DILIrank (vMost-DILI-Concern or vLess-DILI-Concern) [87] (Table 3.2). This overlap with compound-level DILI annotations by the FDA shows that the compounds in this study partially represent known mechanisms of DILI in humans, while also highlighting the fact that a clear classification is not possible.

3.2.3 Pathway and TF activity inference

The activity of pathways and TFs across all doses and timepoints of a treatment including vehicle controls was derived based on the expression of its gene sets members using `GSVA` [175], which computes a gene set enrichment by sample matrix from the gene expression by sample matrix. This was performed using a Gaussian kernel requiring at least 5 genes per gene set, and overall provides the basis for the subsequent pathway- and TF-centric steps. As prior knowledge, pathway maps from Reactome were used [167] which were derived through MSigDB [226] and the `msigdbr` package [251]. TF activity gene sets were derived from `DoRothEA` [170] and mapped from human to rat gene symbols with `biomaRt` [227]. These gene sets describe known, functional TF-gene interactions and are assigned a confidence level based on the strength of evidence of these interactions. Thereby, only the 207 TFs with a high to medium confidence level of A-C were included and the few TF-gene interactions with a negative mode of regulation were removed to better infer TF directionality. To evaluate which pathway or TF is dysregulated, the differential activity in comparison to the vehicle control group, which was treated for the same amount of time and as part of the same experiment, was computed using the moderated t-statistic in `limma` [162]. Significantly dysregulated gene sets were identified using a False Discovery Rate (FDR) < 0.05.

3.2.4 Temporal concordance of events

In this study, the order of events was derived based on each event's timepoint of first activation within each time-series (Figure 3.1A). For pathways and TFs, first activation was defined as the earliest time of measurement at which significant differential regulation was observed (FDR < 0.05) in each direction, while an additional logFC cut-off has been implemented for individual genes. As first evidence of adverse morphological changes in the liver, the first timepoint at which any of the adverse histopathology labels is found was used.

We were then generally interested in potential preceding events ***PE*** which are first activated before or at the same time as a potential later event or outcome ***LE*** and used multiple metrics to quantify the degree of time concordance which can be related to the original work by Bradford Hill (Table 3.3). Thereby, the key later event in this study was adverse histopathology but a more general notation ***LE*** was used, as some of the following criteria

to quantify time concordance are also applied in the TF analysis, where gene expression-derived events are used as later event, and since in general any event may serve as later event.

First, the true positive rate (TPR) was used which describes how frequently ***PE*** is observed before ***LE*** among all time-series with ***LE*** and hence its consistency across compounds. Secondly, the maximal effect size of ***PE*** observed before ***LE***, summarised across time-series by median, was used to characterise the strength of association. To evaluate the significance of the findings, additionally a set of background time-series unrelated to ***LE*** was defined (Figure 3.1B). For adverse histopathology, these unrelated background time-series were the 133 time-series without any observed histological changes. The enrichment of ***PE*** before or at ***LE*** was then computed using the *fisher.test* function of the `stats` R package [221], first estimating the odds ratio using the conditional maximum likelihood estimate and subsequently testing the null hypothesis whether the odds ratio derived from a confusion matrix as described in Figure 3.1 is equal to or smaller than 1. Additionally, the positive predictive value (PPV) of ***PE*** for ***LE*** was computed, which describes how likely ***LE*** is observed at the same or a later time given the observation of ***PE.***

Table 3.3: Metrics quantifying the time concordance between a potential preceding event *PE* and potential later event *LE,* and their relation to the original Bradford Hill (BH) considerations.

BH consideration	Metric	Formula	Description
Consistency	True positive rate (TPR)	p(**PE** →**LE**|**LE**)	Fraction of time-series with event ***PE*** with specified temporal relation among time-series with event ***LE***
Specificity	Positive predictive value (PPV)	p(**PE**→ **LE**|**PE**)	Fraction of time-series with event ***LE*** with specified temporal relation among time-series with event ***PE***
Temporality	Time concordance p-value	One-sided Fisher's Exact test	Likelihood of observing event ***PE*** and ***LE*** with specified temporal relation with equal or higher frequency by chance assuming a hypergeometric distribution.
Strength	Effect size in time-series with ***LE***	Median (logFC)	Median logFC of ***PE*** observed in time-series with ***LE*** (in comparison to vehicle control)

Across all metrics, only time-series were considered in the statistics for which any event of the same type as ***PE***, e.g. TF or pathway, was observed at the included timepoints, so before or at ***LE*** or at any timepoint in the background time-series. This was done to account for the

fact that in some cases no changes are found which may be a consequence of the fact that there isn't a measured timepoint before ***LE*** or that at the available timepoints expression changes cannot be detected. The argument for this is that in these cases this should not be treated as evidence of absence of the given event, but rather as absence of evidence.

3.2.5 Combining time concordance on TF-TF interactions

We used three sources of causal prior knowledge to derive mechanistic hypotheses linking TFs: Protein-protein interaction between TFs derived from Omnipath through `OmnipathR` [252,253], TF-target gene interactions from `DoRothEA` [170] and the link between gene expression and protein levels following the central dogma of molecular biology. Using these interactions as backbone, those additionally supported by time concordance were derived. Thereby, the dysregulation of the nodes was required to match the reported mode of regulation (edge sign) and the source node or upstream event was required to be observed in at least 20% of cases before or at the same time as the target node or downstream event. For induced TFs, significant enrichment of gene expression ($|logFC|>0.5$) and TF activity before adverse histopathology was required, as well as evidence for changes in expression preceding changes in the same direction in regulon activity within the same time-series.

3.2.6 Time dependence

In each adverse time-series, Spearman correlation between timepoint and event activation logFC was evaluated using the `correlation` R package [254] including a logFC of 0 at timepoint 0 h assuming that there are no differences in comparison to the control group before treatment. Then, pathways and TFs were identified which only show significant Spearman correlation in one direction, positive or negative. For those events, the Fisher's combined probability test was applied using the `metap` R package [255] across all adverse time-series to evaluate whether overall significant correlation between event activation and time is found.

3.3 Results and Discussion

In order to derive the time concordance between cellular events and later adverse histopathology, the workflow outlined in Figure 3.3 was used with each step being also

introduced in the subsequent sections and details on their respective implementation in Methods.

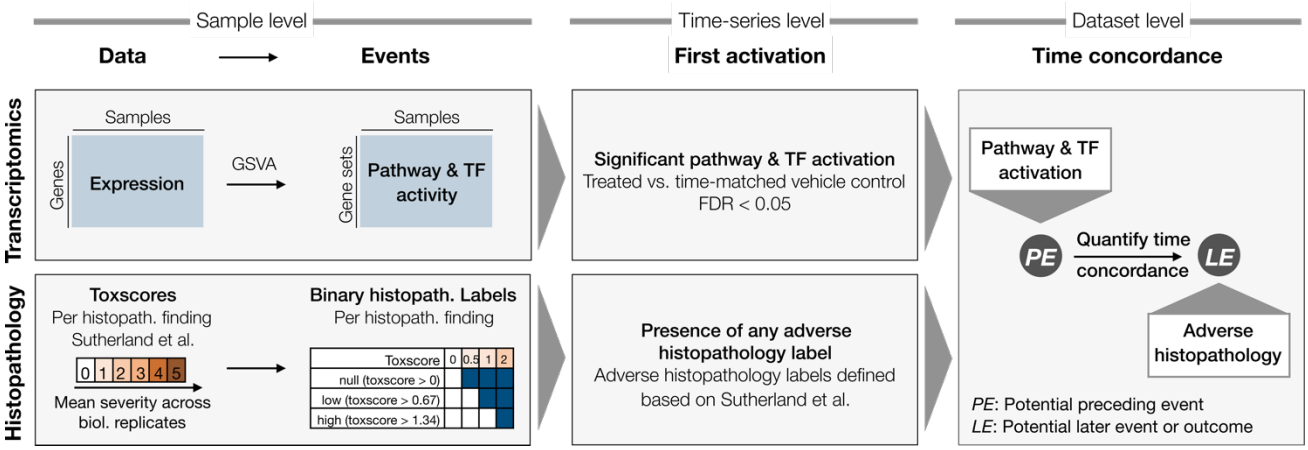


Figure 3.3: Workflow to quantify time concordance between preceding gene expression-derived events and later adverse histopathology.

First, events are derived from the gene expression and histopathology data. Pathway and TF activity is inferred based on the expression of the respective gene sets using GSVA [175] and binary histopathology labels are derived from the continuous toxscores. Secondly, the first activation of expression-based events as well as of adverse histopathology are derived. Lastly, the time concordance between potential preceding events (PE) which are derived from gene expression and adverse histopathology as potential later event (LE) are derived.

First, TF and pathway activities were derived across expression profiles from the same experiment and subsequently defined the first up- or downregulation TFs or pathways as events. Furthermore, binary histopathology labels describing the occurrence of each histopathological finding at different levels of severity and frequency were derived from the toxscores provided by Sutherland *et al.* [242]. Subsequently, the earliest timepoint of each event, e.g. pathways or adverse histopathology, was derived within each time-series. As last step of the time concordance analysis, it was then evaluated which gene expression-derived events are significantly enriched before or at the time where adverse histopathology is found, as well as additional time concordance metrics outlined in Table 3.3.

3.3.1 Adverse histopathological findings and their temporal relation

To define the earliest timepoint of adverse histopathology within each time-series, the annotations of time-series as adverse or non-adverse by Sutherland *et al.* [242] are used, as well as the toxscores, which summarise the severity and frequency for each histological finding and each compound-dose-time combination as mean severity score and range from 0 (normal) to 4 (severe). These toxscores were used to define three levels for each histological finding: “null” (toxscore > 0), “low” (toxscore > 0.67) and “high” (toxscore >1.34).

For example in case of a toxscore of 1, both “null” and “low” are considered to be present. It was then evaluated which histology groups were frequently found in the adverse compound-dose combinations (observed in >10% of adverse time-series corresponding to at least 5 out of 40 cases) with at least 50% of findings being in adverse time-series (Figure 3.4A). All of the included histology groups are significantly enriched in adverse conditions, however, these criteria were implemented to identify findings with a certain specificity and frequency instead of allowing a trade-off between both.

The histology groups which passed the filtering are regarded as adverse histopathological findings and include hepatocellular single cell necrosis and biliary hyperplasia at all toxscore thresholds. In contrast, only some of the three toxscore thresholds were selected with the above criteria for all other findings, e.g. the two higher toxscore cut-offs for hepatocellular necrosis and inflammation and only the “high” cut-off for increased hepatocellular mitosis. In all cases, the lower toxscore level was also frequently observed in non-adverse conditions and hence considered too unspecific. In contrast, only the two milder levels of fibrosis were included in the selection, as severe fibrosis was observed rarely.

While the described definition of adverse histopathological findings is used in this study, the difficulty in summarising a complex phenotype such as DILI into a binary classification, adverse or not adverse, is well established [97,98] and is also demonstrated by the discrepancies between DILI classifications from DILIst [250], DILIrank [87] and those derived by Sutherland *et al.* [242] based on the Open TG-GATEs data. Aware that also broader or more

targeted phenotypes might be of interest, a R/Shiny app was developed in which results for alternative definitions of adverse and non-adverse histopathology groups can be explored.

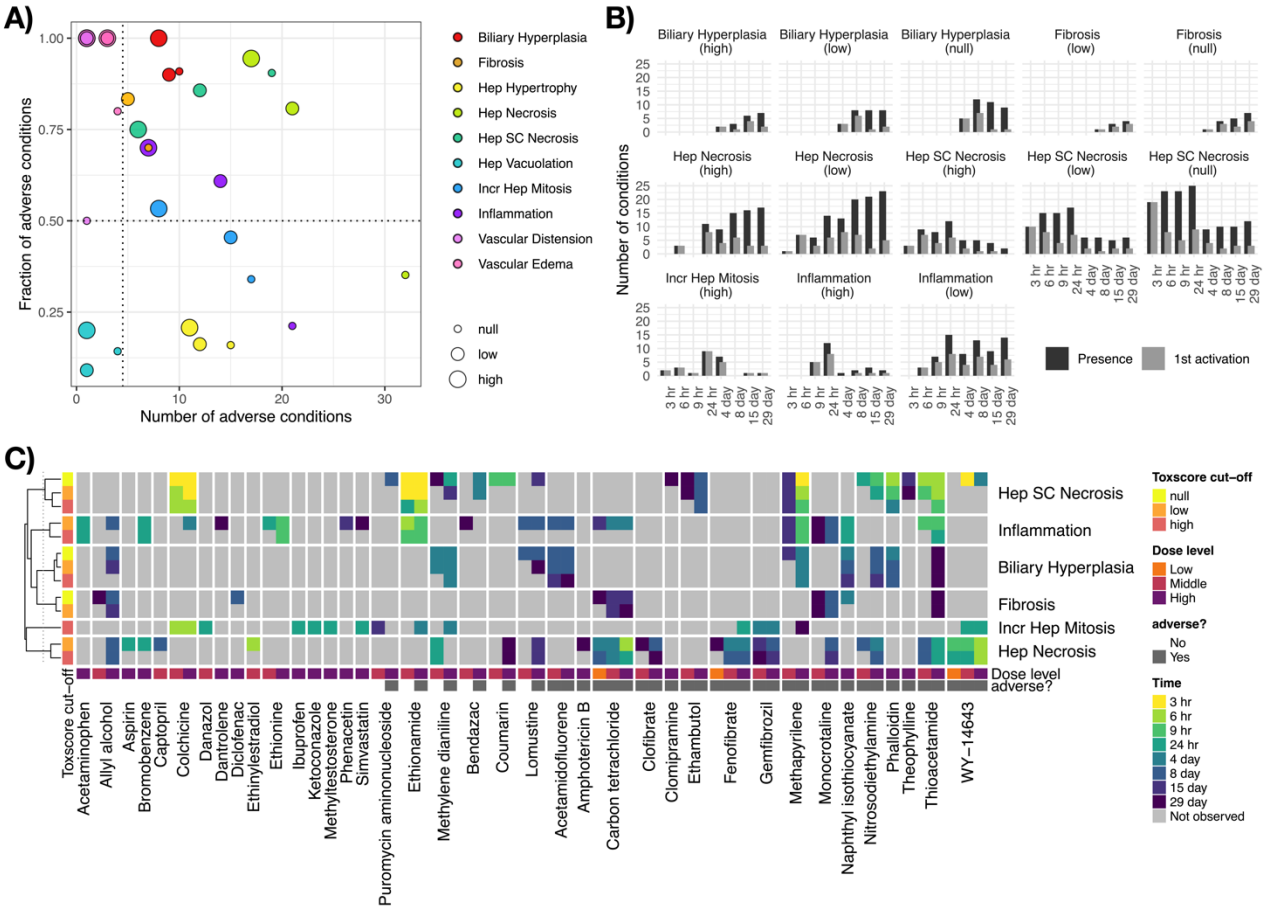


Figure 3.4: Distribution and relation of histopathological findings across time-series.

A) Histopathology labels are defined for each histopathological finding at 3 different toxscore cut-offs, namely “null” (toxscore>0), “low” (toxscore>0.67) and “high” (toxscore>1.34. For each label, the number of occurrences in the 40 adverse time-series and the fraction of adverse time-series among all occurrences of the given histopathology label are shown. Histopathological findings, out of which at least 50% and at least 5 of the occurrences were found in adverse conditions timeseries were considered adverse B) Number of conditions with histopathological findings at different timepoints, as well as the frequency of the respective first activations C) Time of first activation across timeseries labelled as adverse or non-adverse. Each time-series is annotated with the dose level in repeat-dose studies, as well as whether or not the time-series was considered adverse by Sutherland *et al.* [242].

For the adverse histopathology labels, the distribution of toxscores and first activation over time (Figure 3.4B) shows that some findings are predominantly found late, like fibrosis, while others are predominantly found early, e.g. hepatocellular single cell necrosis. Next, out of all

360 time-series with at least 6 measured timepoints, the 61 time-series in which any of the adverse histopathology labels is found were identified, which covered 38 compounds (Table 3.2). In those, the earliest evidence of an adverse phenotype is used to approximate the timepoint of the primary adverse phenotype. Across all time-series with adverse histopathology, hepatocellular single cell necrosis is found most frequently as the primary adverse phenotype, while biliary hyperplasia at any severity is in most cases a secondary effect (Figure 3.4C and Figure B.2).

3.3.2 Known pathways in DILI preceding adverse histopathology

To identify cellular mechanisms in the early pathogenesis of DILI, time-concordant cellular changes preceding later adverse histopathology were studied (see Methods). This identified 911 pathway-level events (37.3%), and 108 TF-level events (33.6%) with significant enrichment (Time concordance p-value<0.05) before or at adverse histopathology. As next step, time concordance was evaluated for a set of ten known events in DILI (Figure 3.5 and Table B.2).

Recycling of bile acids and salts was the most significantly enriched geneset overall and hence also among the ones linked to known events. Also down-regulation of the other bile acid gene sets was significantly enriched (Time concordance p-value < 0.05) pointing to an overall down-regulation of bile acid metabolism. While cell death was also only found to be up-regulated, dysregulation in both directions was found to precede injury for all other key events (Figure 3.5). However, only for peroxisomal processes, namely peroxisomal protein import and beta-oxidation of very long fatty acids, both directions were significantly enriched indicating that dysregulation in either direction might be linked to injury. Overall, significantly enriched gene sets are found for all ten represented known events in DILI (Time concordance p-value < 0.05) indicating that the analysis is able to recover known cellular events.

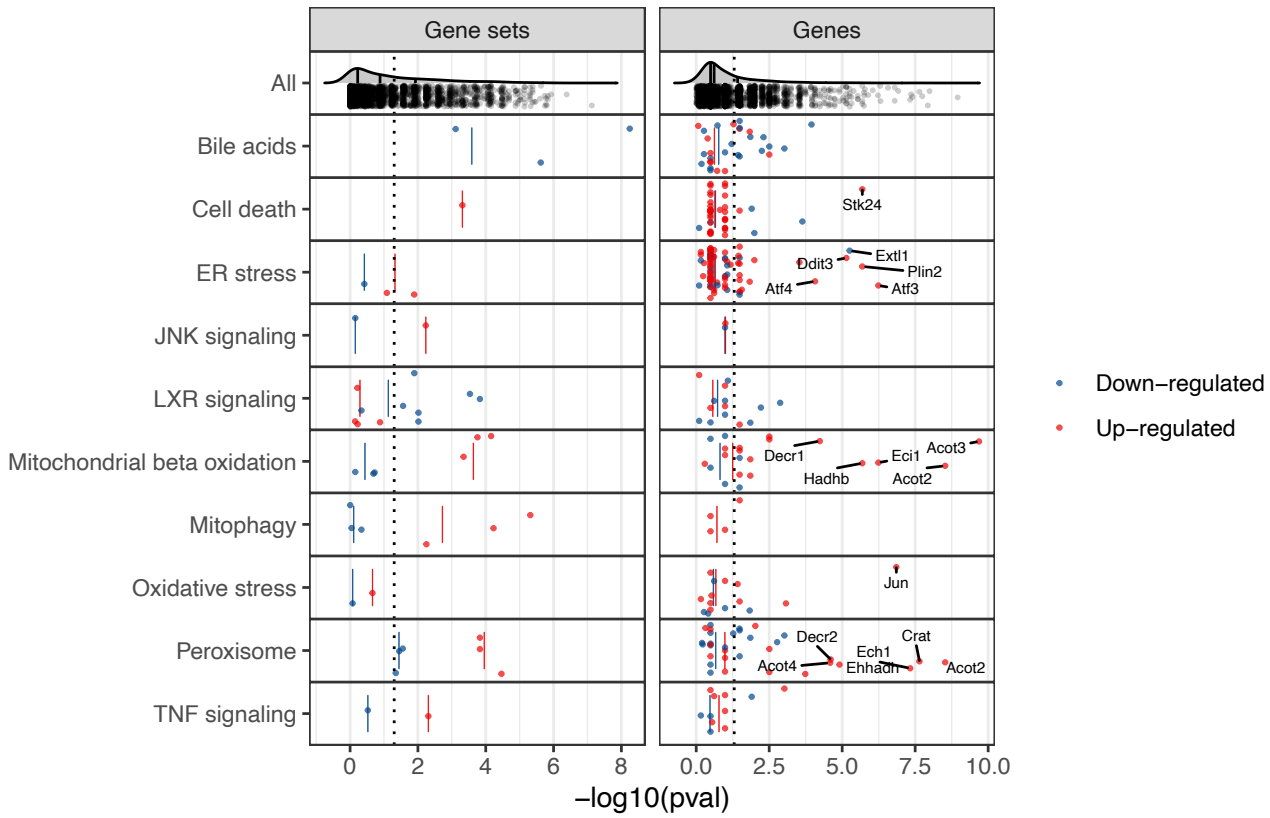


Figure 3.5: Enrichment of known events in DILI before adverse histopathology based on gene sets as well as individual gene members.

The enrichment of first activation before or at adverse histopathology is shown for gene sets mapping to known key events in DILI, for which first activation was defined as first timepoint of differential GSVA-derived gene set activity. Furthermore, also enrichment of individual genes within these genesets is shown and was derived based on the first timepoints of differential expression. Aligning with the expected direction, a significant down-regulation of Liver X Receptor (LXR) signalling and bile acid-related pathways is observed, while all other gene sets were found to be more significantly up-regulated. Only for peroxisomal pathways, both directions were significantly enriched indicating that dysregulation in direction might be linked to adverse histopathology.

To gain insights on a more fine-grained level, the enrichment of significantly and strongly (absolute log fold change > 1) dysregulated individual genes from the above gene set (File B.1) was further analysed. Among the ten most significantly enriched gene-level events, three are involved in known processes, namely the up-regulation of acyl-CoA thioesterase 2 (*Acot2*), acyl-CoA thioesterase 3 (*Acot3*) and carnitine O-acetyltransferase (*Crat*) which are involved in fatty acid beta oxidation [256,257]. Multiple genes among the ten most significantly enriched gene-level events are also involved in mitochondrial and

peroxisomal processes except for *Gadd45a,* Growth Arrest And DNA damage-inducible protein which has a known role in hepatic fibrosis [258], Neutral cholesterol ester hydrolase 1 (*Nceh1*) which is involved in cholesterol metabolism in macrophages [259], ras-related protein Rab-30 (*Rab30*) elevated in early liver regeneration [260], as well as the serine/threonine protein kinase NIM1 (*Nim1k*).

For JNK signalling, no significantly enriched genes were found indicating that while the overall process is changing none of the individual genes shows strong and frequent expression changes. In contrast, the opposite was found for oxidative stress with the Jun proto-oncogene (*Jun*) being one of the most significantly enriched gene-level events but lacking significant changes on the gene-set level. This shows that both gene- and gene-set level analysis can provide complementary insights into cellular changes preceding DILI, and that in some cases effects can be attributed in individual genes which might give more detailed information about the cellular changes.

While significant enrichment before or at adverse histopathology can be regarded as a necessary criterion for time concordance, the temporal event relationship can be further characterised based on the observed behaviour across experimental conditions which may be useful to further prioritize mechanistically relevant pathways in a hypothesis-free manner. Following the Bradford-Hill considerations (Table 1.5), it was hypothesized that this might be the case for observed effect size, frequency and specificity of event occurrence before adverse histopathology. Firstly, it was investigated how strongly pathways were dysregulated comparing the maximal absolute log fold changes (|logFCs|) before or at adverse histopathology in each adverse time-series for significantly time-concordant events (Figure 3.6).

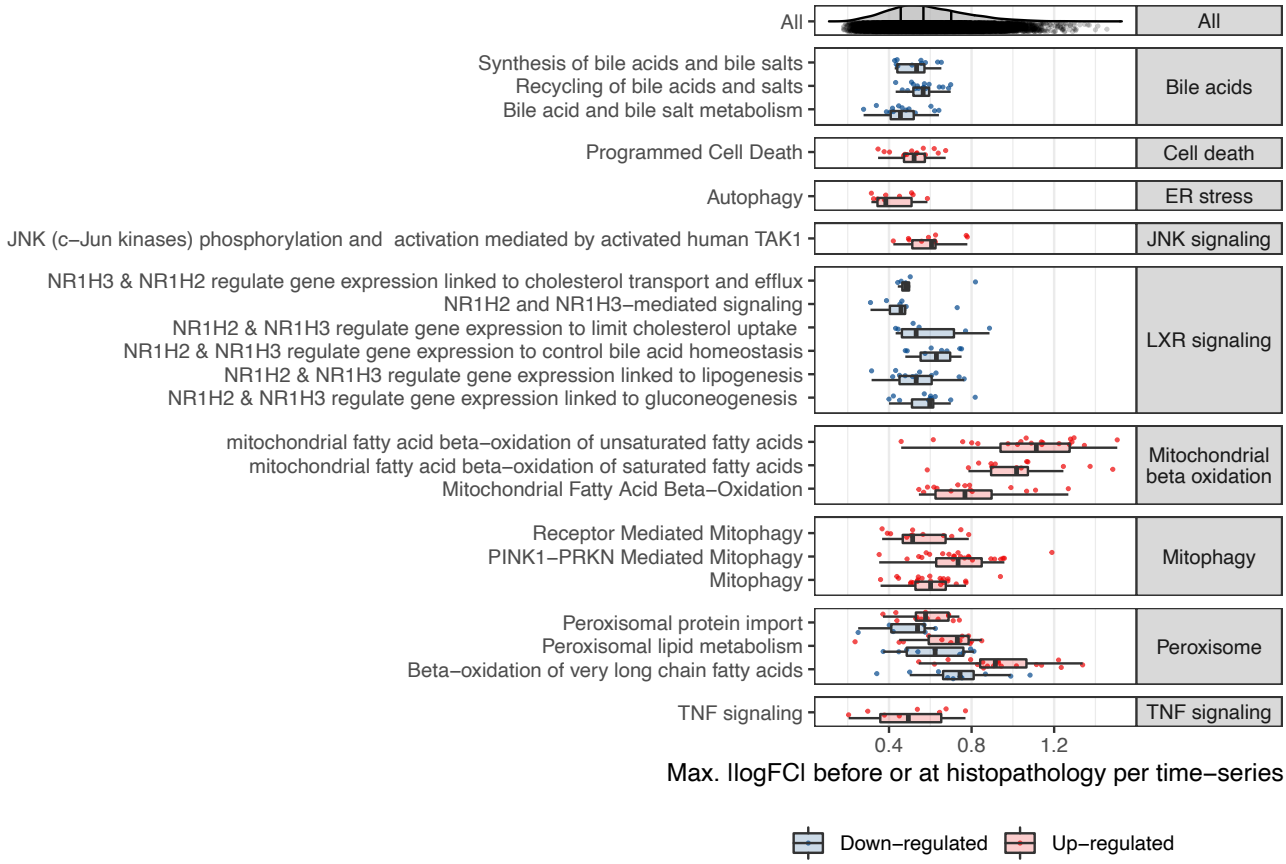


Figure 3.6: Observed max. |logFC| before adverse histopathology.

For known processes in DILI which correspond to significantly enriched events before adverse histopathology, the max |logFC| before adverse histopathology is shown. In comparison to other known pathways and the overall background distribution, a high logFC is found for mitochondrial beta oxidation followed by peroxisomal beta oxidation and mitophagy.

High median maximal |logFCs| were overall found for mitochondrial and peroxisomal pathways and the highest median maximal |logFC| among all significant events was found for mitochondrial fatty acid oxidation of unsaturated fatty acids. At the same time, however, high variance was observed for pathways with high median maximal |logFC| and only a moderately high |logFC|s was observed for other known pathways in DILI, such as programmed cell death. This indicates that a high magnitude of |logFC| is not necessary to contribute to an adverse event, but at the same time can be a useful property to further prioritize important pathways.

It was next analysed to what extent dysregulation in a pathway is predictive for a particular type of histopathology. To this end, it was calculated across how many adverse time-series each pathway is observed, summarised by the true positive rate (TPR), and the positive

predictive value (PPV) indicating whether presence of the key event is a confident indicator for the later adverse event (Figure 3.7). The focus was on significantly enriched events only (Time concordance p-value < 0.05) and a trade-off was found with respect to the highest TPR and PPV (Figure 3.7; for distribution of all events see Figure B.3). This generally shows that either highly frequent events with lower specificity can be identified, e.g. increased mitophagy (TPR: 0.41, PPV: 0.72), or more specific events at the expense of lower relative frequency, e.g. bile acid recycling (TPR: 0.30, PPV: 1).

Surprisingly, lower relative frequencies are particularly observed for stress response and signalling pathways with only liver X receptor (LXR)-dependent gene expression linked to lipogenesis reaching a TPR over 20%. One explanation for the lower observed frequencies is that these pathways are predominantly and initially mediated through post-transcriptional alterations instead of gene expression changes [261,262], making the expression of pathway members a weak proxy for pathway activation in early pathogenesis and explaining the overall low frequencies. In fact, one reason LXR-dependent changes might have achieved higher frequencies as they explicitly include the downstream regulated genes unlike the other signalling and stress response pathways [167].

Due to the previously discussed complementarity of gene- and gene set-level analysis, TPR and PPV are also shown for individual genes with a focus on those which are involved in gene sets mapping to known key events. The most significant genes, already highlighted in Figure 3.5, reveal a high frequency for the up-regulation of the acyl-CoA thioesterases *Acot2*, *Acot3* and *Acot4*, as well as the for the enoyl-CoA hydratase 1 *Ech1* which aligns with the relatively high frequency of pathway-level events linked to mitochondrial and peroxisomal processes. Furthermore, the most frequent gene-level events with a PPV=1 are the up-regulation of activating transcription factor 3 (*Atf3*) which was found to promotes hepatic fibrosis [263] and enoyl-CoA delta isomerase 1 (*Eci1*).

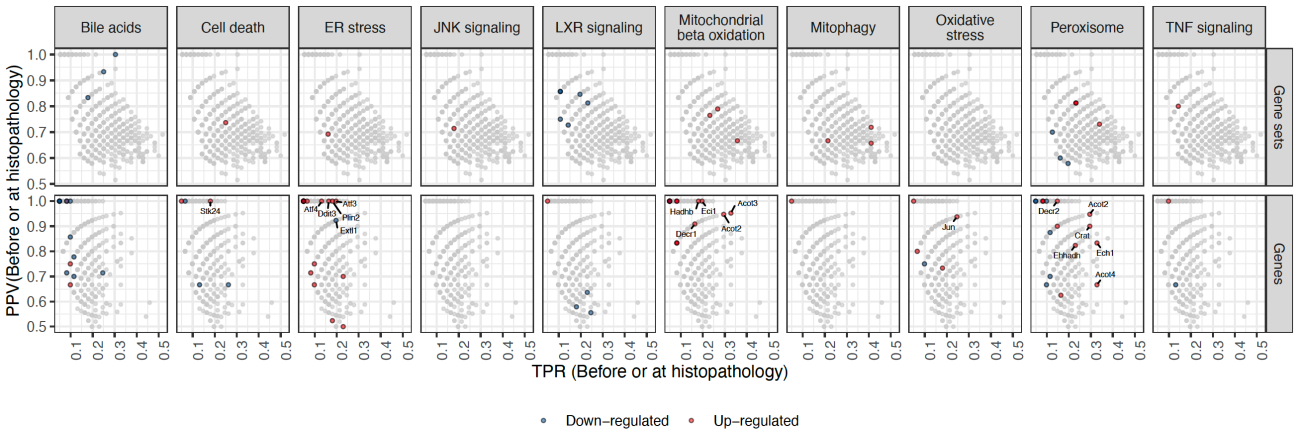


Figure 3.7: True positive rate (TPR) and positive predictive value (PPV) before or at histopathology of genes and gene sets in known key events in DILI.

Events related to the given known key event are shown in red or blue indicating an up- or downregulation, respectively. Genes with a time concordance p-value < 0.0001 involved in known key events in DILI are additionally labelled. The background distribution of all significantly enriched genes or gene sets is shown in grey (Time concordance p-value < 0.05).

3.3.3 Known TFs in DILI preceding adverse histopathology

To gain insight into signalling and expression regulation preceding adverse histopathology, TFs were analysed next as these are involved in early perturbation response preceding downstream gene expression changes and also are likely to show strong signal in transcriptomics data given their direct link to gene expression. As known TFs in DILI, TFs mediating the stress response and signalling pathways already introduced above were included, as well as nuclear receptors which take in important roles in liver physiology and malfunctions and can be, both, MIEs or KEs (mapping shown in Figure 3.8A). Consistent with the pathway-level results, an enriched up-regulation was found for nuclear factor erythroid 2-related factor 2 (Nfe2l2) which is a key mediator of oxidative stress [245,264] as well as the Nf-κB subunits Rela and Nfkb1 indicating inflammation [265], while the oxysterols receptors LXRα (Nr1h3) and LXRβ (Nr1h2) which control lipid metabolism showed enriched down-regulation [266].

For ER stress, three TFs mapping to the three branches of unfolded protein response (UPR) [267] were included: Activating transcription factor 4 (Atf4), activating transcription factor 6 (Atf6) and X-box binding protein 1 (Xbp1). Atf4 up-regulation was found to be most significantly enriched, most frequent and also showing the largest logFC (Figure 3.8). This

highlights its overall importance in mediating ER stress and is consistent with the known role for ATF4 in DILI [268]. While Atf4 is a member of the pro-apoptotic UPR branch, the ATF6 and XPB1-mediated branches tend to be cytoprotective [269]. In agreement with this, Atf6 was not significantly enriched, however, Xbp1 showed rare but significantly enriched down-regulation.

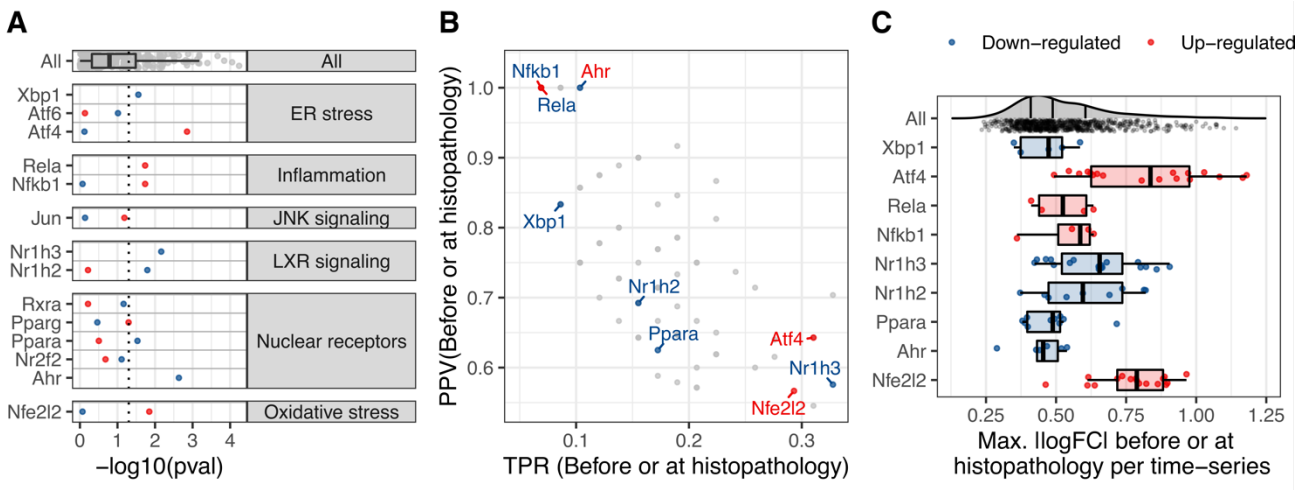


Figure 3.8: Temporal concordance of nuclear receptors and adaptive response transcription factors (TFs) in DILI.

For known TFs in DILI the following time concordance metrics are shown: A) The enrichment significance before or at first adverse histopathology, B) Positive Predictive Value (PPV) and True Positive Rate (TPR), C) Max. mean |logFC| before or at first adverse histopathology. As background distribution in grey, the statistics for all inferred TFs is shown.

Transcription factor AP-1 (Jun) which is one of downstream target TFs of c-Jun N-terminal kinase (JNK) signalling was not significantly enriched in either direction due its rare activation among adverse time-series although JNK signalling up-regulation itself was significantly enriched with *Jun* up-regulation being one of the most significantly enriched gene-level events. However, JNK signalling is particularly known in acetaminophen-induced liver injury and in this context leads to hepatocyte death through interactions with Sab on the mitochondrial outer membrane and not through transcriptional regulation mediated by AP-1 [270,271]. As increased Jun activity is hence known to be a consequence of JNK signalling but not a cause of injury, it would be plausible to see enriched pathway activity but not in TF activity before adverse histopathology. Overall, it was hence possible to show significant enrichment of some of the known TFs in DILI before adverse histopathology and also to biologically reason the absence of significance for others.

While none of the included TF-level events ranked as most significant or most strongly changing before adverse histopathology as in the analysis of pathway-level events, the down-regulation of Nr1h3, which is involved in lipid metabolism, was identified as most frequent event (Figure 3.8B) indicating that the linked physiological changes are commonly but not specifically found before injury. Similarly, the up-regulation of stress response, indicated by Nfe2l2 and Atf4, was found to be frequent aligning with their role in adaptive stress response [272]. Overall, frequency might hence be a useful metric to identify pre-adverse cellular events which precede injury but are not highly specific.

3.3.4 Prioritization of cellular events taking place before adverse histopathology

As many events were found to be significantly enriched before adverse histopathology, the next aim was to identify and characterize events most supported by time concordance, and hence to move closer to the eventual aim of constructing AOPs from data. In our analysis, some known events in DILI ranked highest by time concordance p-value while others rank highest by max. |logFC| before adverse histopathology. In contrast, known TFs in DILI were found as most frequent ones in the dataset. Hence, the top 10 TF- and pathway-level events were identified using max. |logFC|, the time concordance p-value, and the TPR before or at adverse histopathology. These are shown in Figure 3.9 while the time concordance metrics for these events are summarised in Table B.3 and Table B.4, and all time concordance metrics can be found in File B.1. The most significantly enriched pathway-level event is decreased bile acid and salt recycling and also the down-regulation of multiple metabolic pathways, in particular targeting glycosaminoglycans, is found among the most significant pathway-level events pointing towards reduced liver function. Moreover, the most significantly enriched TF-level event was the down-regulation of Transcription factor activating enhancer binding protein 4 (Tfap4) which shows emerging roles in cell fate decisions [273], and is followed by Homeobox B13 (Hoxb13) for which expression has previously been found to correlate with hepatic inflammation in hepatic fibrosis [274].

Among the up-regulated events, the most significant enrichment is found for cell cycle checkpoints and DNA repair among the pathway-level events as well as E2F transcription factor 2 (E2f2), which controls cellular proliferation and liver regeneration [275], and was found

among the most significant TF-level events. E2f2 up-regulation was also identified as 2nd most frequent TF event after the down-regulation of Nr1h3 and among the top 10 most strongly changing TF events further highlighting its strong time concordance.

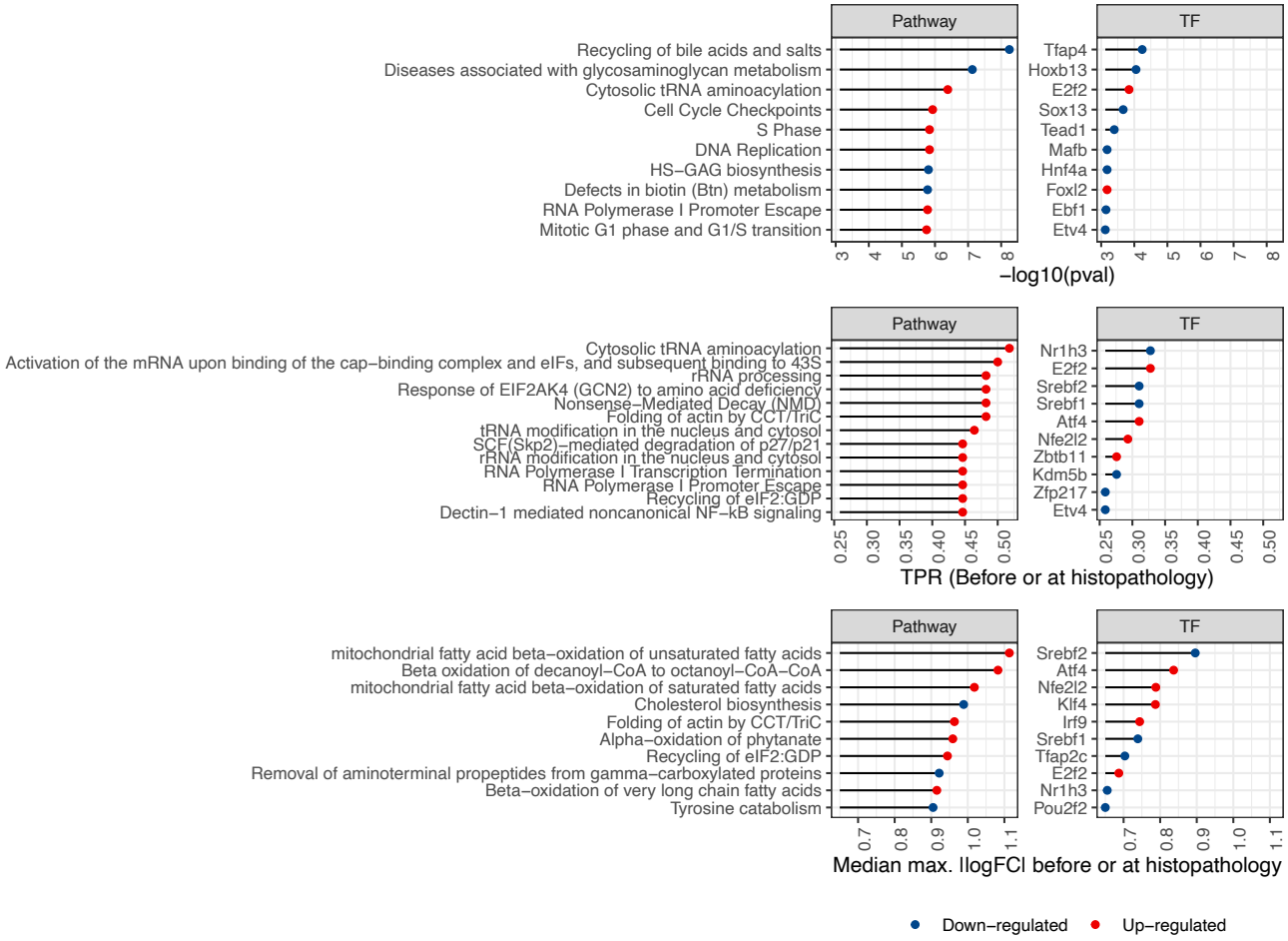


Figure 3.9: Highest ranking events by time concordance metrics.

The ten transcription factor (TF)- and pathway-level events ranking highest by time concordance p-value, median max. |logFC| and true positive rate (TPR) before or at histopathology is shown.

The most frequent genesets point to translation regulation via eukaryotic translation initiation factor 2A (EIF2a) including the upstream response mediated by eIF2α kinase GCN2 and the downstream role in protein translation mediated through interactions with tRNA. EIF2a is part of the same branch of UPR as Atf4 and causes its preferential translation which, among others, mediates autophagy and proapoptotic response [267,276] and is a known predictor of DILI [277]. Furthermore, increased folding of actin by chaperonin containing tailless complex polypeptide 1 (CCT) or tailless complex polypeptide 1 ring complex (TriC) is found frequently and with large effect size. It has been previously linked to proteostasis and autophagy [278,279], but a role in DILI specifically is not yet known.

As most strongly dysregulated events, metabolic pathways are found pointing to increased beta oxidation of fatty acids, as well as decreased cholesterol biosynthesis and tyrosine catabolism. Also the most strongly down-regulated TFs point towards lipid metabolism, i.e. the Sterol Regulatory Element Binding Transcription Factor 1, Srebf1, and the Sterol Regulatory Element Binding Transcription Factor 2, Srebf2, as well as Nr1h3 which controls *Srebf1* expression. Overall, the derived time-concordant events, which take place between the beginning of treatment and onset of adverse histopathology, hence include known and plausible events in liver injury which can be further characterized based on their frequency, significance and logFC.

3.3.5 Mechanistic hypotheses based on known TF functions and time concordance

While both pathways and TFs constitute interpretable events in this study, further prior knowledge is available on how TFs can function on a molecular level allowing us to derive more detailed hypothesis. Firstly, TF activity can generally be modulated through changes in expression or by post-transcriptional regulation as consequence of cellular signalling or environmental changes (Figure 1.7). In case of transcriptional regulation, changes in mRNA levels should precede changes in TF activity estimated based on regulon expression and hence time concordance can be used to gain support for transcriptional TF regulation. Being only interested in TF events with a potential mechanistic link to liver injury, it was studied how significantly concordant expression and activity for each TF are enriched before adverse histopathology.

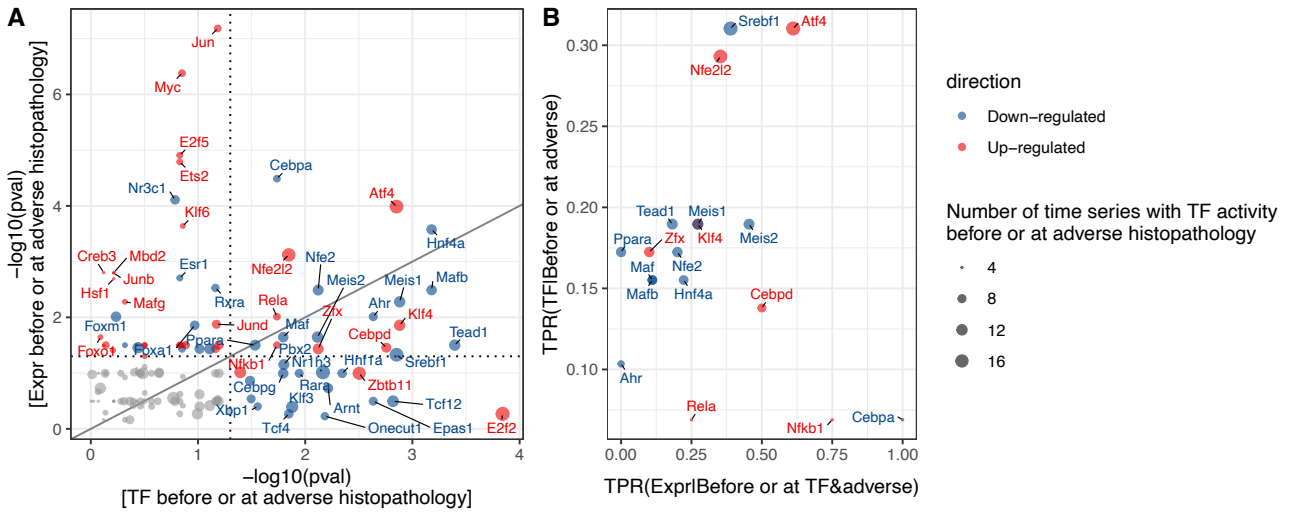


Figure 3.10: Transcription Factor (TF) activity and expression before adverse histopathology.

A) Significance of enrichment in adverse conditions for matched TF activity and expression-based events. Events only found on the expression or TF level are not included in the figure due to the inability to perform a statistical test for those. B) For significantly enriched TF activity-based events, the True Positive Rate (TPR) of observing TF activity before or at the time of adverse histopathology is shown, as well as the TPR for observing TF expression changes before TF activity in the time-series where it precedes adverse histopathology.

The strongest evidence for a role in DILI pathogenesis is found for 18 TF events which show both significantly enriched TF expression and regulon activity, providing complementary evidence of TF importance and hinting at transcriptional regulation (Figure 3.10A). While this is not the case for the 17 TF events which only show significantly enriched TF activity but insignificant enrichment of differential expression, including increased E2f2 activity, this can be explained by post-transcriptional regulation potentially describing earlier response patterns which are a direct consequence of upstream signalling. In contrast, 35 TF events with only significant gene expression, such as increased Jun or Myc, might be already showing changes in expression but not sufficiently large changes in activity yet. As this rather indicates a role in later pathogenesis and expression is only regarded as supporting evidence, these TFs have not been included in the next analysis steps.

To derive stronger mechanistic evidence for induction, it was next evaluated how frequently expression changes precede TF activity in the same adverse time-series and compare this against the overall frequency of TF event occurrences preceding adverse histopathology (Figure 3.10B). Among the events with significant enrichment of TF

expression and activity, the most frequent evidence for induction was found for the down-regulation of CCAAT/enhancer-binding protein α (Cebpa). In humans, decreased expression of the homologous CEBPA is not only known across liver diseases, exogenously increased CEBPA expression has also been shown to reverse liver injury and is explored as therapeutic target in hepatocellular carcinoma [280]. The event with the second-highest relative frequency of expression preceding TF activity as well as the highest frequency of TF activity preceding injury is Atf4, for which expression of the homologous gene in humans is known to be induced as part of the ER stress response contributing to adverse liver phenotypes [281,282]. In contrast, it was found that for the aryl hydrogen receptor (Ahr) and the peroxisome proliferator-activated receptor α (Ppara) changes in expression never preceded those in TF activity which aligns with their roles as nuclear receptors which are generally post-translationally activated via ligand binding [283,284]. As this provides counterevidence for transcriptional induction, these were not included as induced TF in the subsequent analysis.

After investigating the mode of regulation for individual TFs above, it was next considered how these TFs are interlinked. To this end, protein-protein interactions and, for induced TFs, TF-target gene interactions between significantly enriched TFs were identified, which showed significant enrichment before adverse histopathology for both expression and regulon activity, as well as evidence of expression preceding TF activity within the same adverse time-series. Results of this analysis are shown in Figure 3.11, and details on the observed absolute and relative frequencies, as well as the source of the interaction are shown in Table B.5.

One of the two identified interactions by the highest absolute frequency is Nr1h3 down-regulation resulting in reduced Srebf1 activity. Furthermore, Srebf1 is also linked to upstream regulation by Nr1h2 which interacts with peroxisome proliferator-activated receptor (Ppara) in both directions, and this cross-talk between Ppara and LXR regulating Srebf1 expression has been explicitly studied in the context of fatty acid metabolism regulation [285–287].

The 2nd most frequently observed interaction is the down-regulation of transcription factor 12 (Tcf12) inducing reduced activity of TEA domain transcription factor 1 (Tead1). While Tead1 is indeed known to be involved in liver diseases and injury [288,289], the interaction itself has not been reported before in the context of liver injury and the same applies also

for the other upstream Tead1 regulators identified. It should also be noted that for these interactions first activation is only found at the same time but not in the time-concordant order providing weaker evidence than, for example, the interaction between Nr1h3 and Srebf1.

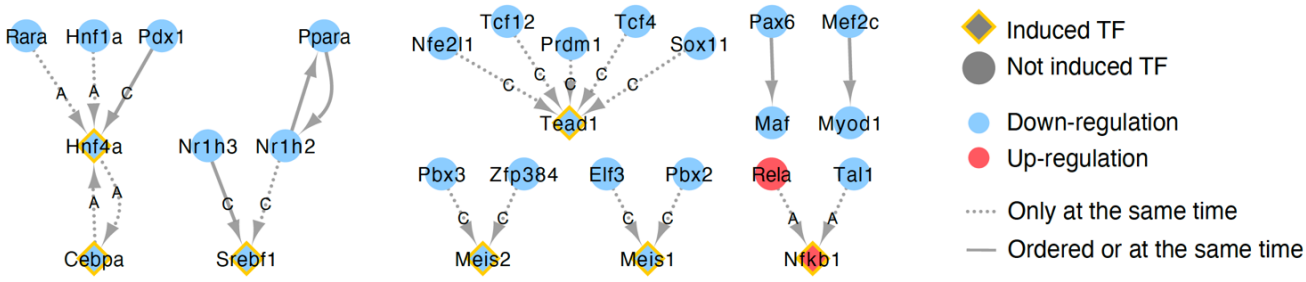


Figure 3.11: Causal relationships between TFs supported by time concordance.

For TFs which are significantly enriched before or at adverse histopathology, known causal relations are shown in which the upstream event is found before or at the downstream event in at least 20% of adverse cases. For induced TFs for which expression is found before regulon activity and significantly enriched, not only protein-protein interactions are considered but also upstream TF-target gene interactions annotated with DoRothEA [170] confidence scores (A: High confidence, C: Medium confidence).

As an additional larger TF cluster, decreased activity of the hepatocyte nuclear factor 1 (Hnf1a), retinoic acid receptor α (Rara) and pancreatic and duodenal homeobox 1 (Pdx1) was found to lead to decreased expression and activity of hepatocyte nuclear factor 4 (Hnf4a) which is linked to reduced expression and activity of CCAAT/enhancer-binding protein (Cebpa) through edges in both directions. This cluster stands out due to the high confidence score of all interactions except the edge between Pdx1 and Hnf4a indicating that there is strong support based on prior knowledge for the involved interactions. Furthermore, it was previously found that artificially increased expression of Hnf4a is able to reverse hepatic liver failure in rats, while also restoring expression of a highly interconnected TF network including Hnf1a and Cebpa which supports the identified interactions [290,291].

Two of the yet unknown TFs in DILI are meis homeobox 1 (Meis1) and meis homeobox 2 (Meis2) which are generally known in a developmental context [292,293]. However, their down-regulation in early pathogenesis is supported by enriched TF activity, differential expression before adverse histopathology as well as upstream regulators which are also enriched

before adverse histopathology. Consequentially, detailed hypotheses are provided which support their mechanistic role in DILI.

3.3.6 Time-concordant events reflecting disease progression

While events do not have to be activated continuously to be causally involved in pathogenesis, events with consistent or increasing activation over time are particularly interesting as biomarkers as they can be experimentally measured without the chance of missing the timepoint of activation, and can potentially reflect disease progression beyond early pathogenesis. Therefore, TFs and pathways were studied which show time-dependent activation by testing for significant Spearman correlation between activation logFC and time in adverse time-series, as well as the overlap between these and the previously derived time concordant events (Figure 3.12). Overall, 118 pathways and 19 TFs were supported by both, significant time concordance and dependence, which represents 86.1% or 70.4% of the time-concordant events, and 59.9% or 48.7% of the time-dependent events, respectively.

On the pathway level, multiple genesets pointed to a reduced level of plasma lipoprotein particle assembly and remodelling which indicates changes in lipid distribution. This aligns with the known dyslipidaemia in chronic liver diseases, including decreasing serum values of LDL, HDL, total cholesterol, and triglycerides with increasing severity of disease, based on which previous studies suggested that routine monitoring of lipid profiles can improve the outcome for CLD patients [294]. Furthermore, a down-regulation of response to metal ions was found which could be related to metallothioneins which protect against oxidative stress and are able to chelate heavy metals [295]. Both directions of dysregulation were previously observed in liver diseases: While a negative correlation with disease progression was found in hepatocellular carcinoma [296], a positive correlation was found in most other liver diseases including acetaminophen-induced liver injury [297]. This indicates that opposite directionality is more plausible based on current literature knowledge, but cannot be fully clarified.

The most time-concordant and -dependent TF event was down-regulation of SRY-box transcription factor 13 (Sox13) which is generally involved in cell fate [298] and embryonal development [299]. As Sox13 does not yet have well understood functions on a more detailed level, experimental validation of a potential role in DILI would be interesting. In contrast, the next most significant time dependence is found for the hepatocyte nuclear factors Hnf1a

and Hnf4a, as well as Cebpa, which are known to negatively correlate with liver cirrhosis in rats [300,301]. Overall, this shows that a mechanistic role for time-concordant and -dependent events is strongly supported by the understanding of adverse liver phenotypes.

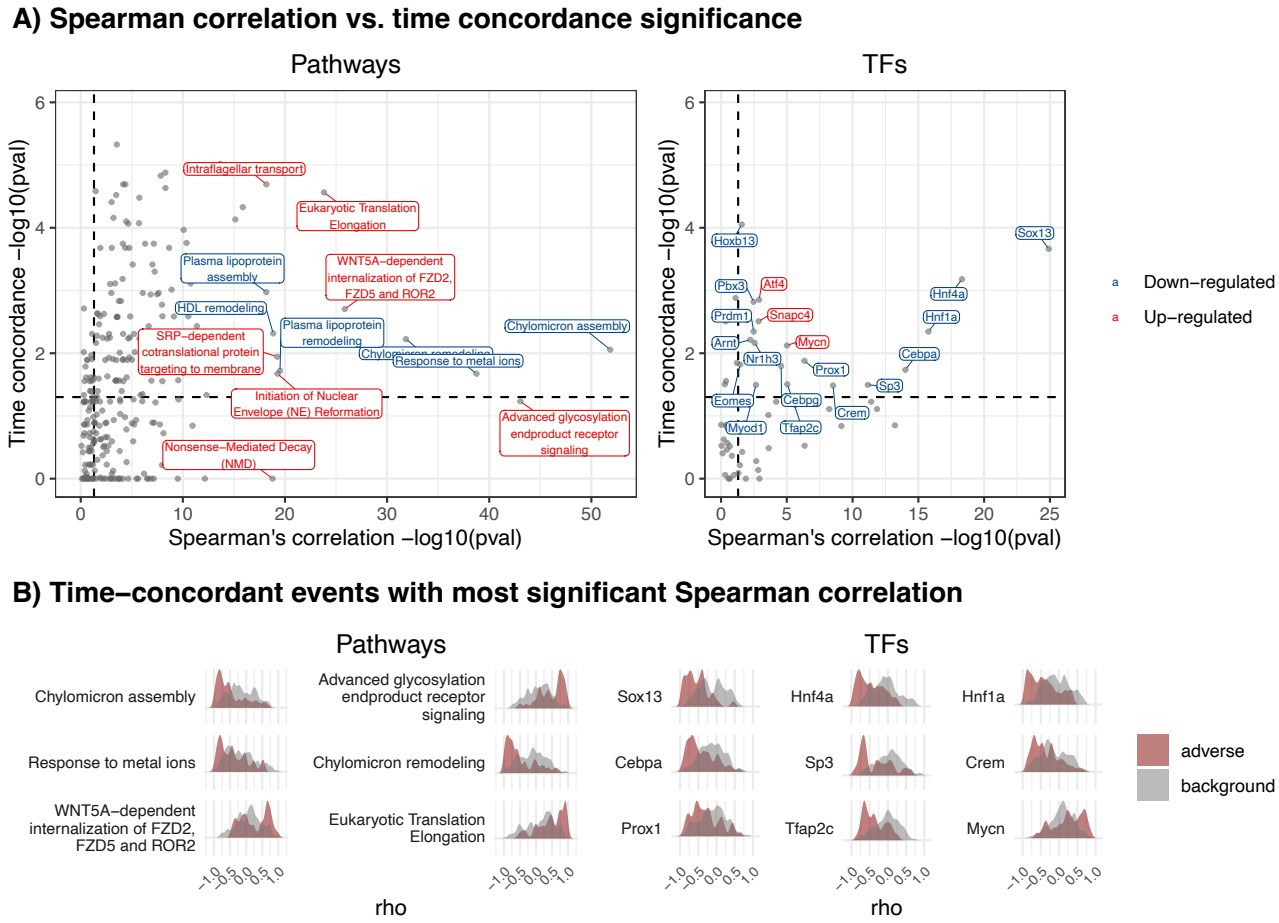


Figure 3.12: Combining time dependence and concordance to identify mechanistically supported biomarkers.

A) The relation between time concordance, quantified by the time concordance p-value for event activation before adverse histopathology, and time dependence, quantified by the meta p-value for Spearman correlation between time and event activation across adverse conditions, is shown. B) For events with the most significant time-dependence, the distribution of correlation coefficients is shown providing further insight into the strength of correlation and consistency across adverse conditions.

While in general events with highly significant time dependence also showed highly significant time concordance, some exceptions were found in which only one of both was highly significant. For instance, the pathway with the 2nd most significant time-dependence (meta p-value for Spearman correlation < 10^{-44}) is signalling via advanced glycosylation end product receptor (RAGE) which contributes to inflammation and oxidative

stress generation and did not pass the significance threshold for time concordance (Time concordance p-value = 0.058). RAGE expression and activity, which are both induced by binding of RAGE ligands, are thereby known to be up-regulated in various hepatic disorders resulting in a positive feedback loop explaining increasing or sustained RAGE activation [302]. This indicates that, while RAGE signalling is correlated with progression, there is no clear evidence for a role in early pathogenesis preceding adverse histopathological changes. In contrast, SUMOylation of TFs, is time-concordant (Time concordance p-value = 0.002) but not -dependent (Meta p-value for Spearman correlation p-value = 0.48) indicating a mechanistic role in early pathogenesis which is not sustained over time. This aligns with the finely regulated and pleiotropic roles of SUMOylation in post-transcriptional regulation which have also been found to be involved in the context of liver diseases [303].

3.4 Limitations of this study

A time-concordance based approach was introduced to derive mechanistic insight from gene expression and histopathology data. Known mechanisms in DILI were recovered and it was also possible to propose potentially novel and detailed mechanistic hypotheses. However, the present analysis is based on a limited number time-series as well as only few timepoints within each time-series. This does not only mean that rare events might be missed as they occur between measured timepoints and that small effects might not be identified as significant, but also that there is potentially a bias based on the tested compounds towards the represented modes of toxicity.

Furthermore, the analysis is limited by how confidently biological processes are inferred from the data. This was for instance demonstrated by the differences between pathway and TF activation for signalling and stress response pathways highlighting the discrepancy between protein activation and gene expression. As only pathways induced through changes in gene expression or their downstream expression footprints [170] can be confidently detected, this means that good estimates of time concordance can predominantly be derived for intermediate or later key events while preceding key events or molecular initiating events which are not mediated by transcriptional regulation cannot be estimated based on the data. Additionally, it should be noted that all measured genes were used in GSVA. As GSVA approximates the variance of the gene set across samples based on the variance of all of

its gene members, this may underestimate the dysregulation for gene sets in which some members are not expressed and hence invariant.

Moreover, multiple choices were made to align our analysis to the AOP concept prioritizing mechanisms supported by prior knowledge over purely data-driven hypothesis. First, detailed insights might be lost by summarising results to the pathway level. While generally measurements for individual genes can be noisy, this can be summarised in different ways e.g. based on similarity in expression profiles [242]. In this study, however, curated gene sets were used due to their interpretability and to derive modular events as defined in the AOP framework. Additionally, prior knowledge was taken as ground truth, both in the gene set and interaction analysis, meaning that only generally known pathways and interactions could be discovered. Like all methods based on curated gene set and interactions, it was hence informed and biased by the current understanding of biology. However, this prior biological knowledge contributes to the biological plausibility of the derived events and relationships contributing to the weight of evidence of our findings in the context of AOPs.

Lastly, it should be highlighted that time concordance is necessary for causal relations but not sufficient to prove it. For instance, two events may be time-concordant because they are causally linked to a shared preceding cause. To distinguish these effects, the additional Bradford-Hill considerations can be helpful, but only prior knowledge has been considered in parts of this study. In particular essentiality would provide strong evidence for causality, however, requires targeted experiments and hence is unsuitable for hypothesis generation. In contrast, dose and incidence concordance are generally feasible from a data-driven standpoint but were not pursued in this case study due to the low number of doses and replicates.

3.5 Conclusion

In this study, “first activation” was introduced as concept to quantify the strength of temporal concordance between events across time-series with the assumption that each activated event may have downstream effects irrespective of whether it is continuously or only transiently activated. With this approach, gene expression-based TF and pathway-level events were studied which are found before adverse histopathology indicating liver injury in repeat-dose studies in rats from Open TG-GATEs. Some known processes in DILI were

found to be highly confident, e.g. bile acid recycling, while others are highly frequent but less specific including adaptive response pathways such as the eIF2α/ATF4 pathway [276].

Beyond quantifying time concordance for known and potentially novel events in DILI, it is additionally demonstrated how time concordance can be combined with prior biological knowledge to generate hypothesis on potentially causal gene-regulatory cascades in DILI. Amongst others, this identifies LXRα down-regulation leading to decreased Srebf1 expression, an interaction known to regulate fatty acid synthesis in the liver [266], but also characterizes yet unknown TFs based on their time concordance, their mode of regulation (either transcriptional or post-transcriptional) and potential upstream regulators and downstream effectors. Two of the identified induced TFs are Meis1 and Meis2 which is supported by significantly enriched decrease in expression and activity before adverse histopathology, as well as upstream regulators which also show significant enrichment of regulon activity and are found within the same time-series. On top of time concordance, also each event's time dependence was computed showing that events mechanistically involved in early pathogenesis do not necessarily reflect disease progression and vice versa. However, for some events, e.g. Sox13, both properties are found and these may be useful biomarkers which reflect injury progression and already change preceding histopathological manifestation.

We believe that the described analysis can provide supporting evidence for mechanistic links between events in line with the evolved Bradford-Hill considerations on time concordance and biological plausibility and can hence e.g. support AOP development. Furthermore, the approach is not limited to a particular adverse event and can instead quantify the interaction between any two events represented in time-series in a data-driven and automatable fashion. Consequentially, this type of analysis could also be of interest to study the mechanism of action particular compound classes or patterns of disease progression.

4 DILI Cascades: A web app to study time concordance in the TG-GATEs liver data

This work builds on the study published as research article in PLOS Computational Biology [240].

4.1 Introduction

In the preceding chapter, a new approach to derive time-concordant and hence potentially mechanistically relevant relationships between gene-expression-derived cellular events and adverse histopathology is presented. While this largely focussed on the newly established methodology, also the analysis on the Open TG-GATEs rat liver data and the established workflow can benefit the DILI community beyond the scope of the results discussed. Therefore, `DILI Cascades`, an open-source Shiny app was developed which provides a Graphical User Interface (GUI) enabling researchers to explore evidence for time concordance in the Open TG-GATEs data without requiring programming experience.

Firstly, researchers may be interested in particular definitions of adverse and background histopathology, which may prioritize other time-concordant cellular processes than the definition used originally, which combined multiple histopathological findings at different severity cut-offs. In the app, the previously introduced time concordance metrics, as well as additional ones, can be directly computed and interrogated through an interactive table as well as visualizations.

Secondly, researchers may want to evaluate evidence for a specific potential key event relationship across all tested compounds, e.g. in case a specific hypothesis was generated experimentally using only a few compounds or also using time concordance analysis. In the app, the relation can be explored for two events of interest, or two combinations of events of the same event type, by querying across which compounds each of the events is observed and, if both are observed, with which temporal relation.

In this chapter, technical implementation of the app is described and its application is then demonstrated using a case study on events preceding fibrosis (at any severity score) while light inflammation (toxscore ≤ 0.67) is regarded as background.

4.2 Implementation

DILI Cascades was written in R 4.1.2 [221] using the `shiny` and `tidyverse` [222,223] frameworks. The layout of the user interface was further refined with `shinyWidgets` [304], `shinyBS` [305], and `shinycssloaders` [306]. The `plotly` [307], `ggiraph` [308], and `shinyHeatmaply` [309] are utilized to generate interactive figures, and the `DT` [310] R package provides an interface to the JavaScript library DataTables, enabling sorting, filtering and other useful features to interact with tables.

The app can be accessed in two ways. Firstly, it is deployed at ShinyApps and can be directly accessed via `https://anikaliu.shinyapps.io/dili_cascades` without the need for any programming skills. Secondly, it can be locally deployed from the open-source Github repository under a GNU General Public License v3.0 (`https://github.com/anikaliu/DILICascades_App`). All R packages needed to run the app, including their versions, are documented using the `renv` R package and can be installed by the user through a single command (`renv::restore()`).

For the time concordance analysis, multiple input parameters (Table 4.1) are needed which can be provided by the user through the GUI. Both adverse and background histopathology can be defined as a combination of the histopathology labels, as introduced in 3.2.2, which describe the annotated histopathological findings at multiple toxscore cut-offs: “null” (toxscore > 0), “low” (toxscore > 0.67) and “high” (toxscore >1.34). Time-series without any histopathology are always regarded as background so the definition of additionally tolerated background histopathology is optional. As additional parameters, the user can define whether only events taking place before the timepoint of activation are considered or if activations at the same time are also included, and the event type of interest can be selected to show either the results for histopathology-derived events, TF or pathway activations (Table 4.1).

Table 4.1: User-defined input parameters to compute time concordance metrics.

Object	Description
Adverse histopathology	Histopathology which is regarded as adverse
Background histopathology	Histopathology which is regarded as background
Temporal filter	Either only events “before” or “before or at” the time of adversity
Event type	Either events of the type “TF”, “Pathway” or “Histopathology”
Significance threshold	Time concordance p-value cut-off applied to filter out insignificant events

The time concordance metrics are then computed with the workflow outlined in Figure 4.1. based on the user-defined parameters and the pre-computed timepoints of first activation, as described in 3.2.3, across all time-series and events for each event type. From the therein included histopathology data, adverse and background time-series are identified using the user's definition of adverse and background histopathology, and subsequently the frequency of each event of the user-selected event type with the desired temporal relation is computed (Figure 4.1). This is then used to generate a temporally filtered contingency table for each event, which describes how frequently the preceding event (***PE***) is either found or not found before the later event (***LE***) and in background time-series without ***LE*** at any time.

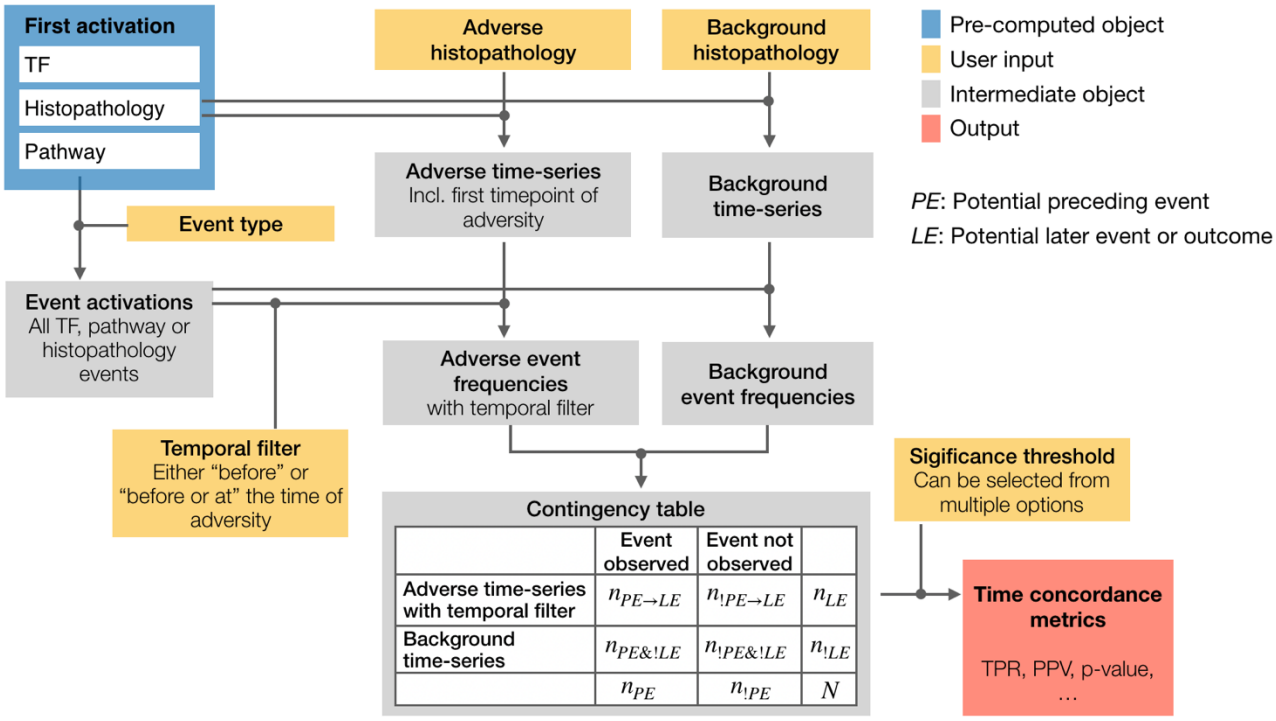


Figure 4.1: Workflow for computing time concordance metrics.

Using the user's definition of adverse and background histopathology, adverse and background time-series are identified. For the selected event type, pre-computed timepoints of first activation are derived, and subsequently, event frequencies of the two time-series sets are identified. Thereby, activation can take place at any time in the background time-series but needs to fulfil the temporal filter for adverse time-series. The event frequencies then serve as the foundation for the confusion matrix from which the time concordance metrics are derived.

From the contingency table, multiple measures of the temporal association between ***PE*** and ***LE*** are derived (Table 4.2). This includes the true positive rate (TPR), also known as recall,

the positive predictive value (PPV), also known as confidence or precision, and the time concordance p-value already introduced in our previous study. Furthermore, additional interestingness measures are included which are commonly used in rule mining [311] and these will be introduced here in the context of the temporally filtered contingency table: The odds ratio (OR) is defined as the ratio of the odds that ***PE*** is found (before ***LE***) given that the ***LE*** is present, and the odds that ***PE*** is found (at any time) given that the ***LE*** is absent; Lift, originally referred to as interest [312], quantifies how many times more ***PE*** is observed before ***LE*** than expected under the assumption that both events are independent; and the Jaccard similarity describes the ratio of intersection over union of ***PE*** and ***LE***.

Table 4.2: Time concordance metrics.

Multiple association metrics are computed based on the event frequencies in the temporally filtered contingency table. The Odds Ratio (OR) and time concordance p-value were computed using the *fisher.test* function of the `stats` package, indicated by "*". This provides the conditional maximum likelihood estimate for the Odds Ratio and the time concordance p-value describes the likelihood that the random variable $C_{PE,LE}$ is larger or equal to $n_{PE \to LE}$ for $C_{PE,LE} \sim Hypergeometric(N, n_{PE}, n_{LE})$.

Interest measure	Formula	Range
True positives	$TP = n_{PE \to LE}$	$[0, N]$
False positives	$FP = n_{PE\&!LE}$	$[0, N]$
True positive rate	$TPR = \frac{n_{PE \to LE}}{n_{LE}}$	$[0,1]$
False positive rate	$FPR = \frac{n_{PE\&!LE}}{n_{!LE}}$	$[0,1]$
Positive predictive value	$PPV = \frac{n_{PE \to LE}}{n_{PE}}$	$[0,1]$
Odds Ratio*	$OR = \frac{n_{PE \to LE} / n_{!PE \to LE}}{n_{PE\&!LE} / n_{!PE\&!LE}}$	$[0, \infty]$
Time concordance p-value*	$p - value = p(C_{PE,LE} \geq n_{PE \to LE})$ $C_{PE,LE} \sim Hypergeometric(N, n_{PE}, n_{LE}) \Longrightarrow p(C_{PE,LE}) = \frac{\binom{n_{PE}}{C_{PE,LE}} * \binom{N - n_{PE}}{n_{LE} - C_{PE,LE}}}{\binom{N}{n_{LE}}}$	$[0,1]$
Lift	$Lift = \frac{n_{PE \to LE}}{n_{PE} * n_{LE}} * N$	$[0, \infty]$
Jaccard Similarity	$Jaccard = \frac{n_{PE \to LE}}{(N - n_{!PE\&!LE})}$	$[0,1]$

Besides information solely based on the temporally filtered confusion matrix, the max. logFC observed before adversity is provided as an additional column in the time concordance table as this was previously found to be an additional useful metric to prioritize time-concordant events (Figure 3.6). To obtain this, significant TF and pathway activations were pre-

computed and it should be noted that the timepoint of earliest activation may not be the timepoint of highest activation preceding adversity.

The resulting table can then be analysed within the app allowing users to search for specific events, to set limits for specific time concordance metrics, and to export the information for further evaluation outside of the app. In our case study, filtering by time concordance significance was used as a filter before prioritizing events. To enable users to explore all time-concordance results but also to highlight this filter, an option was included in the sidebar to choose commonly used p-value cut-offs (on top of the option to filter the table interactively) with no filtering as default.

4.3 Case study

To give an overview of the app's functionalities, a case study is presented in which specifically events preceding liver fibrosis were investigated instead of combining multiple adverse histopathology labels as in the preceding chapter. This means that only ten instead of 61 time-series are regarded as adverse, which limits the statistical power but may identify signals which are more specific to fibrosis. The adverse time-series are derived from repeat-dose studies with carbon tetrachloride (all doses), allyl alcohol and monocrotaline (middle and high dose), as well as naphthyl isothiocyanate, diclofenac and thioacetamide at the high dose. Furthermore, light inflammation (at the toxscore cut-off "null") is considered as background which adds 37 time-series as background in which only light inflammation but not any other histopathological changes are observed. All results and figures presented in the chapter were generated using the app.

4.3.1 Time-concordant pathway events

In the first tab, both background and adverse histopathology can be selected through the sidebar, events taking place before or at the same time as adverse histopathology are regarded as time-concordant and additionally a p-value cut-off of 0.05 was selected (Figure 4.2). Furthermore, it is possible to load the definitions implemented in our original analysis [313], either to explore these results further or simply as an example. With the provided settings, time concordance metrics are provided in the form of a summary table which can be filtered and sorted interactively (Figure 4.2).

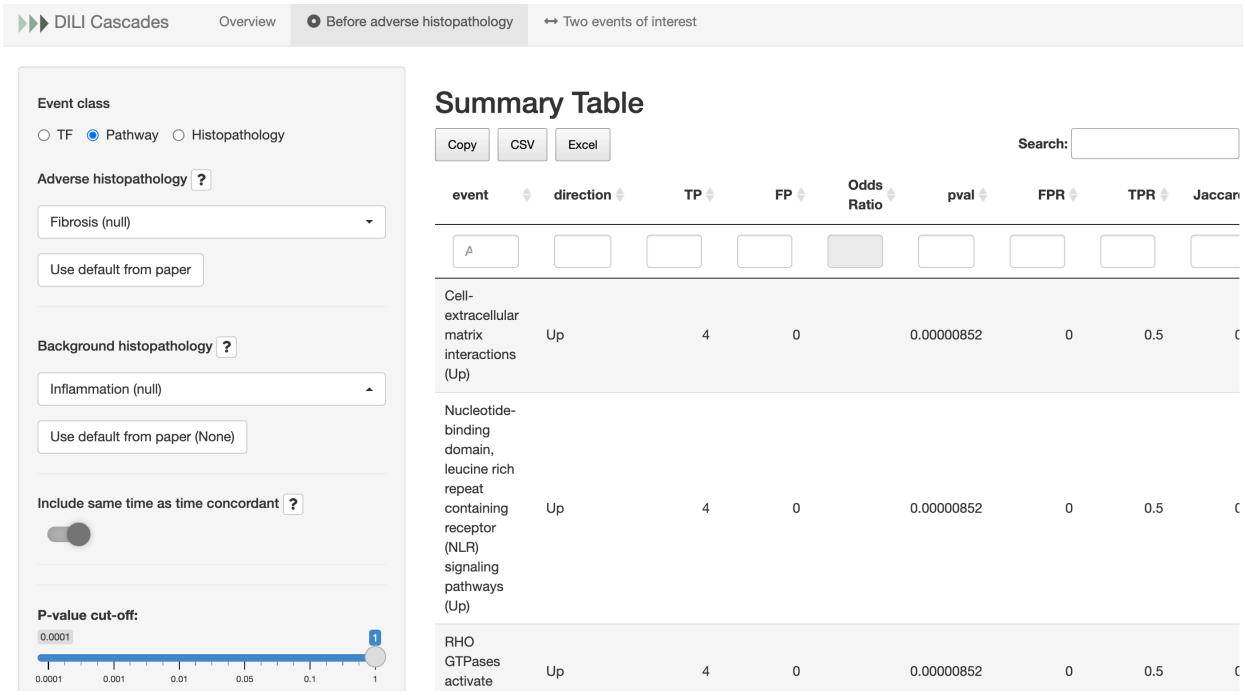


Figure 4.2: DILI Cascades GUI to derive time concordance for user-defined adverse and background histopathology.

Time concordance metrics, introduced in Table 4.2, can be computed for any event class of interest, and any definition of adverse and background histopathology. Furthermore, the temporal filter can be set to only regard events at an earlier timepoint as time-concordant, or to additionally include events activated at the same time.

As the analysis of TF- and pathway-level events is implemented in the same way, the focus will only be on pathways, but the most significantly enriched TFs are provided as supplementary information (Table C.1). The five most significantly enriched pathways, shown in Table 4.3, are observed across four adverse conditions and not in the background time-series, and point to known processes in liver fibrosis such as increased cell-ECM interactions [314] and smooth muscle contraction which is similar to the contractile phenotype of activated myofibroblasts [315]. These myofibroblasts can originate from multiple cell types, in the liver predominantly from hepatic stellate cells (HSCs), and take in a key role in hepatic wound healing and fibrosis [316,317]. Furthermore, nucleotide-binding domain leucine-rich repeat containing (NLR) signalling is identified, which has been linked to HSC activation [318], and as additional most significantly time-concordant pathways, activation of kinectin (KTN1) and p21-activated kinases (PAKs) via RHO GTPases is identified which contributes to cell migration and also has been linked to HSC activation and hepatic fibrosis. As the next most significant (and most frequent) pathways, gene sets related to cell cycle checkpoints are

found and related gene sets have also been identified in our previous study (Figure 3.9). Overall, this shows that while some pathways overlap with the general processes in liver injury also previously identified using a broader definition of adversity, the most significant (and specific) pathways are more specifically linked to pathways in hepatic fibrosis.

Table 4.3: Ten most significantly enriched pathway events before or at the time of liver fibrosis.

Event	TP	FP	OR	p-value	Lift	Jaccard	TPR	PPV	logFC
Cell-extracellular matrix interactions (Up)	4	0	-	8.52E-06	15	0.5	0.5	1	0.583
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways (Up)	4	0	-	8.52E-06	15	0.5	0.5	1	0.451
RHO GTPases activate KTN1 (Up)	4	0	-	8.52E-06	15	0.5	0.5	1	0.44
RHO GTPases activate PAKs (Up)	4	0	-	8.52E-06	15	0.5	0.5	1	0.6
Smooth Muscle Contraction (Up)	4	0	-	8.52E-06	15	0.5	0.5	1	0.436
G2/M Checkpoints (Up)	6	7	41.4	1.18E-05	6.92	0.4	0.75	0.462	0.586
Mitotic G1 phase and G1/S transition (Up)	6	7	41.4	1.18E-05	6.92	0.4	0.75	0.462	0.584
Regulation of localization of FOXO transcription factors (Up)	5	3	53.4	1.54E-05	9.38	0.455	0.625	0.625	0.513
Response of Mtb to phagocytosis (Up)	5	3	53.4	1.54E-05	9.38	0.455	0.625	0.625	0.559
HIV Infection (Up)	6	8	36.1	2.03E-05	6.43	0.375	0.75	0.429	0.529

In the app, the dependencies between selected metrics are visualized which helps to understand how these are related to each other and how frequently each combination of metrics is found. For instance, this shows that positive predictive value (PPV) and lift are directly related which can be explained by the constant number of adverse time-series n_{LE} and total time-series N (Figure 4.3).

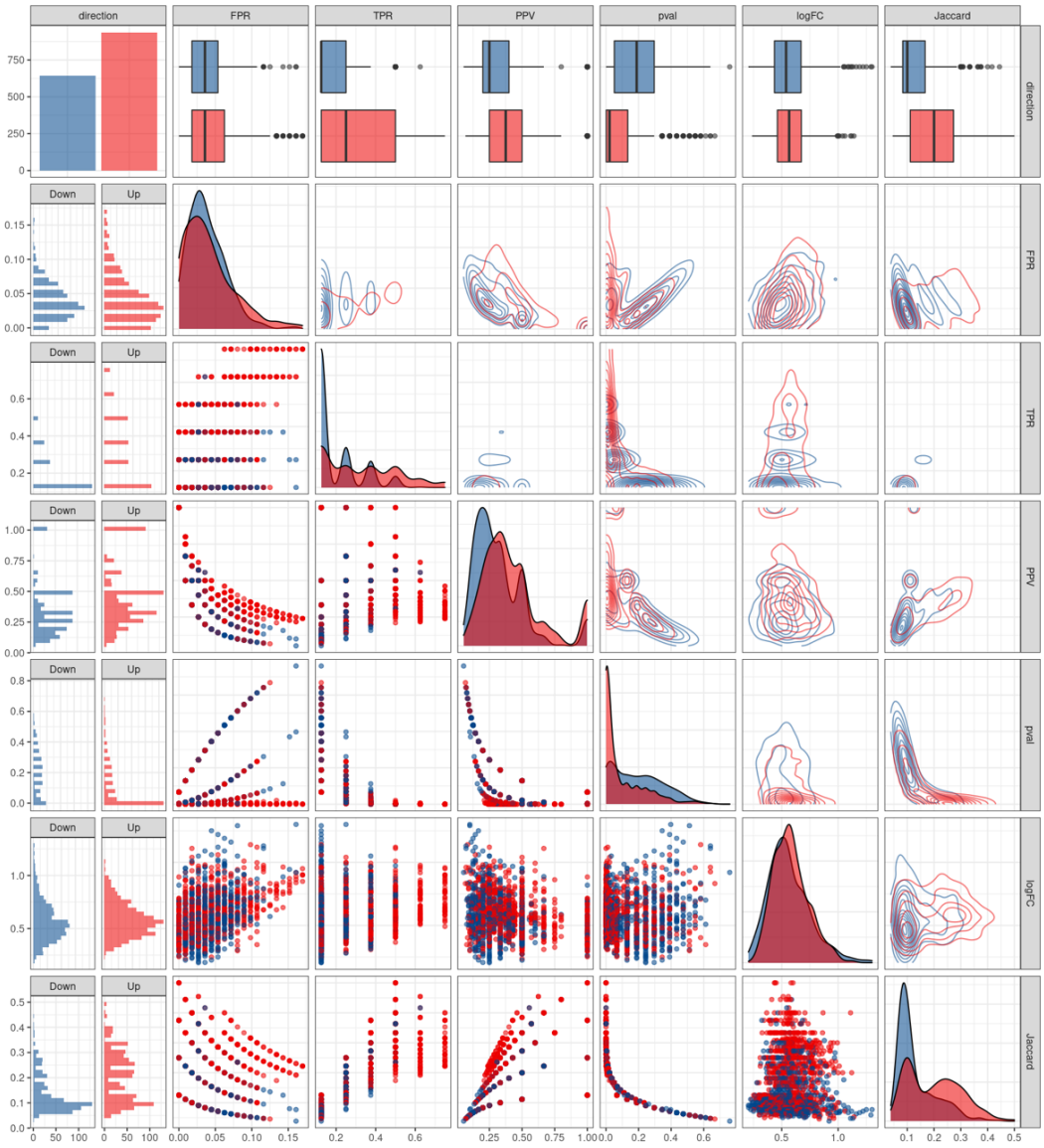


Figure 4.3: Distributions of and dependencies between time concordance metrics.

The diagonal shows the density distribution for each metric individually and the upper triangle the density distribution for each combination of metrics. Furthermore, the lower triangle shows the observed combinations as scatter plots revealing dependencies between the metrics. Which metrics are plotted can be selected by the user in the app.

Besides this rather global view on the distribution of time concordance metrics, it is possible to interactively investigate specific metrics and events in more detail as demonstrated in Figure 4.4. In this tab, two scatterplots are shown: One with PPV and TPR as axes and one where both axes can be specified by the user and can be log-transformed which is particularly useful to inspect time concordance p-value distributions as it visually highlights differences between the small p-values. By hovering over the plots, the event name and their location in the respective other scatter plot will be indicated which further helps to understand the time concordance of individual events. For instance, the labelled "Rho GTPases activate PAKs (Up)" shows the highest logFC among the events with the same level of significance by time concordance p-value, but at lower significance, larger logFCs are observed.

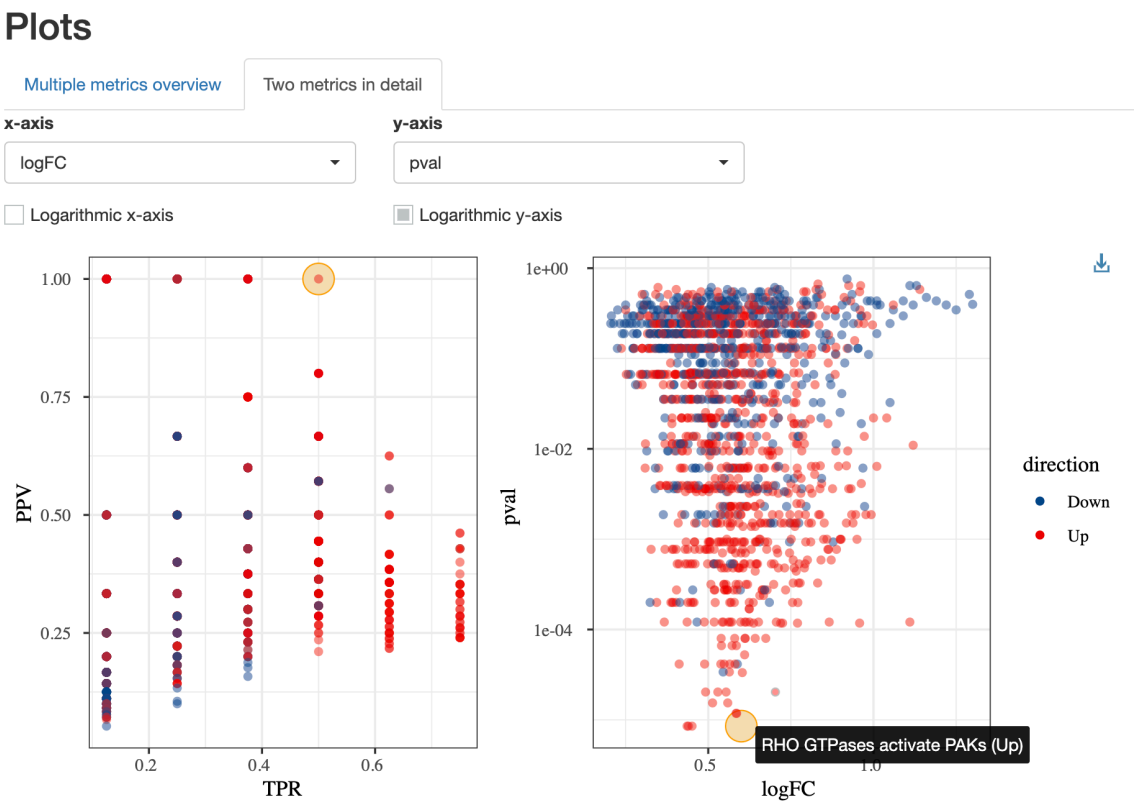


Figure 4.4: Interactive exploration of the time dependence metrics.

Two scatter plots enable a more detailed exploration of the time dependence metrics and their relationships. While the axes in the left scatter plot are fixed to show the True Positive Rate (TPR) and Positive Predictive Value (PPV) of events, the axes of the scatter plot on the right can be set by the user.

To gain more insights on the temporal relation between two specific events, e.g. “Rho GTPases activate PAKs (Up)” and fibrosis, the 2nd tab of the app (Figure 4.5) can be used. Generally, both the preceding event and later event can be described by multiple events of the same class, e.g. enabling the combination of multiple pathways with shared functionality or multiple histopathology labels. Similar to the previous tab, the originally used adverse histopathology definition can be loaded as later event through a button. Based on the known activations across all time-series, a table is then generated which summarises how frequently only either the preceding or later event is observed and, if both are present within the same time-series, in which temporal order they are found (Figure 4.5), as well as multiple visualizations.

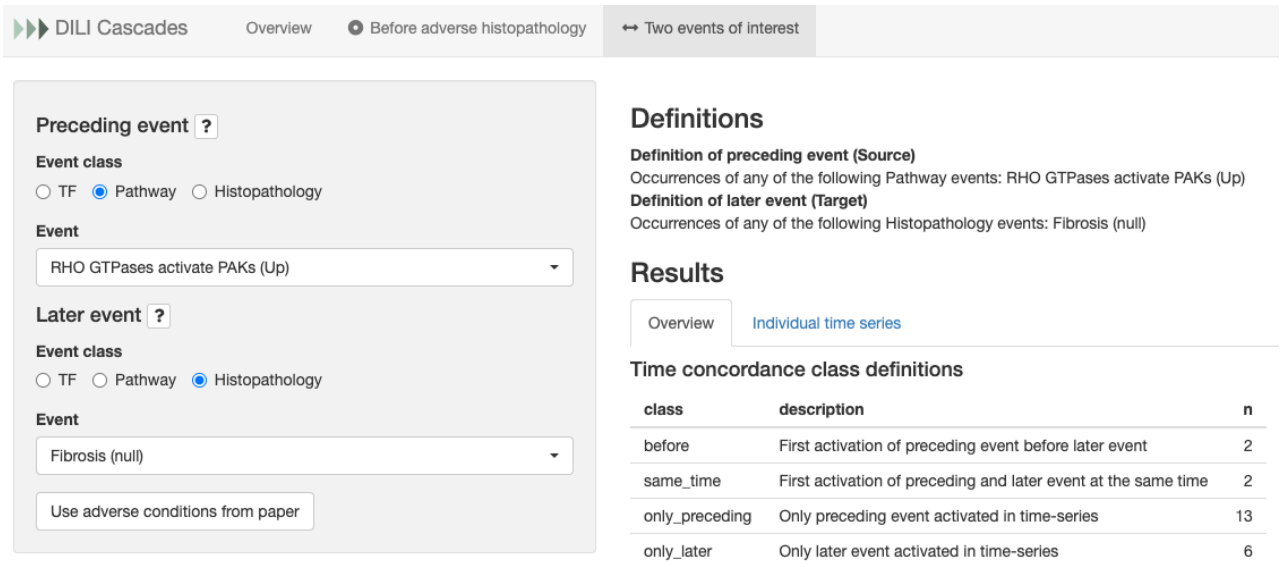


Figure 4.5: DILI Cascades GUI to derive details on time concordance for two events of interest.

For two events of interest, which can be from any event class and also can be combination of events, details on the temporal relation observed in the Open TG-GATEs database are shown.

In our case study, these for example show that that our events of interest are not highly frequent in comparison to other events of the same class. Furthermore, the distribution of the logFC associated with significant dysregulation of the preceding event is shown in time-series with or without the later event with additional subgroups for time-series with the later event indicating the max. logFC per condition before, at the same time, and after the later event (Figure 4.6).

It is generally found that the Rho GTPase-mediated activation of PAKs is increasing over time in fibrotic conditions, but overall not larger in magnitude than in events without fibrosis. Hovering over an individual points reveals that the shown logFC was observed in the time-series describing response to carbon tetrachloride treatment at the middle dose, and also highlights other data points from the same time-series.

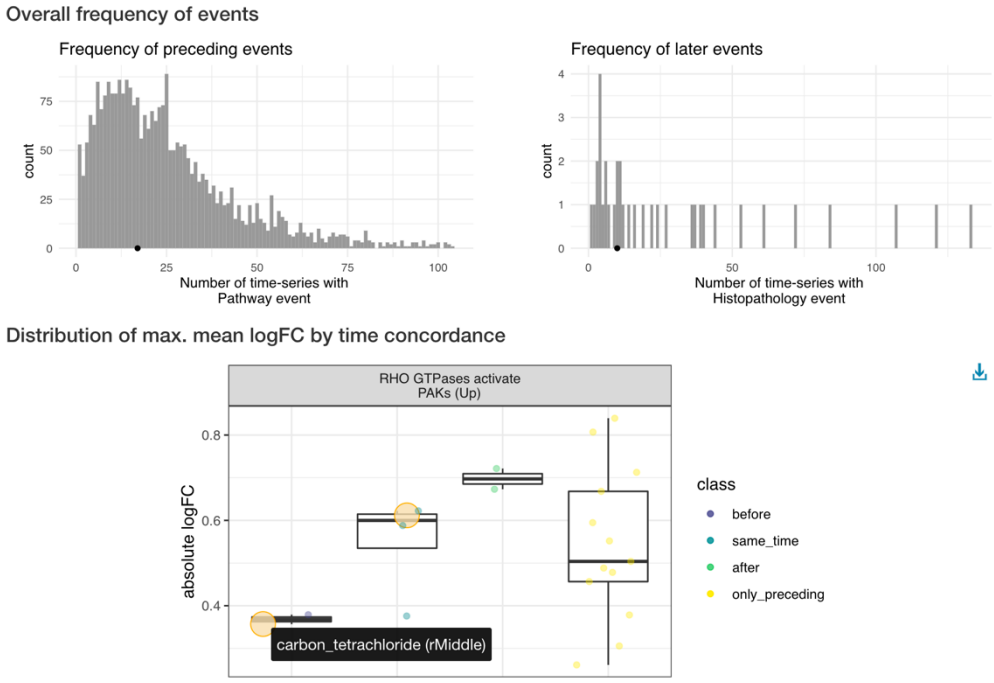


Figure 4.6: Overview figures on temporal relation between two events of interest.

The two upper histograms show the frequency of the preceding and later event in the TG-GATEs database and the overall distribution of frequencies for events of the same event type. The lower plot shows the logFC of the preceding events before, at the same time, and after the later event, as well as the distribution in time-series where only the preceding event was found. Interactively hovering over a datapoint reveals the time-series as text label and highlights other data points from the same time-series.

It is furthermore possible to summarise the individual time-series in which the preceding and later event are observed as a data table, and to visualise this as an interactive heatmap, where the timepoint of activation is shown for preceding and later event, and in which time-series are sorted based on the temporal relationship between both. In our case study, this for example shows that Rho GTPase-mediated activation of PAKs is found after 4 days in three of the adverse time-series and after 15 days in one, while it is never observed only after the observation of fibrosis (Figure 4.7).

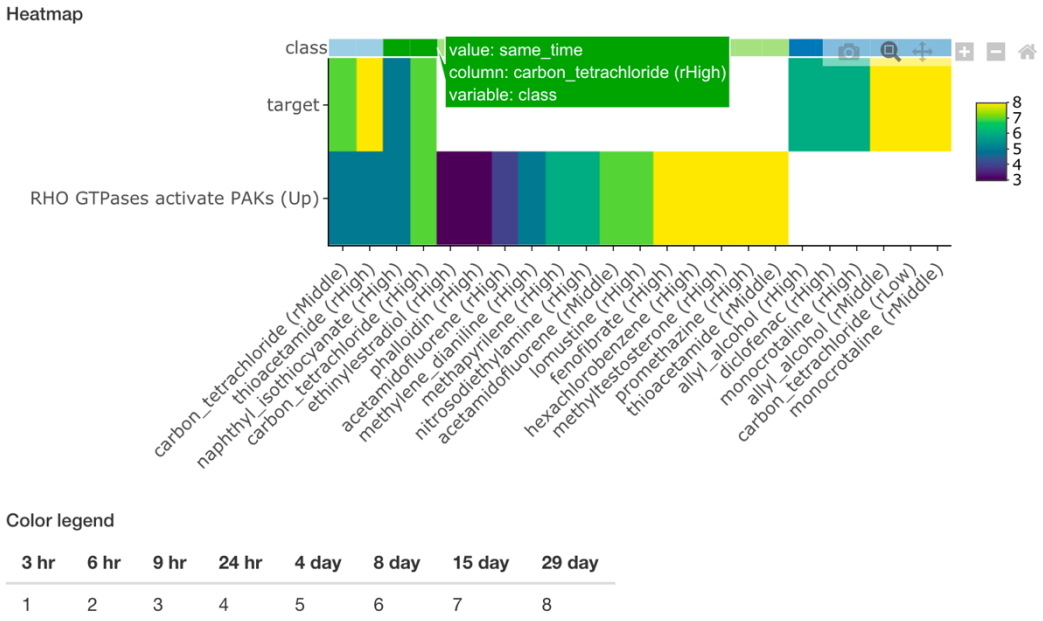


Figure 4.7: Co-occurrence and time of activation for Rho GTPase-mediated activation of PAKs and fibrosis.

4.3.2 Time-concordant histopathology events

As for TF- and pathway-level, time concordance analysis is also possible for histopathology events for all metrics except the logFC as the extent of histopathology has instead been encoded in separate labels characterised by the three toxscore cut-offs. Hepatocellular necrosis (at a toxscore > 0) followed by inflammation (at a toxscore > 0.67) are found to most frequently and significantly precede fibrosis (Table 4.4) and both are indeed generally known to precede fibrosis [319,320].

While information on the temporal order of histopathology events can be useful to better understand the pathogenesis, it should be noted that the definition of adverse and background time-series is based on histopathology which means that the contingency table is fundamentally biased which will be further elaborated. Firstly, unless specifically stated otherwise, histopathology events are not found in background time-series resulting in the absence of false positives while the overall frequencies of histopathology are not considered. For instance, hepatocellular necrosis (null) is found to most frequently precede fibrosis (Table 4.4), but is also overall the 2nd most frequent histopathological finding, after inflammation (null) which was considered as background, which means that numerically it

has hence a higher chance of co-occurring in the same time-series at random than rarer histological findings).

Table 4.4: Ten most significantly enriched histopathology events before or at the time of liver fibrosis.

The concurrently most significant and frequent histopathology events are shown for both possible temporal filters, "before" or "before or at".

Event	Before or at the same time			Before		
	TP	p-value	TPR	TP	p-value	TPR
Fibrosis (null)	10	1.93E-10	1	0	-	-
Hepatocellular Necrosis (null)	8	1.43E-07	0.8	7	1.91E-06	0.7
Inflammation (low)	8	1.43E-07	0.8	5	1.64E-04	0.5
Hepatocellular Necrosis (low)	6	1.96E-05	0.6	4	1.18E-03	0.4
Hepatocellular Necrosis (high)	5	1.64E-04	0.5	4	1.18E-03	0.4
Fibrosis (low)	4	1.18E-03	0.4	0	-	-
Inflammation (high)	4	1.18E-03	0.4	2	4.16E-02	0.2
Biliary Hyperplasia (null)	3	7.4E-03	0.3	0	-	-
Vascular Edema (null)	3	7.4E-03	0.3	1	0.213	0.1
Hepatocellular Hypertrophy (high)	2	4.16E-02	0.2	1	0.213	0.1

Secondly, if the temporal filter is set to "before or at the same time", which was used also for the TF- and pathway-level, the histopathology labels used to define adverse time-series are favoured and e.g. the observed TPR of 1 for fibrosis in this case study is a direct consequence of this (Table 4.4). Hence, all time concordance metrics should be treated with caution, which is also indicated in the app as a text warning when histopathology is selected as the event type of interest.

4.4 Conclusion

We have developed a R/Shiny app called DILI Cascades through which time concordance in the Open TG-GATEs data can be explored in more detail and beyond the results discussed in the preceding chapter for user-defined events of interest. The app leverages pre-computed results and the established methodology from our previous work [313], and can be accessed via `https://anikaliu.shinyapps.io/dili_cascades`, or can be deployed locally based on the publicly available open-source GitHub repository (`https://github.com/anikaliu/DILICascades_App`).

Considering fibrosis as adverse histopathology while tolerating mild inflammation (null) as background histopathology as case study, it is demonstrated how the implemented

functions can provide further insight into the time-concordant events, and indeed pathways which are more specific to fibrosis are identified in comparison to the previous chapter which used a more general definition of adverse histopathology. Also, the temporal relationship can be analysed in more detail and hepatocellular necrosis and inflammation are found to most frequently precede fibrosis. While the number of adverse time-series in this case study is small, which overall limits the statistical power of the results and can be a general limitation depending on the definition of adverse histopathology of interest, hence pathways and histopathology are identified which align with the current understanding of fibrosis.

5 scRNA-Seq based drug repurposing targeting idiopathic pulmonary fibrosis (IPF)

This work was published as a pre-print on bioRXiv [321].

5.1 Introduction

Single-cell transcriptomics has fundamentally revolutionised biological research by enabling us to study gene expression at cellular resolution instead of averages across samples. With this technology, introduced in 1.4.2.2, it is now possible to study questions on cellular heterogeneity and compositional changes, and to identify rare cell populations which may be overlooked in bulk transcriptomics [322,323]. Furthermore, transitional cell states and cell trajectories can be uncovered with scRNA-Seq, e.g. in perturbation response or differentiation, overall providing us with a better understanding on how dynamic cellular processes are orchestrated [322].

Single-cell transcriptomics is already advancing our systems-level understanding of diseases [324,325] and treatments [326,327], which will likely impact the development of future therapeutics. However, scRNA-Seq cannot only guide drug development indirectly through mechanistic insights, but also the data itself can potentially be exploited to prioritize compounds in drug discovery. The approach revisited in this study towards this aim is signature matching to identify compounds inducing perturbation responses which match a defined disease signature. This has already been successfully applied to bulk transcriptomics data for drug repurposing despite its simplicity and intrinsic limitations (described in 1.6.3). As scRNA-Seq data describes cellular instead of global transcription, it can partially overcome these limitations enabling an even more successful application of the approach.

Previous studies applying signature matching to scRNA-Seq data follow the idea of reversing the diseased state of individual cell clusters instead of bulk expression, which mitigates compositional changes as covariate [328,329]. Matching results are then available for each cluster, and a final drug ranking can then be obtained, e.g. by giving higher priority to compounds detected in multiple clusters, which was implemented by Wang *et al.* [328], or by

summarising to an overall score, as implemented in ASGARD (A Single-cell Guided pipeline to Aid Repurposing of Drugs) [329].

While the assumption that compounds targeting more cell clusters are more promising is plausible in the absence of further information, we often know which cell types are causally involved in pathogenesis and can use this to direct the analysis. For example, Alakwaa [330] focussed on type II alveolar cells (AT2 cells) in their repurposing study on COVID-19, or more precisely they used the differential expression comparing ACE2-expressing AT2 cells and other AT2 cells as disease signature because ACE2 was already hypothesized to be a receptor for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) at the time of the study [97]. While this hence includes prior information on relevant cell types, the underlying hypothesis that reducing the expression of the target receptor may help to treat COVID-19 is not further elaborated and is not an established strategy to tackle infectious diseases.

Instead, inducing disease-relevant transitions poses a more promising application of signature matching, although not purely data-driven, given that it is then known that 1) the transition can take place which is not always true, e.g. due to lineage commitment [331], and that 2) inducing the transition is affecting pathogenesis and e.g. not merely reversing symptoms (Table 1.1). This has already been successfully applied in previous bulk transcriptomics studies, e.g. osteoblast differentiation has been targeted to identify treatments for osteoporosis [332,333], and candidates for differentiation therapy have been proposed based on the transition signature of leukaemia cells to granulocytes [203]. With bulk transcriptomics, however, it was only possible to characterise the transitions for particular cell types *in vitro*, which limits the physiological relevance, while transitions were difficult to characterize *in vivo* given that expression changes cannot be assigned to particular cell populations and rare cell types may be overlooked. However, with the advent of scRNA-Seq, it is now possible to discover and characterize disease-relevant and rare cell types and transitions *in vivo* which opens up new opportunities, which to our knowledge have not been explored yet.

This study focusses on the transition from a rare disease-enriched intermediate progenitor cell state, recently discovered aided by scRNA-Seq [135,136], to AT1 cells, which are the alveolar epithelial cells involved in gas exchange in the lung. This transition was found to contribute to the regeneration of the alveolar lung tissue upon injury and its induction has since

emerged as a potential target in IPF (and potentially other lung diseases) as outlined in 1.3.3.3.

While additional studies in the bleomycin injury model have already elucidated some signals involved in the AT2 to intermediate progenitor to AT1 differentiation further in mice [22,23,138], specific proteins or pathways with established ability to induce the transition have yet to be uncovered in IPF suspending target-based drug discovery (1.1.1). Using a signature matching-based approach as alternative, perturbation signatures from LINCS were leveraged to prioritize compounds for drug repurposing. As summarised in Figure 5.1 and further described in Methods, transition signatures and consensus compound perturbation signatures were first derived for this and then matched. Additionally including compound bioactivity data from ChEMBL [334], this was further interpreted through the identification of target proteins, and a target miRNA, which may be causally linked to the cell transition.

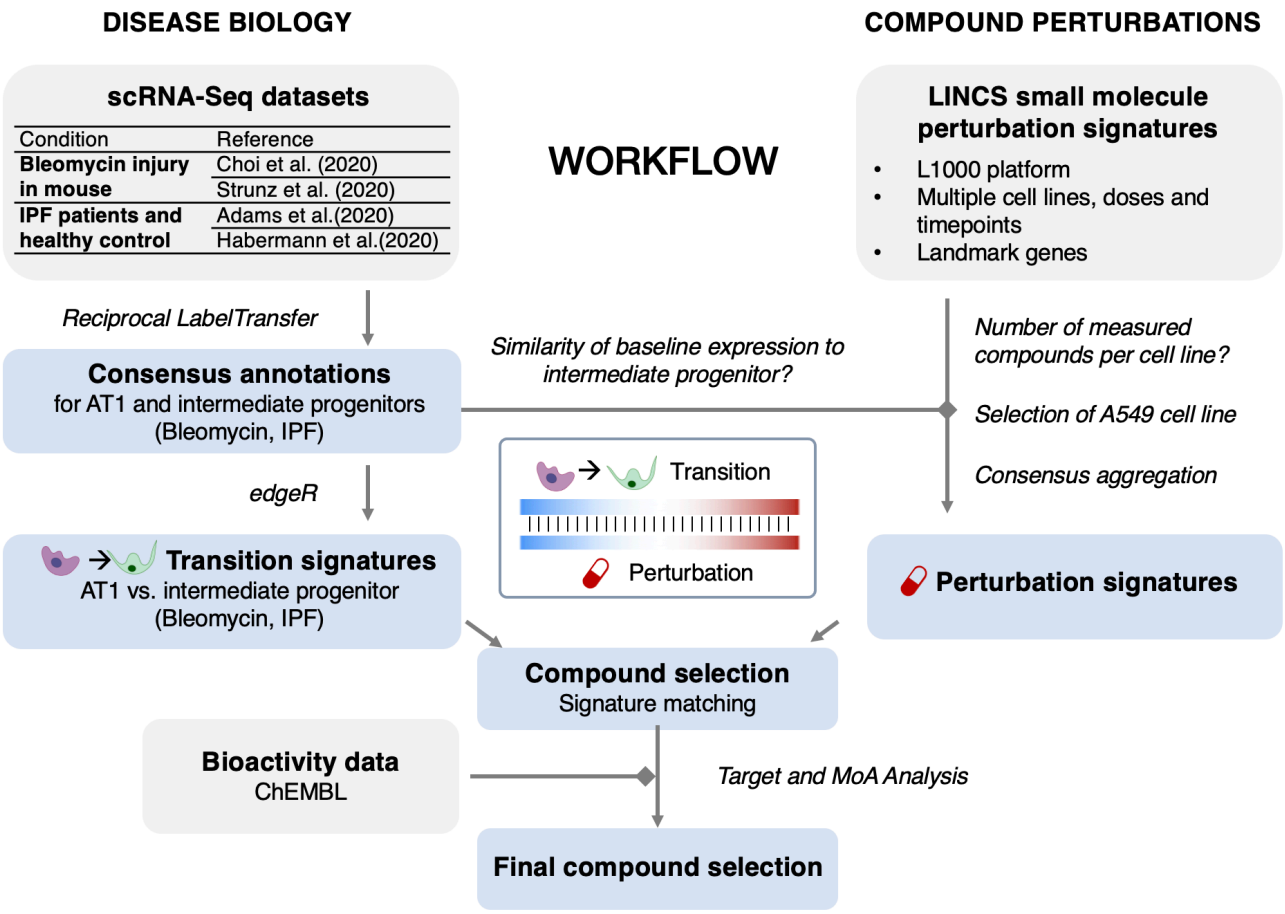


Figure 5.1: General workflow to prioritize drug repurposing candidates.

First, intermediate progenitor to AT1 cell transition signatures were characterized in IPF, as well as in bleomycin-induced pulmonary fibrosis in mice which is a commonly used animal model for IPF. Based on these, perturbations which induce similar transcriptional changes were matched, identifying multiple distinct compound classes. To further interrogate the potential targets and mechanisms, also bioactivity data from ChEMBL was leveraged [334].

5.2 Methods

5.2.1 Deriving transition signatures from scRNA-Seq data

In this study, scRNA-Seq data from two studies on bleomycin injury in mice and two studies in IPF patients was used, respectively (Table 5.1). For these datasets, count matrices and metadata were derived via the GEO FTP server [219] for all datasets except for the metadata by Choi *et al.* [133] which was provided directly by the authors instead (File D.1). Throughout the study, the single-cell data was handled in `Seurat` [335]. The provided count matrices have been pre-filtered based on number of genes, unique molecular identifiers (UMIs), and mitochondrial fraction as outlined in Table D.1. The distribution of these metrics is additionally shown in Figure D.1. From the study by Habermann *et al.* [129], only samples from IPF patients and healthy controls were used although the study covered multiple lung diseases, in order to avoid potential confounding factors.

Table 5.1: scRNA-Seq dataset origins.

Paper	Condition	Cells	GEO ID	Platform
Choi *et al.* [133]	Bleomycin injury in mouse	Alveolar epithelium	GSM4304609, GSM4304611, GSM4304613	10x Genomics Chromium
Strunz *et al.* [132]	Bleomycin injury in mouse	Epithelium	GSE141259	Dropseq
Adams *et al.* [130]	IPF patients and healthy control	Whole lung	GSE136831	10x Genomics Chromium
Habermann *et al.* [129]	Pulmonary fibrosis patients (incl. IPF) and healthy control	Whole lung	GSE135893	10x Genomics Chromium

For each sample with at least 1,000 cells, doublets were removed using `scDblFinder` [336] using default parameters (Figure D.2). This first generated artificial doublets between randomly selected cells based on the 1,000 top expressed genes and subsequently classified cells as doublets if a large number of artificial doublets were identified in its neighbourhood taking into account an expected doublet rate of 1% per thousand cells captured with an uncertainty of 40%. The classifier was retrained three times removing cells labelled as doublets after each iteration. Subsequently, all samples from the same publication were merged, normalized and variance stabilized using `SCTransform` [337]. IPF datasets were subsetted to epithelial cells based on the original annotations as these were distinct from other cell types and are at the centre of this study.

5.2.1.1 Harmonize cell annotations through reciprocal label transfer

To identify a consensus cell annotation for both IPF and both bleomycin datasets, respectively, it was predicted for each dataset, referred to as “query”, whether each cell is labelled as AT1 cell, intermediate progenitor or another cell type based on the cell annotations from the respective other dataset, referred to as “reference” as described by Stuart *et al.* and implemented in `Seurat` [338]. First, pairs of cells were identified between two datasets which are mutual nearest neighbours (MNN) using the *FindTransferAnchors* function. These transfer anchors were among each other’s 5-nearest neighbours based on the PCA reduction of the reference (30 PCs) and its projection onto the query dataset, and were further scored and filtered using default parameters. Subsequently, the *TransferData* function was used to construct a weight matrix which quantifies the association between each query cell and each anchor, which was then multiplied with the anchor classifications to compute the label predictions. Then, only AT1 cells and intermediate progenitors were included in the downstream analysis which were regarded as such based on both the original and the predicted annotation.

Details on annotation concordance were visualized in Figure 5.2 using an integrated UMAP representation for IPF and bleomycin, respectively, derived through `Harmony` [339]. For the bleomycin datasets, the integration (and visualization) included cell types expected to be present in both datasets, namely AT1, AT2 and intermediate progenitor cells, to account for the fact that Choi *et al.* specifically studied cells originating from the AT2 cell lineage. In contrast, all epithelial populations were included for IPF except for ciliated cells which formed distinct clusters in both datasets and are not studied in this work. Overall, good agreement was found between the cross-predicted cell labels and the original annotations with an accuracy of 84.1% and 97.4% for the datasets by Strunz *et al.* and Choi *et al.*, respectively, while an even better agreement was found for the IPF datasets by Adams *et al.* (98.5%) and Habermann *et al.* (99.1%). The confusion matrices in Figure 5.2C additionally show that there are intermediates progenitors which were rather regarded as AT1 by Strunz *et al.* than by Choi *et al.* and vice versa. As such annotation discrepancies can affect how well a joint signature can be derived, only consensus intermediate progenitors and AT1 cells, which were regarded as such based on the original and predicted annotation, were included in the downstream analysis.

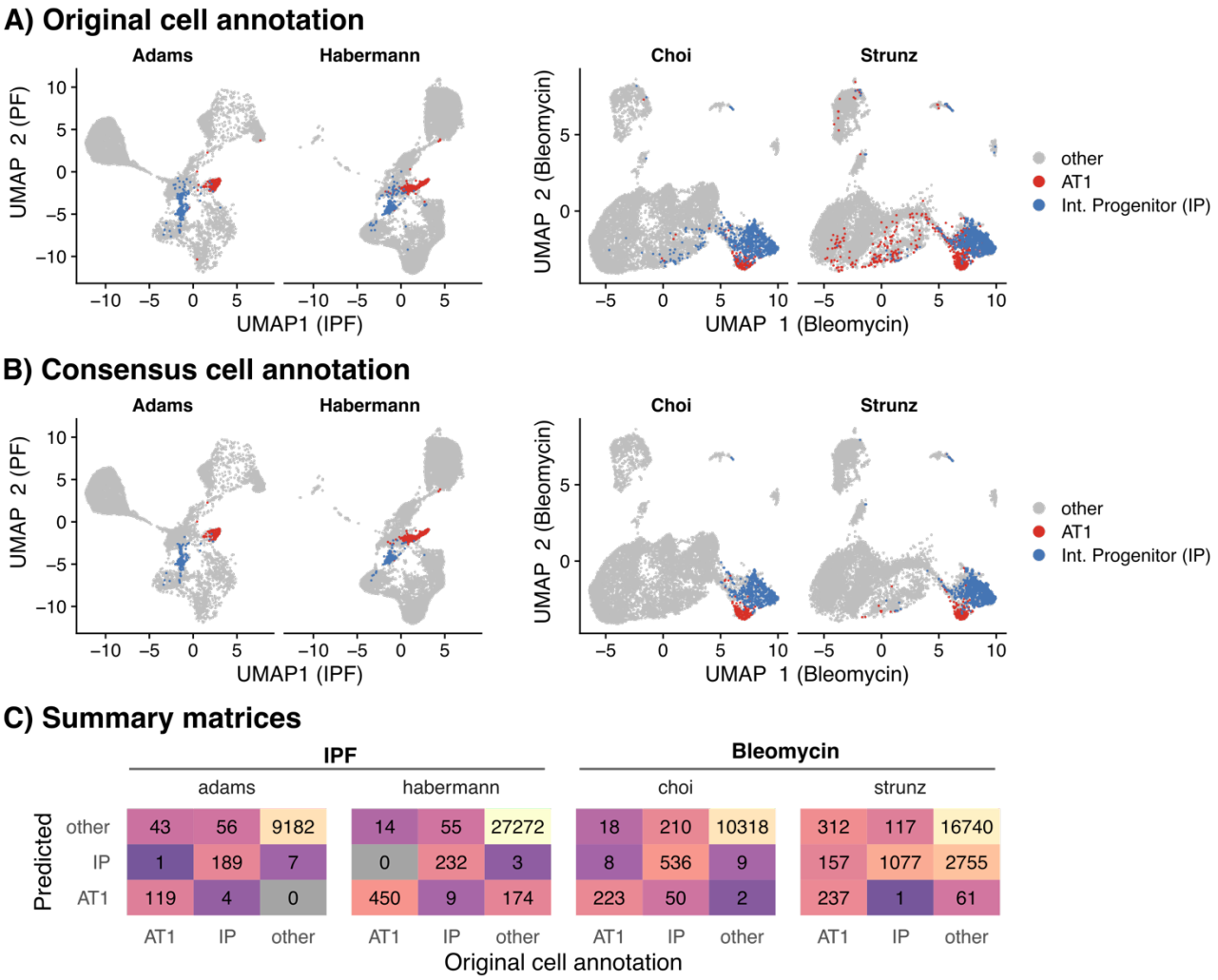


Figure 5.2: Comparison of original and consensus annotations for AT1 cells and intermediate progenitors (IP).

A) Original cell annotations for each dataset plotted onto the UMAP after integration with `Harmony`. IPF datasets were first subset to epithelial non-ciliated cells to improve integration. B) Same UMAP projection with only consensus AT1 and IP cells indicated, which show matching original and predicted cell labels. C) Comparison of original and predicted labels.

5.2.1.2 Differential expression and disease signature analysis

Based on the derived consensus annotations, transition signatures were derived by comparing AT1 cells to intermediate progenitors in IPF or bleomycin injury, respectively, using the `edgeR` [340] implementation of the likelihood ratio test (LRT) as part of the `Libra` package [341]. In this pseudo-bulk approach, individual samples were specified as replicates which translates to individual timepoints of sampling after treatment in case of bleomycin injury or individual patients in case of IPF. For downstream analysis, the derived bleomycin

signature was mapped from mouse gene symbols to HGNC symbols using `biomaRt` [227]. In 1.85% of HGNC symbols, the mapping was not unique and only the most significant differential expression statistics was kept. For Reactome pathway maps [167], derived using `msigdbr` [251], enrichment was computed using the `fgsea` package [342] for all gene sets with 10-200 genes per geneset in order to capture processes which can be statistically inferred and are biologically informative.

Signalling pathway activity was inferred with `PROGENy` based on the expression of the top 100 genes involved in the pathway at hand according to significance [262]. To gain further insights into proteins involved in upstream signalling, TF activity was inferred with the *msviper* function from the `viper` package [343] for all `DoRothEA` [170] regulons with confidence levels A-C for which at least 10 genes were represented in the differential expression signature.

Subsequently, upstream signalling networks were inferred using `CARNIVAL` [261] based on the inferred TF activities as measurements, as well as a signed and directed protein-protein interaction network from Omnipath [252] as prior knowledge. `CPLEX` [344] was used to solve the ILP problem formulated based on the causal reasoning implementation by Melas *et al.* [345] in which each node *j* can take in 3 states, namely -1 (down-regulated), 0 (inactive) and 1 (up-regulated). Across all nodes *j*, node activities ($x_j \in \{-1,1\}$) are penalized except for those where this matches the inferred TF activity (m_j), as described in Equation 5.1.

Equation 5.1: $\mathbf{min}\left(\sum|\boldsymbol{\alpha}| * \left|\boldsymbol{x_j} - \boldsymbol{m_j}\right|\right)$

Thereby, mismatches for nodes with inferred activity were additionally penalized through α and equals the normalised enrichment scores (NES) for inferred TFs, resulting in a higher penalty for the mismatch of more dysregulated TFs. Default parameters for this were loaded using the *defaultCplexCarnivalOptions* function. Additional information on targets and pathway activation was not included, to not bias the inference towards these network regions.

5.2.2 Signature matching

5.2.2.1 Baseline expression

Baseline expression profiles for the cancer cell lines used in LINCS were obtained from the Cancer Cell Line Encyclopedia [346] (CCLE) through the depmap portal (21Q3, DOI: `10.6084/m9.figshare.15160110.v2`). For the intermediate progenitor populations, the baseline expression in TPM (transcripts per million) was determined in `Seurat` based on the consensus labelled cells [335].

5.2.2.2 Deriving perturbation signatures from LINCS

LINCS drug perturbation signatures (Level 5) from the alveolar A549 cell line were obtained from GEO [219] using GEO ID GSE92742. All signatures from the same compound, dose and time combination were combined into a single consensus signature using moderated z-scoring which was first introduced for summarising biological replicates from the same experiment in the LINCS analysis pipeline [143] and was also implemented to summarise multiple signatures from the same condition across experiments by Szalai *et al.* [347].

As the 978 landmark genes are chosen based on complementarity and other genes are inferred based on these, only the measured genes are used for signature matching while all genes are used to infer TF activity using `DoRothEA` [170], to increase the coverage of genes annotated in regulons, using the same settings as for the disease signatures.

5.2.2.3 Signature matching

Pearson correlation between each drug and disease signature was computed using `cTRAP` [348], which provides multiple measures used in signature matching. While significant (p-value < 0.05) and positive Pearson correlation was required for both transition signatures, the magnitude of Pearson correlation was subsequently used to further reduce the list of drug repurposing candidates for experimental validation.

5.2.3 Target bioactivity

To derive bioactivity data on the perturbations from CHEMBL 30 [334], the Python client for the ChEMBL API [349] was used. First, molecules with the same connectivity were identified based on the canonical SMILES representation provided as LINCS metadata, and subsequently all pCHEMBL values for these compounds where the target organism was *Homo sapiens* were retrieved. Besides the target metadata from CHEMBL, additional information on the protein families of the targets was obtained from the Uniprot knowledgebase (UniProtKB) [350], and the UniProt ID was also used to map targets to gene symbols using BioMart [227] via the `biomaRt` R package.

To compare bioactivity across compounds, mean pCHEMBL values were used, which are defined as $-\log_{10}$(molar IC_{50}, XC_{50}, EC_{50}, AC_{50}, K_i, K_d or Potency), and hence enables integration of multiple metrics of half-maximal response [334,351]. To establish enrichment of target activity for matched compounds, pCHEMBL values > 5 were defined as “active” while pCHEMBL values ≤ 5 were regarded as “inactive”. For all targets with at least 20 retrieved pCHEMBL values, a one-sided Fisher’s Exact test, as implemented in the `stats` package, was then used to compare activity in matched compounds to unmatched compounds with perturbation signatures in the A549 cell line while considering all compounds with measured bioactivity as background.

Furthermore, the Pearson correlation between pCHEMBL and the strength of signature matching was computed and tested using the `correlation` R package [254]. In both, the correlation and enrichment analysis, the maximal signature matching correlation per disease signature was used to summarise signature matching across LINCS profiles on the same compound, and only targets with at least 20 data points were included.

Baseline expression for targets was derived as outlined in 5.2.2.1, and genes with a minimal TPM of 0.5 were regarded as expressed as the same cut-off was used in the Expression Atlas [352].

5.3 Results and discussion

5.3.1 Transcriptional characterization of intermediate progenitor to AT1 transition

To characterize the transition signature from intermediate progenitors to AT1 cells, the cell annotations were first harmonized through reciprocal label transfer as described in 5.2.1 and subsequently differential expression signatures were derived comparing AT1 cells to intermediate progenitors (Figure 5.3A and B). In these signatures, up-regulation indicates higher expression levels in AT1 compared to intermediate progenitors and, vice versa, down-regulation indicates higher expression intermediate progenitors compared to AT1 cells.

First, the expression of previously identified markers for AT1 cells and intermediate progenitors was investigated as a quality check for the signatures and markers (Table D.2). In both signatures, up-regulation for all previously identified AT1 markers was found, such as advanced glycosylation end-product specific receptor (*AGER*), podoplanin (*PDPN*), and HOP homeobox (*HOPX*), except for insulin-like growth factor binding protein-2 (*IGFBP2*). Interestingly, *IGFBP2* was previously found to be expressed later in AT1 differentiation than other markers, such as *HOPX*, and at this earlier IGFBP2$^-$ cell state AT1 cells can transdifferentiate to AT2 cells in alveolar regeneration, while IGFBP2$^+$ AT1 cells do not show cellular plasticity physiologically or in injury [353,354]. Hence, this indicates that the IPF transition signature is not capturing this terminal step of AT1 differentiation, potentially because the AT1 IGFBP2$^-$ cells are enriched, because mature IGFBP2$^+$ AT1 cells are destroyed in IPF lungs, or because they have been favoured in the cell annotation in comparison to bleomycin injury.

For the intermediate progenitor markers, decreased expression in both signatures is found for amphiregulin (*AREG*), urokinase plasminogen activator surface receptor (*PLAUR*) and cyclin dependent kinase inhibitor 1A (*CDKN1A*). For four additional markers, significantly decreased expression was only observed in bleomycin injury indicating that these are not applicable to IPF, potentially because of differences between the respective intermediate progenitor populations, e.g. linked to species differences, or their annotations (Table D.2).

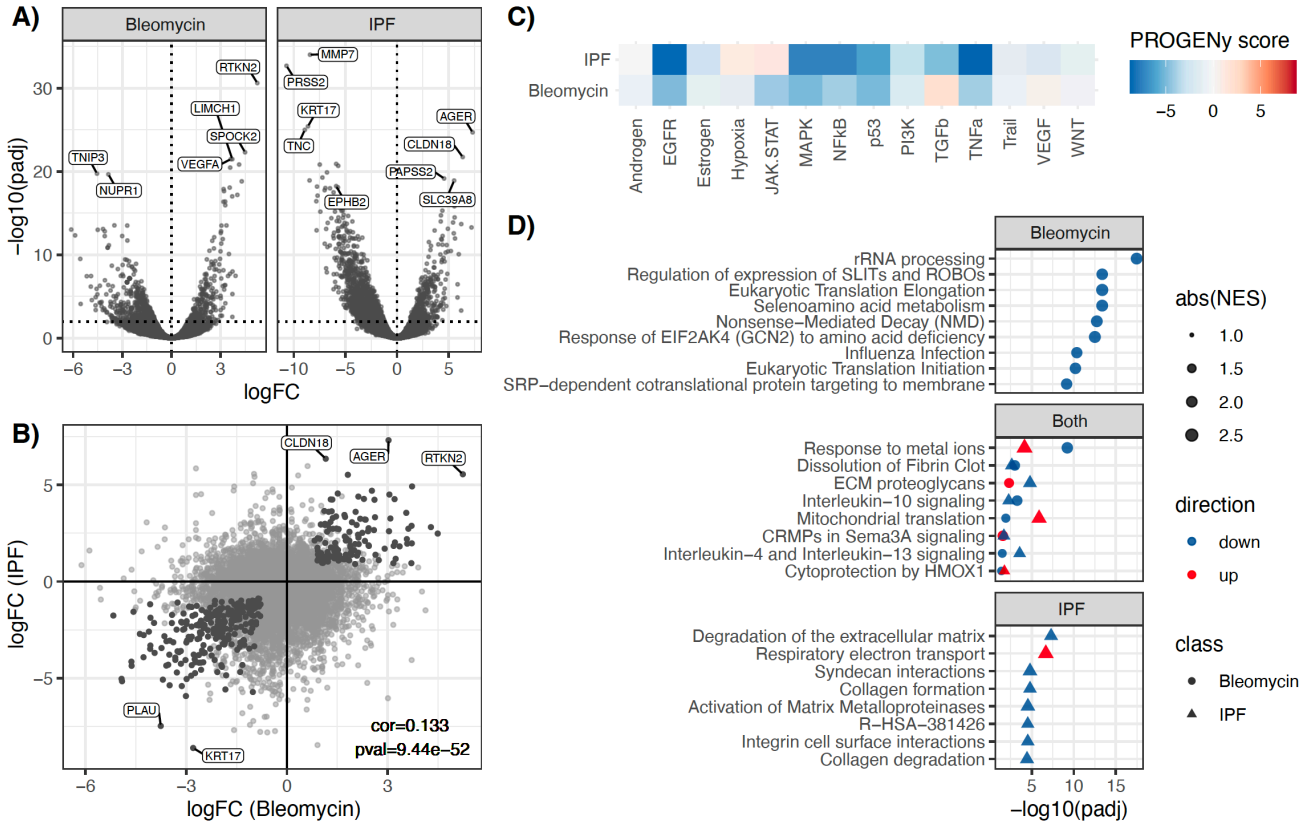


Figure 5.3: Transition signatures in bleomycin injury and IPF.

A) Differentially expressed genes (DEGs) in bleomycin and IPF signatures. B) Comparison of DEGs identifying significant but low correlation. C) PROGENy identifies consistent and strong down-regulation of EGFR, MAPK, NFκB, p53, and TNFα signalling (|score| >3) D) Reactome pathway enrichment identifies most significant down-regulation of pathways linked to ribosomal proteins in bleomycin injury, and extracellular matrix- and mitochondria-related processes in IPF. Significant down-regulation in both is found for signalling via interleukins 4,10 and 13, as well as dissolution of fibrin clots. Opposite directionality is found for multiple stress-related and developmental processes.

Globally comparing the transition signatures derived for bleomycin injury in mice and IPF in humans, a significant but low correlation is found between the differential expression profiles indicating that the transition signatures are similar but that there are also discrepancies between both conditions which were investigated further (Figure 5.3B). Subsequent gene set analysis revealed that the most strongly and significantly dysregulated pathways in bleomycin injury are linked to increased expression of ribosomal proteins, which are generally involved in the translation of mRNA to proteins (Figure 5.3D). They epigenetically regulate protein translation through so-called “heterogeneous ribosomes” [355], have previously been linked to cell fate decisions [356], such as proliferation, differentiation or tumorigenesis, and increased expression of ribosomal proteins was found to be a

characteristic of stem cells [357]. Consequentially, higher expression in intermediate progenitors is expected and absence thereof in IPF is pointing towards less differentiation between the defined AT1 and intermediate progenitor populations.

In contrast, the most pronounced pathways in IPF indicate increased levels of proteins linked to the extracellular matrix (ECM), including both collagens and matrix metalloproteinases which degrade collagen, as well as interactions on the cell surface mediated by syndecans and integrins which mediate cell adhesion to the ECM via collagen [358] (Figure 5.3D). In pulmonary fibrosis, simultaneously increased collagen production and degradation are known, however, effectively resulting in collagen accumulation which is a known driver of fibrosis via TGFβ signalling [359]. The absence of significant changes in these processes in the bleomycin injury model may be attributed to the fact that the intermediate progenitor population is less profibrotic, potentially because the injury is acute and not chronic, or also due to the known differences in collagen regulation between mice and humans [360].

Only a few pathways are found to be dysregulated in bleomycin injury and IPF (Figure 5.3D). These point to higher expression in intermediate progenitors of genes involved in the dissolution of fibrin clots, which is consistent with the role of fibrinolysis modulation via *PLAUR* in alveolar repair [361], and for inflammation-related pathways linked to interleukins IL-4, IL-10 and IL-13. Consistent with that an increased activity of NFκB and TNFα signalling is found with PROGENy, as well as of p53, which generally points to stress response, and of MAPK and EGFR signalling indicating proliferation (Figure 5.3C). However, also pathways with opposite directionality were found such as ECM proteoglycans, mitochondrial translation, HMOX1 and metallothioneins, as well as hypoxia TGFβ, VEGF and JAK-STAT signalling in PROGENy (Figure 5.3C and D). Overall, this sheds further light on the similarities and discrepancies between both conditions.

To further study which TFs are potentially involved in the transition from intermediate progenitor to AT1 cells in bleomycin injury and IPF, their activity was inferred based on regulon expression (Figure 5.4). Consistent down-regulation in AT1 cells compared to intermediate progenitors is observed for a strongly interlinked set of TFs, amongst others, related to stress, inflammation and proliferation with the strongest dysregulation being found for Jun proto-oncogene (JUN) and SMAD family member 3 (SMAD3). Furthermore, the

strongest TF which is activated in bleomycin injury but reduced in IPF is PR-domain containing protein 14 (PRDM14) which was previously identified as an epigenetic regulator of pluripotency in mice [362]. In contrast to that, a higher activation in IPF but lower activation in bleomycin injury is observed for MYC proto-oncogene (MYC) and zinc finger protein 263 (ZNF263), highlighting differences in transcriptional regulation between both conditions.

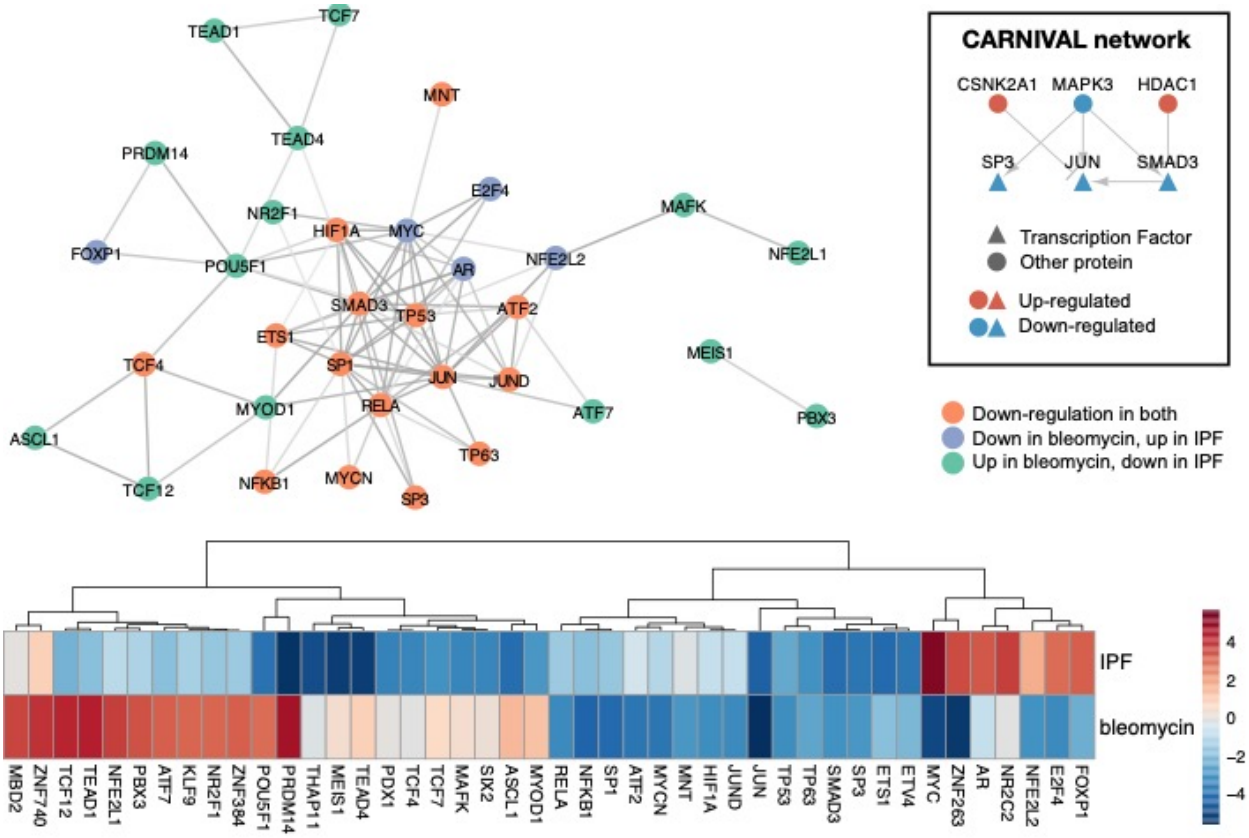


Figure 5.4: Transcription factors (TFs) involved in transition signature.

TFs inferred with DoRothEA regulons confidence A-C and an absolute normalized enrichment score (NES) > 3 in at least one transition signature are shown revealing shared down-regulation for some, e.g. JUN, HIF1A and TP53, while others show opposite directionality, such as PRDM14 or MYC. Protein-protein associations with high confidence in STRING (>0.7) reveal that the core cluster largely consists of consistently down-regulated TFs, while oppositely regulated ones form separate clusters. Furthermore, the consensus CARNIVAL consensus subnetwork is shown including CSNK2A1, MAPK3 and HDAC1 as upstream regulators.

Additionally, causal networks were inferred based on both transition signatures to further investigate upstream signalling (complete networks shown in Figure D.3 and Figure D.4). A consistent subnetwork was identified which consisted of the down-regulated TFs JUN, SMAD4 and SP3 as TFs as well as their potential upstream regulators. These included the down-regulated MAPK3, as well as the up-regulated histone deacetylase 1 (HDAC1) and

casein kinase 2 alpha 1 (CSNK2A1) shedding further light on potential upstream signalling involved in the transition.

5.3.2 Identification of repurposing candidates by signature matching

In this study, perturbation signatures from LINCS are used as it contains most drugs and includes cancer cell lines of lung origin, e.g. unlike its predecessor Cmap (Table 1.4). To identify compounds with the ability to induce transition-related gene expression changes in intermediate progenitors, perturbation signatures are required from a model system which is ideally similar to the intermediate progenitor population.

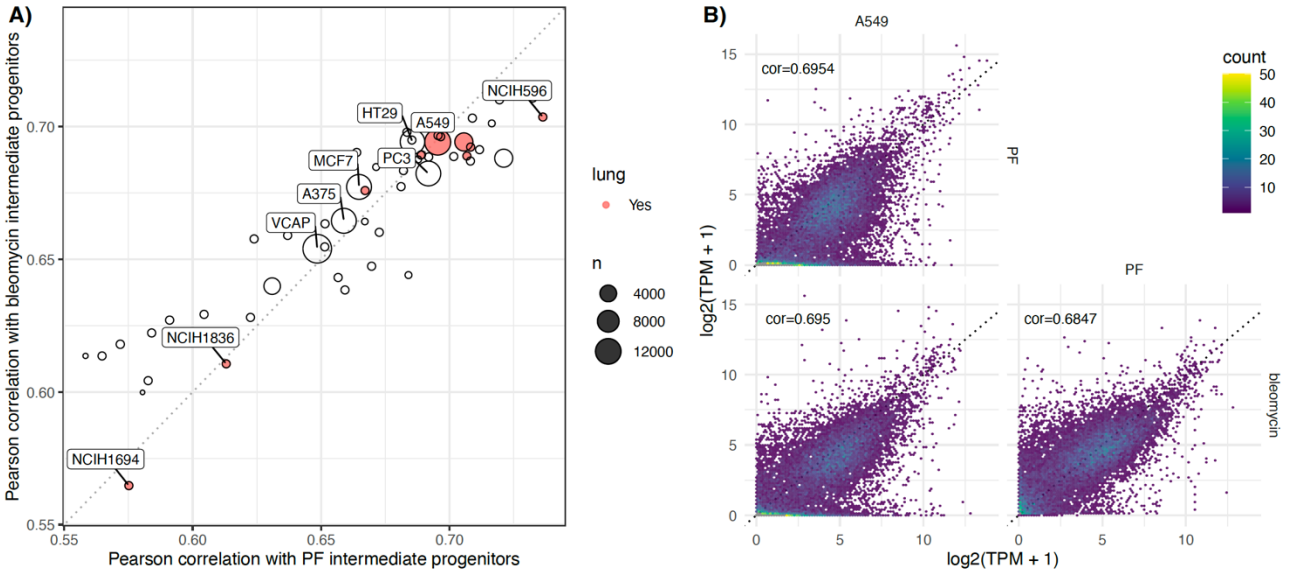


Figure 5.5: Comparison of baseline transcriptional profiles between LINCS cell lines and intermediate progenitor cells.

The correlation between the baseline expressions of the intermediate progenitor population from IPF or bleomycin, respectively, and each cell line in LINCS is shown, and the number of perturbation signatures is shown as point size. Among the cell lines with at least 10,000 measured perturbations, the highest correlation is found for the A549 lung adenocarcinoma cell line. The highest correlation among cell lines of lung origin (red) is found for the NCIH596 adenosquamous carcinoma cell line, while the lowest ones are found for the only two small cell lung cancer (SCLC) cell lines NCIH1694 and NCIH1836. B) Comparison between the baseline expression of the A549 cell line and the intermediate progenitor populations reveals that the A549 cell line is as transcriptionally similar to the intermediate progenitor populations as they are to each other. All correlations between baseline expression profiles in $\log_2$(TPM+1) were significant (p-value < 2.2e-16).

To evaluate the similarity between intermediate progenitors and cell lines, the intermediate progenitor baseline expression, for bleomycin injury and IPF, respectively, was compared to the baseline expression of the cell lines used in LINCS (Figure 5.5). Interestingly, the two lung cancer cell lines which are not highly correlated to the intermediate progenitors are small cell lung cancer (SCLC) cell lines which are of neuroendocrine origin while the others are derived from non-small cell lung cancer (NSCLC) which can originate from intermediate progenitors [363,364]. The highest correlation was found for the NCIH596 cell line which is of NSCLC origin, however given the higher number of measured perturbations, the A549 cell line was used in this study which is also of NSCLC origin. These have been previously used as *in vitro* model for AT2 cells [365,366], however, do not show all characteristics of AT2 cells, such as the expression of surfactant protein or the same phospholipid content [367,368].

Consensus perturbation signatures were derived as described in 5.2.1.2 and matched to the transition signatures for bleomycin injury in mice and IPF in humans, respectively. 89 perturbation signatures mapping to 84 compounds were found to significantly correlate (p-value < 0.05) with the bleomycin and IPF transition signatures (Figure 5.6). Multiple compound classes with correlated perturbation signatures were identified, which are shown in Figure 5.6 and Figure 5.7. The largest cluster contained corticosteroids, including fludroxycortide and halcinonide, which are generally anti-inflammatory and immunosuppressive [369], and have been widely used to treat IPF [370]. However, despite decades of use their efficacy in IPF remains questionable as no evidence was found for an association with improved clinical outcome [371,372]. Also, multiple kinase inhibitors with overall weaker correlated perturbation signatures were identified (Figure 5.7). This includes the highest-ranking compound by average correlation to both signatures, the spleen tyrosine kinase (SYK) inhibitor fostamatinib which is an approved treatment for chronic immune thrombocytopenia. While this has not been studied in the context of IPF, it was also prioritized in a phenotypic screen for acute lung injury [373] and was found to improve the clinical outcome in COVID-19 [374]. Also nintedanib, one of the approved treatments for IPF, is a tyrosine kinase inhibitor which is thought to be anti-fibrotic by targeting vascular endothelial growth factor (VEGFR), platelet-derived growth factor receptors (PDGFR) and fibroblast growth factor receptors (FGFR) [7]. However, as kinases take in broad functions in cellular signalling, these cannot be reduced to a single mechanism of action (MoA).

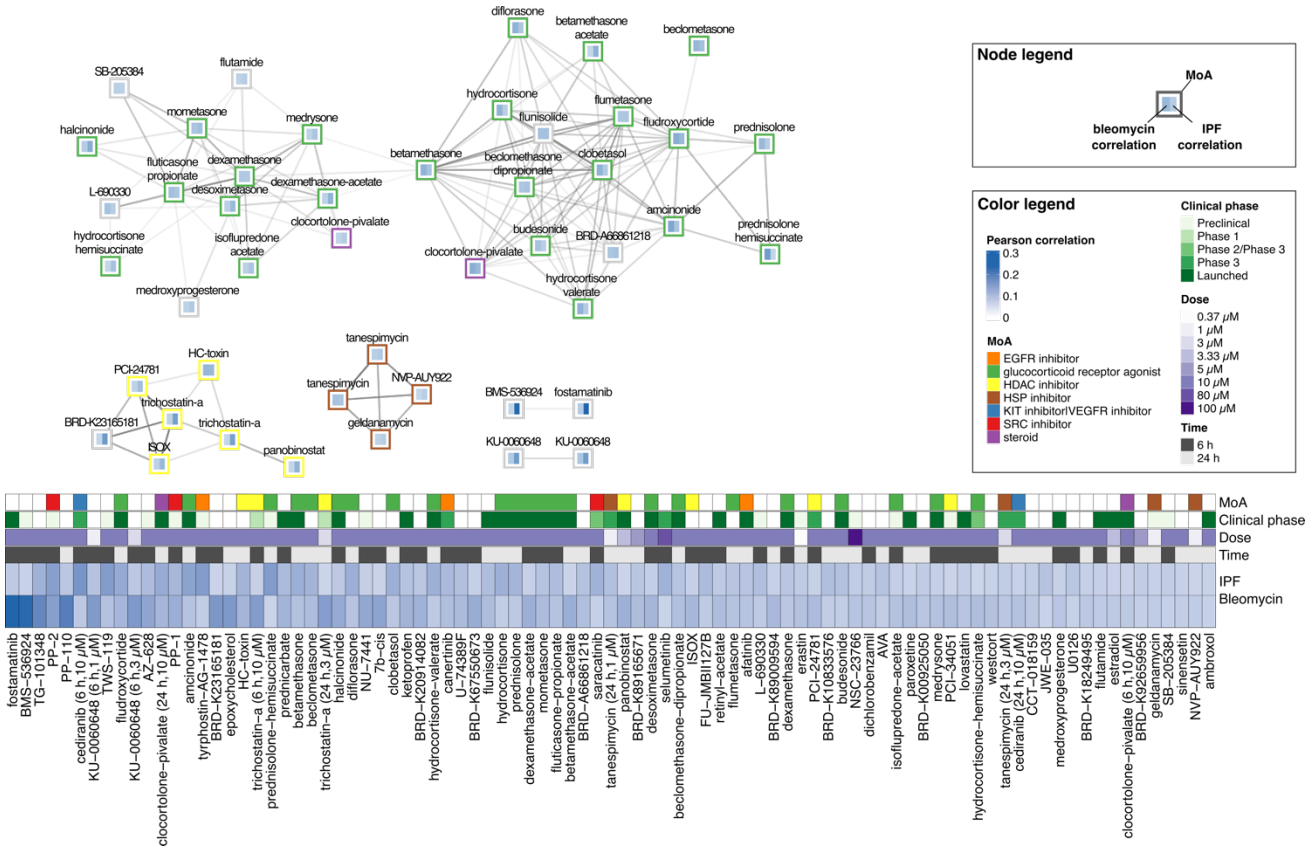


Figure 5.6: Perturbations matched to the bleomycin and IPF transition signatures.

The ranking by mean correlation is shown in the heatmap with the highest average correlation being observed for fostamatinib and the lowest one for ambroxol. Multiple transcriptionally similar compound clusters were identified based on the Pearson correlation between the perturbation signatures (Pearson correlations > 0.6 indicated as edges), namely HSP90 inhibitors, HDAC inhibitors and a large cluster including glucocorticoid receptor agonists and steroids.

Further clusters include inhibitors targeting histone deacetylases (HDACs), such as trichostatin A, and the molecular chaperone HSP90, such as tanespimycin, respectively. Both have not been explored heavily as targets in IPF, but have shown positive impacts on pulmonary fibrosis in animal studies. Histone deacetylases (HDACs) can acetylate histones and other proteins and hence can epigenetically and post-translationally modulate processes [375]. Hyperacetylation through HDAC inhibitors was found to alter gene expression in a tumour-suppressive manner countering the aberrantly high expression of HDACs generally found in cancers, and multiple HDAC inhibitors have hence been established as treatments for neoplastic diseases [376]. Due to their antifibrotic properties, they have also been studied in the context of pulmonary fibrosis [377], and trichostatin A (TSA), one of the matched HDAC inhibitors, was previously found to prevent pulmonary fibrosis in rats [378].

Furthermore, pirfenidone, one of two approved treatments for IPF with yet unknown MoA, was found to modulate HDAC expression and increase histone acylation [379].

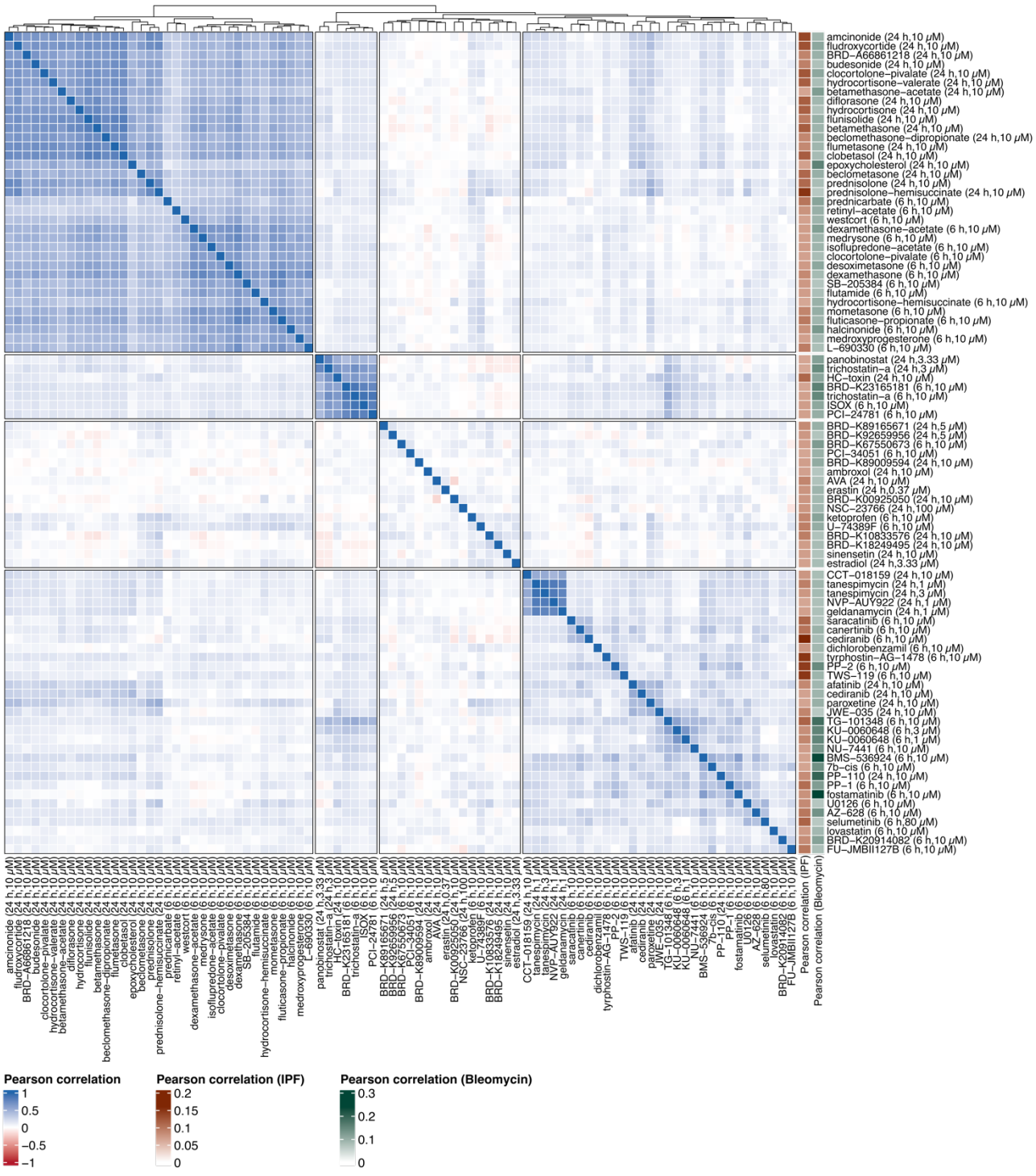


Figure 5.7: Pearson correlation between signatures of matched perturbations.

The correlation between the consensus signatures for landmark genes is shown identifying multiple transcriptional compound clusters.

HSP90 can support the folding of damaged proteins as well as their proteasome-mediated degradation [38], and overall contributes to proteostasis as part of the cellular stress response (Table 1.3). HSP90 inhibitors were found to be efficacious in treating cancer, as HSP90 stabilizes multiple oncogenic proteins and in particular kinases, and in treating viral infections as reduced protein folding results in reduced viral activity [380]. Also, the TGFβ receptor is a client protein of HSP90 and disruption of TGFβ signalling via HSP90 inhibition was found to ameliorate bleomycin-induced pulmonary fibrosis in mice [380,381]. While the MoA is not known for all compounds, e.g. for the aminosteroid U-74389F and ketoprofen, this shows that compound classes are identified for which a role in pulmonary fibrosis is already being investigated or established, although not primarily due to their effects on alveolar regeneration.

5.3.3 Deconvolution of potential targets and downstream TFs

As next step, the aim was to gain further insights into the potential mechanism of action through which drugs may induce the cell transition, which may also identify new targets not yet considered for targeting alveolar regeneration and pulmonary fibrosis. First, it was investigated which of the TFs involved in the disease transition, shown in Figure 5.4, were in fact modulated in the matched conditions and may hence be downstream effectors mediating drug action. To this end, TF activity inferred from the perturbation signatures was compared between matched signatures and signatures characterising unmatched compounds (FDR < 0.01). This identifies the most significant enrichment for the Jun proto-oncogene (JUN), followed by the NF-κB subunits encoded by RELA and NFKB1, all of which are down-regulated by matched compounds (Figure 5.8). NKFB1 [382] and HIF1A [383] have previously been implicated in alveolar regeneration, and Choi *et al.* demonstrated that the transient activation of both is essential for the AT2 to AT1 transition [133]. In contrast, e.g. the tumor proteins TP53 and TP63 are not differentially modulated in matched signatures, indicating that these are not targeted by matched compounds, despite dysregulation in the cell transition.

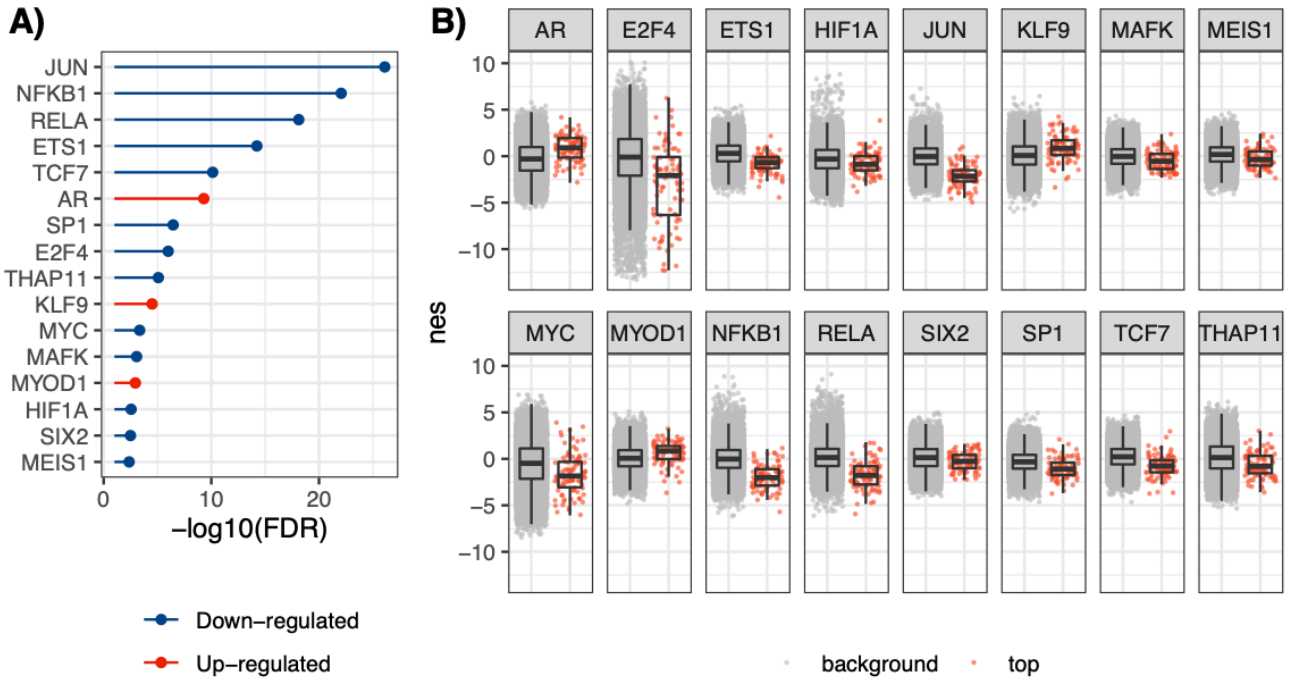


Figure 5.8: Transcription Factors (TFs) linked to disease signatures with differential regulon activity in matched LINCS signatures.

Among TFs with strong dysregulation in at least one of the disease signatures (|NES| > 3), the ones with differential regulon activity between matched perturbation signatures and signatures on unmatched compounds were identified using a t-test (FDR < 0.01). The most significant differential TF activity was found for JUN, followed by RELA and NFKB1 which are downregulated in matched signatures.

To identify potential direct targets involved in the induction of the intermediate progenitor to AT1 cell differentiation, *in vitro* target activity from CHEMBL was then combined with the signature matching correlation which serves as a proxy for activity on the cell transition. First, targets with significantly enriched activity among matched compounds were identified (p-value < 0.05) and here activity is defined as pCHEMBL ≥ 5 corresponding to at least half-maximal activity at 10 µM, which was the most frequently used concentration in the matched perturbation signatures.

Next, the Pearson correlation between bioactivity, estimated as pCHEMBL, and maximal signature matching correlation per compound was computed identifying six targets with significant positive correlations for both transition signatures (p-value < 0.05) which are shown in Figure 5.9: The glucocorticoid receptor NR3C1, which showed the strongest correlation between target activity and signature matching (Figure 5.10) and is known to be

involved in alveolar maturation in lung development [384], the TFs NFKB1 and HIF1A, which were also identified in Figure 5.8 to show down-regulation in matched compounds based on regulon activity, and the tyrosine kinases LCK, FYN and SRC.

While the matched compounds included multiple NR3C1 agonists and one SRC inhibitor (PP-2), the other targets were not reported as primary MoA for matched compounds. This shows that including the ChEMBL bioactivity data provides additional information on potential direct targets, and that this (unintended) polypharmacology of the compounds may contribute to their prioritization for repurposing.

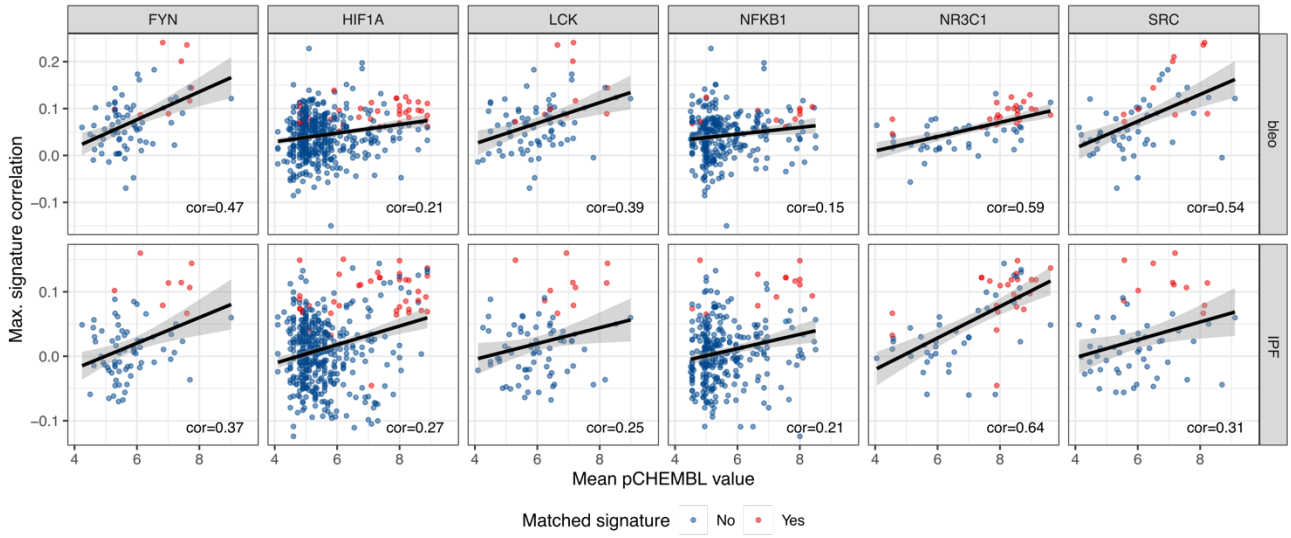


Figure 5.9: *In vitro* targets correlated with transcriptional matching.

Six protein targets were identified with significant (p-value < 0.05) and positive correlations between pCHEMBL values and signature matching correlation for both IPF and bleomycin, namely nuclear factor NF-κB p105 subunit (NFKB1), hypoxia-inducible factor 1α (HIF1A), glucocorticoid receptor (NR3C1), and the tyrosine kinases LCK, FYN and SRC.

To gain a broader overview on potentially involved targets, a less conservative filtering was applied in which only significance for one of the two transition signatures was required. This identified 36 potential direct targets, shown in Figure 5.10, with HIF1A showing the strongest enrichment of bioactivity (pCHEMBL ≥ 5), while the strongest correlation was found for NR3C1. Multiple targets already implied in the previously identified transcriptional clusters were recovered (Figure 5.6), namely NR3C1, HSP90AA1, two histone deacetylases (HDAC3 and HDAC6) and the HDAC3 and nuclear receptor corepressor 2 protein complex (HDAC3/NcoR2). Furthermore, 24 protein kinases were identified including 17 tyrosine kinases as well as the casein kinase CSNK2A1, which was also inferred in upstream

signalling based on the transition signatures (Figure 5.4). These were also found to be frequently co-identified, which can be explained by the known polypharmacology of kinase inhibitors [385] (Figure D.5). Additional targets included phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta (PIK3CB) and miRNA 21 which was the only identified target which was not a protein or protein complex, and is a pleiotropic miRNA involved in pulmonary remodelling [386]. Thereby, overall the strongest enrichment of bioactivity (pCHEMBL ≥ 5) was found for HIF1A, while the strongest correlation was found for NR3C1. Hence, this less stringent filtering is able to suggest additional targets, and recovers the primary targets linked to the previously transcriptional clusters.

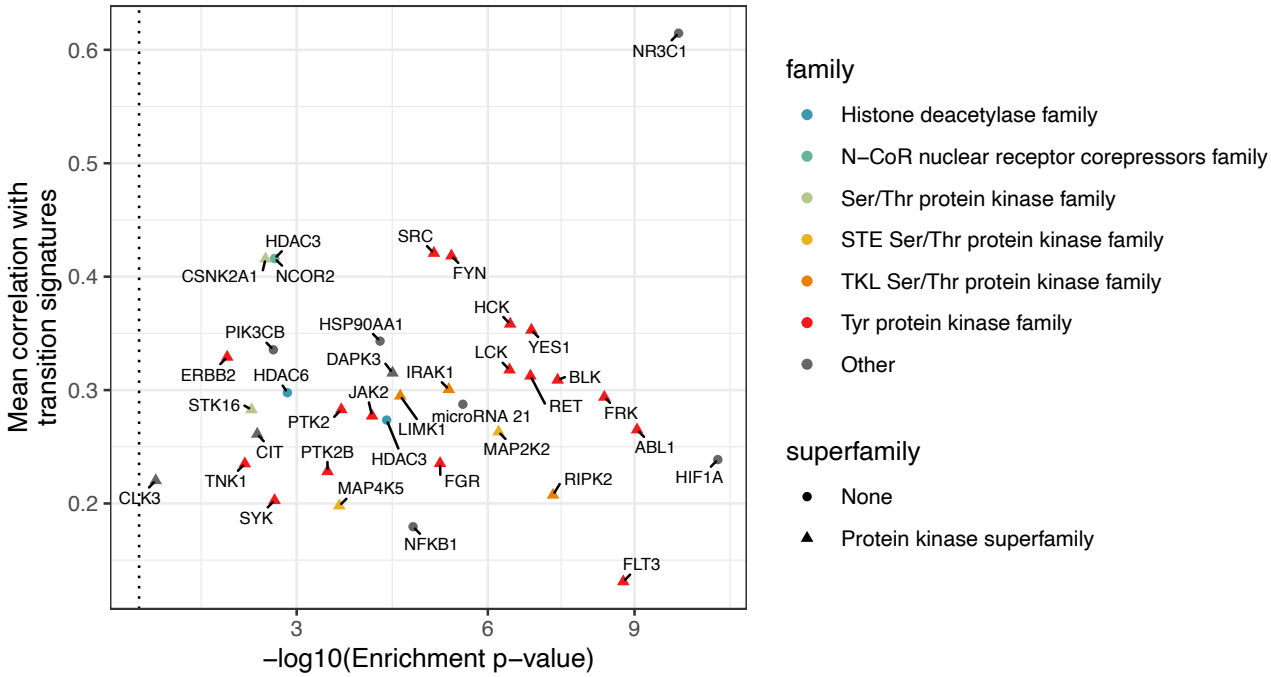


Figure 5.10: Correlation and enrichment of *in vitro* targets.

For all targets with positive correlations between bioactivity and signature matching for both signatures and significance for at least one, the mean correlation for both signatures is shown, as well as the enrichment FDR which describes if high bioactivity (pCHEMBL > 5) was over-represented among the matched compounds, compared to all other compounds with perturbation signatures in the A549 cell line. The most significant enrichment is found for HIF1A and the highest correlation for NR3C1.

To gain a more complete view on how the inferred targets may interact, physical interactions between the identified targets were derived from Omnipath [252], as shown in Figure 5.11. Among the 30 targets for which interactions were found, the highest (undirected) degree centrality was found for SRC (23 edges) followed by PTK2 (20 edges) indicating that modulation of these may also be linked to the modulation of many other proposed targets. Besides the kinases, which were found to be strongly interlinked, the highest degree centrality is found for HSP90AA1 (10 edges) and miRNA 21 (6 edges). Here it should be noted that the analysis may have underestimated the role of HDACs and HSP90, as these do not primarily act by interacting with or modulating individual proteins but by modulating a wide range of entities through indirect effects, namely, respectively, by epigenetically modulating their expression [375] and by stabilizing or degrading them [38]. Although not all types of functions are described equally well, protein-protein interactions may provide useful insights into targets with central functions.

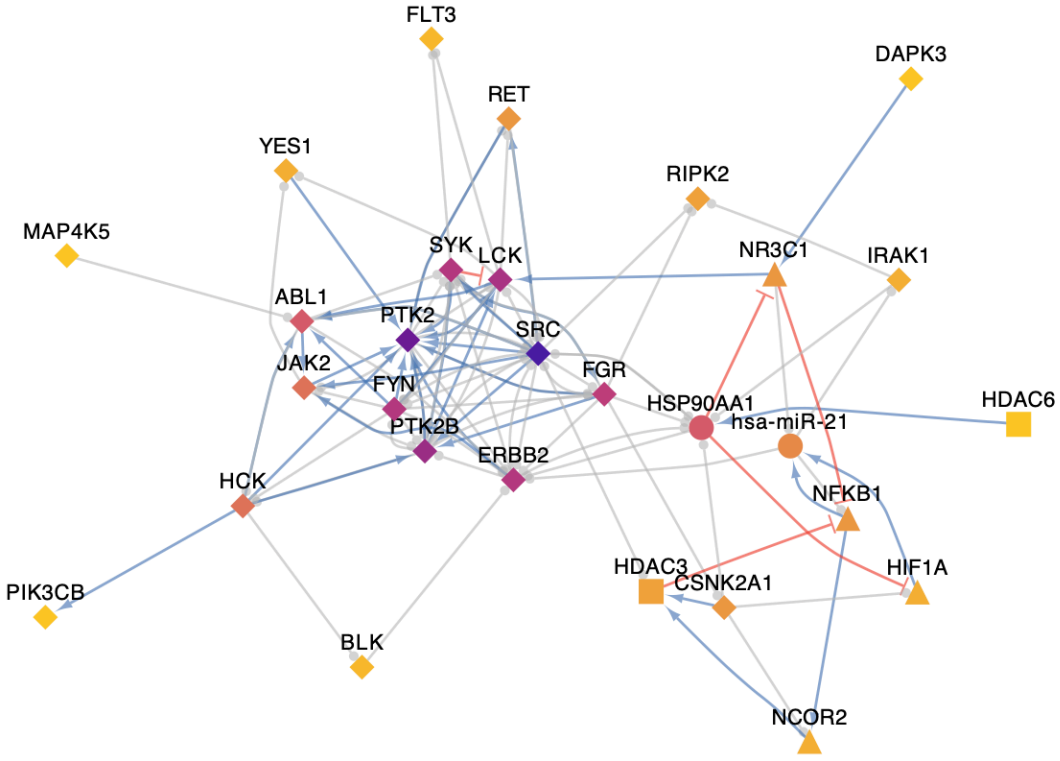


Figure 5.11: Protein-protein interactions between inferred *in vitro* targets.

Physical protein-protein interactions derived from Omnipath between potential targets are shown. Activatory (red) and inhibitory (blue) edges are indicated as colour, and edges without sign information are shown in grey. The highest degree centrality, indicated as darker node colour, is found for SRC followed by PTK2. Transcription factors (▲), epigenetic regulators (■), and kinases (◆) are indicated through distinct node shapes given the specific interest in these targets in this study.

As various kinase inhibitors and kinases have been highlighted by the analysis, their phylogenetic relation and whether these overlap with the kinome previously studied in IPF was investigated using KinMap [387] (Figure 5.12). Eight out of nine members of the SRC kinase family were identified including the SRC, LCK and FYN which were correlated with the matching correlation for both transition signatures. Other prominent families included the phylogenetically related SYK and FAK kinase families.

To investigate whether these overlap with or are related to kinases already studied in the context of IPF, the respective association scores were obtained from OpenTargets [14]. The strongest associations with IPF by far were found for the vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF) and fibroblast growth factor (FGF) receptor families (Figure 5.12), which are also the targets of kinase inhibitor nintedanib [7], one of two approved treatments for IPF. Besides these already established therapeutic targets, the highest association score is found for the tyrosine kinase ABL (ABL1), as two ABL inhibitors have previously been tested in clinical trials, namely imatinib [388] (ineffective in phase II trial) and dasatinib [389] (tested in combination with quercetin in a first-in-human study) since both were found to limit bleomycin-induced pulmonary fibrogenesis in mice [390,391]. ABL1 was also identified as potential target in this study and its inhibition was previously found to promote alveolar regeneration after pathogen-induced pulmonary injury [392] indicating that it may modulate both fibrogenesis and regeneration. There is hence some overlap between the kinases previously studied in the context of IPF, which were primarily of interest due to their anti-fibrotic properties, and the kinases identified to be relevant for alveolar regeneration.

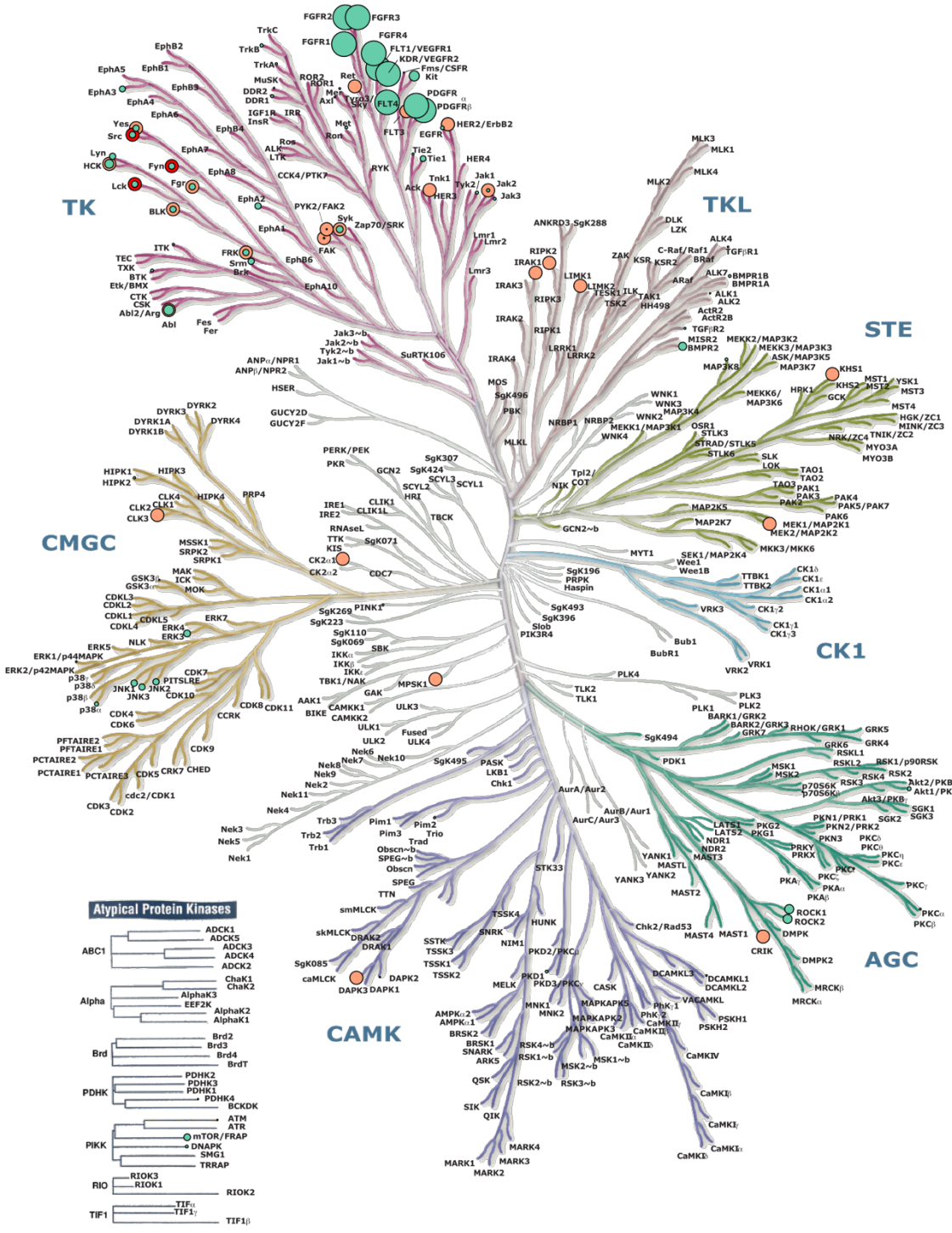


Figure 5.12: Protein kinases inferred as targets and those already studied in IPF on the phylogenetic tree of the human kinome.

The three tyrosine kinases with significant correlation to matching for both transition signatures are shown as red circles, and protein kinases significantly correlated to matching for only one are shown as orange circles. OpenTargets [14] association scores for IPF are indicated as the size of the turquoise circles. The figure was generated using KinMap [387] based on the kinome tree illustration by Cell Signaling Technology Inc.

To further refine the target list to those for which engagement is likely in the cell population of interest, their baseline expression in both intermediate progenitor populations from bleomycin injury and IPF was derived, as well as in the A549 cell line given that it was used for matching (Figure D.6). This identified the overall highest expression among inferred targets for *HSP90AA1*, consistent with the reported increased amount of HSP90 in fibrotic foci [380], followed by *HIF1A*, which is known to be a key mediator in alveolar regeneration [383]. The highest expression among kinases was found for *PTK2* and *CSNK1A1*, while seven other tyrosine kinases, including the four SRC family members *LCK*, *HCK*, *FGR* and *BLK* as well as *SYK*, *RET* and *FLT3*, were not expressed in all three cell populations. However, as expression was found in at least one intermediate progenitor population, the inferred targets were not excluded from the preceding analysis. It should be noted that phylogenetically close and highly expressed targets are identified in the analysis for the SRC family members and *SYK* (Figure 5.12). Hence, they may have been identified due to their functional similarity to potential direct targets resulting in similar *in vitro* bioactivity despite lesser relevance *in vivo*.

5.4 Limitations of this study

In this chapter, a first case study was presented using scRNA-Seq data for drug repurposing by targeting a mechanistically disease-relevant cell transition via signature matching. This not only identified potential compounds, but also potential targets and downstream effectors, which can serve as valuable starting points for drug discovery already before additional mechanistic information is available.

One limitation of this study is that signature matching was performed using the transition signatures from bleomycin injury and IPF jointly. Consequentially, promising candidates for either signature are neglected although discrepancies between the conditions were observed and are known, such as inter-species differences in the regulation of collagen [360] or the fact that bleomycin injury is rather acute than chronic [393]. This joint approach was still chosen to increase the likelihood for translation since the ultimate goal is to target the cell transition in IPF, while the most clinically relevant preclinical model for human IPF is bleomycin injury in mice [394] which will also serve as model system for the planned experimental validation.

Furthermore, as already noted in 1.6.3, there are multiple limitations to signature matching. While these can partially overcome using scRNA-Seq data, by characterising disease-relevant signatures *in vivo*, a fundamental limitation is that inducing transcriptional changes does not guarantee that the phenotype is modulated, e.g. if genetic or epigenetic factors are involved. Furthermore, it should be noted that multiple computational approaches for signature matching exist and it is not clear which one is most accurate. This is why e.g. `cTRAP` implements different metrics to characterise global changes, e.g. Pearson or Spearman correlation, as well as metrics focussing on the most extreme changes [348].

Consequentially, the identified signature matching correlation is both dependent on the chosen computational approach and also, it's meaningfulness from the biological perspective is not guaranteed which is a particularly important limitation in this study as signature matching is not only used for compound ranking but also quantitatively in the downstream deconvolution of targets and TFs. Although the most promising compounds and targets identified in this study are indeed supported by literature, experimental testing is hence crucial to evaluate the approach at hand.

5.5 Conclusion

Reinstating endogenous alveolar regeneration is an emerging therapeutic target to treat pulmonary fibrosis, due to its potential ability to not only reduce disease progression but to reverse it. Thanks to the advances in single-cell technologies, it is now possible to characterise the molecular identity of stem cell states involved in regeneration as well as their trajectories, and this has also led to the discovery of an intermediate progenitor cell state in the AT2 to AT1 cell differentiation in alveolar regeneration which is at the centre of this study. However, how to modulate these tightly regulated processes is not yet clear, also because key targets and mechanisms have yet to be discovered in the context of human IPF or lung disease.

This study aims to computationally prioritize small molecules with the ability to promote the intermediate progenitor to AT1 cell transition which may concurrently act as senotherapeutics [397] by reducing the level of senescent intermediate progenitors and as regenerative medicine [398] by restoring the AT1 population and potentially lung function. To this end, publicly available scRNA-Seq datasets, based on which the transition initially

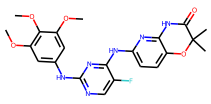
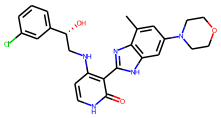
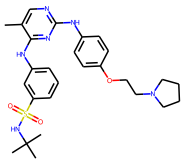
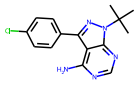
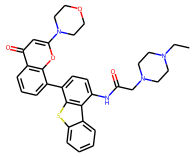
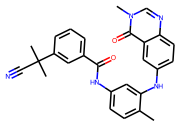
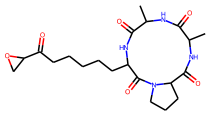
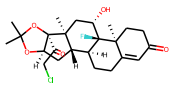
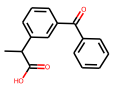
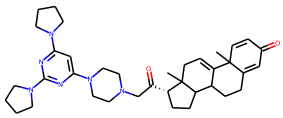
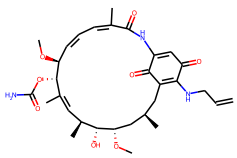
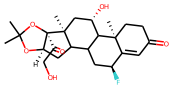
discovered, were used to characterise the transcriptional changes associated with the transition and match this to compounds from the LINCS database which induce similar transcriptional changes and hence may promote the transition. Using this approach, fostamatinib was identified as most promising candidate which aligns with other studies which have prioritised it for acute lung injury [373] and showed that it improves the clinical outcome in COVID-19 [374]. Overall, multiple kinase inhibitors, glucocorticoid agonists, HDAC inhibitors and HSP90 inhibitors were identified as matched, and bioactivity for the respective targets of these compound classes was also found to correlate with signature matching providing complementary evidence that these targets are mechanistically linked to the cell transition (Figure 5.10).

As additional promising targets, NFKB1 and HIF1A are identified which are key TFs based on the transition signatures (Figure 5.4, Figure 5.8), and show a correlation between bioactivity and signature matching for both transition signatures (Figure 5.9). HIF1A additionally showed the most significant enrichment of activity (pCHEMBL ≥ 5) in matched compounds (Figure 5.10) and JUN was identified as most significantly differentially regulated TF (Figure 5.8). Indeed, termination of NKFB1-mediated inflammation [382], as well as hypoxia mediated by HIF1A [383], have previously been implicated in alveolar regeneration, and transient activation of NFκB and HIF1A was previously found to be essential for the AT2 to AT1 transition [133].

Among the kinases, the SRC kinase family is particularly well represented, with three members being among the six targets which correlate to both transition signatures (Figure 5.12). Overall, SRC is also the most central node in the PPI network connecting potential targets, followed by PTK2 (Figure 5.11), which also showed the highest expression in intermediate progenitors making it overall a likely target in the FAK kinase family (Figure D.6).

As a follow-up, we aim to experimentally test selected compounds in organoid models in collaboration with the Lee Lab at the Wellcome-MRC Cambridge Stem Cell Institute, and prioritised compounds for this based on the strength of signature matching, but also aimed to select mechanistically diverse and structurally interesting compounds (Table 5.2).

Table 5.2: Summary of compounds selected for experimental validation.

Compound	Structure	MoA	Clinical phase	Pearson correlation		indication
				IPF	Bleo-mycin	
Fostamatinib		SYK inhibitor	Launched	0.079	0.238	chronic immune thrombocytopenia (ITP)
BMS-536924		IGF-1 inhibitor	Preclinical	0.067	0.234	-
TG-101348 (Fedratinib)		JAK2 inhibitor	Launched	0.114	0.182	-
PP-2		SRC inhibitor	Preclinical	0.144	0.144	-
KU-0060648		DNA-PK/PI3K inhibitor	Preclinical	0.088	0.154	-
AZ-628		RAF inhibitor	Preclinical	0.099	0.135	-
HC-toxin		HDAC inhibitor	-	0.112	0.109	-
Halcinonide		glucocorticoid receptor agonist	Launched	0.080	0.127	corticosteroid-responsive dermatoses
Ketoprofen		cyclooxygenase inhibitor	Launched	0.088	0.113	rheumatoid arthritis, osteoarthritis
U-74389F		-	-	0.094	0.100	-
Tanespimycin		HSP90 inhibitor	Phase 3	0.093	0.089	-
Fludroxycortide		glucocorticoid receptor agonist	Launched	0.064	0.051	skin infections, psoriasis

Based on the transcriptional and mechanistic clustering, shown in Figure 5.6, HC-toxin and tanespimycin were chosen as representatives for HDAC and HSP90 inhibitors, respectively. Furthermore, fludroxycortide and halcinonide were chosen as two glucocorticoid receptor agonists represented the two transcriptional subclusters. Additionally, the six kinase inhibitors which showed the highest correlations to the transition signatures and different selectivity profiles were included (Figure 5.9). Finally, ketoprofen and the aminosteroid U-74389F were chosen due to their high ranking and the fact that they are mechanistically and structurally distinct (Figure 5.6).

While the effect of the tested compounds on alveolar regeneration remains to be validated experimentally, this study hence demonstrates how scRNA-Seq data can be used for computational drug repurposing and how it is then possible to further prioritise potentially involved targets and downstream TFs. It should be highlighted, that similar cases where transitions are first understood transcriptionally can be expected to arise, thanks to single-cell transcriptomics. In these cases, signature matching might provide candidates for drug repurposing and, on a more general level, valuable starting points for drug discovery, also when additional mechanistic information is not available yet.

6 Conclusion

The aim of this thesis was to use transcriptomics to deepen our understanding on how compounds relate to injury with the ultimate goal to derive findings which can advance the drug discovery process. To this end, historical transcriptomics datasets were integrated with other sources of information by combining already existing tools into new analysis pipelines which take advantage of the inherent strengths of the respective datasets. While the results and limitations are discussed in more detail within the respective chapters, the broader context of the presented work and findings will be summarised here.

In Chapter 2, safety biomarker candidates for DIVI were identified using a computational filtering pipeline across transcriptomics data from repeat-dose studies in rats extending previous analysis performed by Dalmas *et al.* [66]. Here, not only consistency, specificity and dose-response were evaluated across compounds but expression changes were also linked to observed and anticipated histopathological changes providing further insights into the genes' prognostic properties. Overall, 33 biomarker candidates were identified and characterised with the most predictive ones encoding secreted proteins. While only results for the most promising candidate biomarkers based on the available data and the implemented filtering criteria are shown, it is clear that biomarker development in practice is heavily informed by broader knowledge on DIVI as demonstrated by ongoing work of the PSTC VIWG and TransBioLine, and is also dependent on additional criteria, such as the availability of suitable and reliable assays, the translation to serum biomarkers, or the specificity for DIVI in comparison to other types of injury. To also support future efforts on the discovery of DIVI biomarkers, this work also provides a publicly available web application (`https://anikaliu.shinyapps.io/divi`) allowing visualization and exploration of gene-level results associated with the presence and/or absence of MAN in rat mesentery for genes of interest beyond the 33 biomarker candidates prioritized in this study.

As second safety-related endpoint, mechanistic insights on the pathogenesis of DILI preceding adverse histopathology was derived from time-resolved gene expression and histopathology data in Chapters 3 and 4. This presents a conceptually new approach based on systematic time concordance analysis, which leverages the large number of time-series linked to different *in vivo* perturbations in the Open TG-GATEs data [150]. In Chapter 3, it was

shown how this automatable methodology prioritises and characterises known and mechanistically relevant events preceding adverse histopathology in DILI. Furthermore, the combination with causal prior knowledge, provides detailed hypothesis on TF mode of action and interactions, and the combination with time dependence prioritizes potential mechanistic biomarkers. While the presented work focusses on events preceding a specific definition of adverse histopathology, also other definitions are possible or the temporal relation between cellular events may be of interest. This can be explored interactively in the `DILI Cascades` Shiny app (`https://anikaliu.shinyapps.io/dili_cascades`), presented in Chapter 4. However, it should be noted that this can be limited by the number of time-series available and the limited temporal resolution, which was also the reason why a systematic analysis of the temporal order of cellular events within adverse time-series was only pursued for TF-level events additionally supported by prior knowledge. Overall, time concordance hence provides valuable starting points for AOP development and evidence towards causality as outlined in the Bradford Hill considerations.

In Chapter 5, scRNA-Seq datasets, originally generated to better understand IPF and alveolar regeneration [22,137,138,140], are reused to prioritize small molecules as regenerative medicine. Thereby, the single-cell resolution enables the transcriptional characterisation of a therapeutically relevant *in vivo* differentiation process. This not only identifies multiple compound classes as candidates for repurposing by signature matching, but further suggests potentially involved direct targets and downstream effectors. The usefulness of the presented pipeline depends on whether indeed promising repurposing candidates are identified and testing of selected repurposing candidates in mouse-derived organoids is still pending. However, fostamatinib, the most highly ranked compound, was found to be promising in other conditions linked to lung injury, which supports the presented approach. Generally, this demonstrates how targeting disease-relevant cell transitions can be directly discovered from scRNA-Seq data, which is particularly interesting when additional mechanistic information is not available yet.

Overall, transcriptomics was hence used to study pathological processes across multiple organs and tissues in combination with other sources of information. Thereby, the presented approaches study dynamic biological processes across scales, from angles beyond those originally intended in the study designs, and the R/Shiny apps make the derived results

easily accessible to the research community. In summary, this provides new starting points to detect, understand and treat injury in the context of adverse effects and fibrotic disease. Ultimately, this aims to derive insights from historical data to increase efficiency in drug discovery, either through a better identification of safety risks or by prioritizing repurposing candidates with already established safety.

7 References

1. Paul, S. M. *et al.* How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**, 203–214 (2010).

2. Harrison, R. K. Phase II and phase III failures: 2013-2015. *Nat Rev Drug Discov* **15**, 817–818 (2016).

3. Bjornsson, T. D. A classification of drug action based on therapeutic effects. *J Clin Pharmacol* **36**, 669–673 (1996).

4. Ison, M. G. Antiviral Treatments. *Clin Chest Med* **38**, 139–153 (2017).

5. Silver, L. L. Challenges of antibacterial discovery. *Clin Microbiol Rev* **24**, 71–109 (2011).

6. Vetter, V., Denizer, G., Friedland, L. R., Krishnan, J. & Shapiro, M. Understanding modern-day vaccines: what you need to know. *Ann Med* **50**, 110–120 (2018).

7. Wollin, L. *et al.* Mode of action of nintedanib in the treatment of idiopathic pulmonary fibrosis. *European Respiratory Journal* **45**, 1434–1445 (2015).

8. Taniguchi, H. *et al.* Pirfenidone in idiopathic pulmonary fibrosis. *European Respiratory Journal* **35**, 821–829 (2010).

9. te Riet, L., van Esch, J. H. M., Roks, A. J. M., van den Meiracker, A. H. & Danser, A. H. J. Hypertension: Renin-Angiotensin-Aldosterone System Alterations. *Circ Res* **116**, 960–975 (2015).

10. Power, I. An update on analgesics. *Br J Anaesth* **107**, 19–24 (2011).

11. Kim, D., Kim, H. J. & Ahn, S. Anesthetics Mechanisms: A Review of Putative Target Proteins at the Cellular and Molecular Level. *Curr Drug Targets* **19**, 1333–1343 (2018).

12. Endo, A. A historical perspective on the discovery of statins. *Proc Jpn Acad Ser B Phys Biol Sci* **86**, 484–493 (2010).

13. Stancu, C. & Sima, A. Statins: Mechanism of action and effects. *J Cell Mol Med* **5**, 378–387 (2001).

14. Koscielny, G. *et al.* Open Targets: A platform for therapeutic target identification and Validation. *Nucleic Acids Res* **45**, D985–D994 (2017).

15. van Drie, J. H. Hit diffusion: limitations to drug discovery and structure-based design. *J Comput Aided Mol Des* **36**, 373–379 (2021).

16. Hwang, T. J. *et al.* Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern Med* **176**, 1826–1833 (2016).

17. Liebler, D. C. & Guengerich, F. P. Elucidating mechanisms of drug-induced toxicity. *Nat Rev Drug Discov* **4**, 410–420 (2005).

18. Uetrecht, J. Idiosyncratic drug reactions: Past, present, and future. *Chem Res Toxicol* **21**, 84–92 (2008).

19. Jenkinson, S., Schmidt, F., Rosenbrier Ribeiro, L., Delaunois, A. & Valentin, J. P. A practical guide to secondary pharmacology in drug discovery. *J Pharmacol Toxicol Methods* **105**, (2020).

20. Bowes, J. *et al.* Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev Drug Discov* **11**, 909–922 (2012).

21. Beilmann, M. *et al.* Optimizing drug discovery by Investigative Toxicology: Current and future trends. *ALTEX* **36**, 289–313 (2018).

22. Törnqvist, E. *et al.* Strategic focus on 3R principles reveals major reductions in the use of animals in pharmaceutical toxicity testing. *PLoS One* **9**, e101638 (2014).

23. Ankley, G. T. *et al. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment*. *Environmental Toxicology and Chemistry* vol. 29 730–741 (Wiley Blackwell, 2010).

24. Meek, M. E. *et al.* New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. *Journal of Applied Toxicology* **34**, 1–18 (2014).

25. Meek, M. E. *et al.* Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. *Journal of Applied Toxicology* **34**, 595–606 (2014).

26. Bai, J. P. F. & Abernethy, D. R. Systems pharmacology to predict drug toxicity: Integration across levels of biological organization. *Annu Rev Pharmacol Toxicol* **53**, 451–473 (2013).

27. Leist, M. *et al.* Adverse outcome pathways: opportunities, limitations and open questions. *Arch Toxicol* **91**, 3477–3505 (2017).

28. Galluzzi, L., Yamazaki, T. & Kroemer, G. Linking cellular stress responses to systemic homeostasis. *Nat Rev Mol Cell Biol* **19**, 731–745 (2018).

29. Pakos-Zebrucka, K. *et al.* The integrated stress response. *EMBO Rep* **17**, 1374–1395 (2016).

30. Chovatiya, R. & Medzhitov, R. Stress, inflammation, and defense of homeostasis. *Mol Cell* **54**, 281–288 (2014).

31. Alberts, B. *Molecular biology of the cell*. (Garland Science, 2015).

32. Shweiki, D., Itin, A., Soffer, D. & Keshet, E. Vascular endothelial growth factor induced by hypoxia may mediate hypoxia-initiated angiogenesis. *Nature* **359**, 843–845 (1992).

33. Kumari, R. & Jat, P. Mechanisms of Cellular Senescence: Cell Cycle Arrest and Senescence Associated Secretory Phenotype. *Front Cell Dev Biol* **9**, 485 (2021).

34. Tang, D., Kang, R., Berghe, T. vanden, Vandenabeele, P. & Kroemer, G. The molecular machinery of regulated cell death. *Cell Res* **29**, 347–364 (2019).

35. Tang, D., Kang, R., Berghe, T. vanden, Vandenabeele, P. & Kroemer, G. The molecular machinery of regulated cell death. *Cell Research 2019 29:5* **29**, 347–364 (2019).

36. Simmons, S. O., Fan, C. Y. & Ramabhadran, R. Cellular Stress Response Pathway System as a Sentinel Ensemble in Toxicological Screening. *Toxicological Sciences* **111**, 202–225 (2009).

37. Ursini, F., Maiorino, M. & Forman, H. J. Redox homeostasis: The Golden Mean of healthy living. *Redox Biol* **8**, 205–215 (2016).

38. Schopf, F. H., Biebl, M. M. & Buchner, J. The HSP90 chaperone machinery. *Nat Rev Mol Cell Biol* **18**, 345–360 (2017).

39. Zhou, B. B. S. & Elledge, S. J. The DNA damage response: Putting checkpoints in perspective. *Nature* **408**, 433–439 (2000).

40. Tirpe, A. A., Gulei, D., Ciortea, S. M., Crivii, C. & Berindan-Neagoe, I. Hypoxia: Overview on hypoxia-mediated mechanisms with a focus on the role of hif genes. *Int J Mol Sci* **20**, 6140 (2019).

41. Hetz, C., Zhang, K. & Kaufman, R. J. Mechanisms, regulation and functions of the unfolded protein response. *Nature Reviews Molecular Cell Biology 2020 21:8* **21**, 421–438 (2020).

42. Cheung, C. Y. K. & Ko, B. C. B. NFAT5 in cellular adaptation to hypertonic stress - regulations and functional significance. *J Mol Signal* **8**, 1–9 (2013).

43. Medzhitov, R. Origin and physiological roles of inflammation. *Nature* **454**, 428–435 (2008).

44. Tang, D., Kang, R., Coyne, C. B., Zeh, H. J. & Lotze, M. T. PAMPs and DAMPs: Signal 0s that spur autophagy and immunity. *Immunol Rev* **249**, 158–175 (2012).

45. Wynn, T. A. & Vannella, K. M. Macrophages in Tissue Repair, Regeneration, and Fibrosis. *Immunity* **44**, 450–462 (2016).

46. Jopling, C., Boue, S. & Belmonte, J. C. I. Dedifferentiation, transdifferentiation and reprogramming: Three routes to regeneration. *Nat Rev Mol Cell Biol* **12**, 79–89 (2011).

47. Weiskirchen, R., Weiskirchen, S. & Tacke, F. Organ and tissue fibrosis: Molecular signals, cellular mechanisms and translational implications. *Mol Aspects Med* **65**, 2–15 (2019).

48. Darby, I. A., Zakuan, N., Billet, F. & Desmoulière, A. The myofibroblast, a key cell in normal and pathological tissue repair. *Cellular and Molecular Life Sciences* **73**, 1145–1157 (2016).

49. Wynn, T. A. Cellular and molecular mechanisms of fibrosis. *Journal of Pathology* **214**, 199–210 (2008).

50. Zhang, D. Y. & Friedman, S. L. Fibrosis-dependent mechanisms of hepatocarcinogenesis. *Hepatology* **56**, 769–775 (2012).

51. Karampitsakos, T. *et al.* Lung cancer in patients with idiopathic pulmonary fibrosis. *Pulm Pharmacol Ther* **45**, 1–10 (2017).

52. Stengel, B. Chronic kidney disease and cancer: a troubling connection. *J Nephrol* **23**, 253 (2010).

53. Haak, A. J. *et al.* Selective YAP/TAZ inhibition in fibroblasts via dopamine receptor D1 agonism reverses fibrosis. *Sci Transl Med* **11**, (2019).

54. Gu, L. *et al.* Targeting Cpt1a-Bcl-2 interaction modulates apoptosis resistance and fibrotic remodeling. *Cell Death Differ* **29**, 118–132 (2022).

55. Underwood, J. C. E. More than meets the eye: the changing face of histopathology. *Histopathology* **70**, 4 (2017).

56. Kilty, C. G., Keenan, J. & Shaw, M. Histologically defined biomarkers in toxicology. *Expert Opin Drug Saf* **6**, 207–215 (2007).

57. Histopathology is ripe for automation. *Nat Biomed Eng* **1**, 925 (2017).

58. Chan, J. K. C. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. *Int J Surg Pathol* **22**, 12–32 (2014).

59. Elmore, S. Apoptosis: A Review of Programmed Cell Death. *Toxicol Pathol* **35**, 495–516 (2007).

60. Dalmas, D. A. *et al.* Transcriptional Profiling of Laser Capture Microdissected Rat Arterial Elements: Fenoldopam-induced Vascular Toxicity as a Model System. *Toxicol Pathol* **36**, 496–519 (2008).

61. Kerns, W. *et al.* Drug-induced vascular injury - A quest for biomarkers. *Toxicol Appl Pharmacol* **203**, 62–87 (2005).

62. Morton, D. & Houle, C. D. Perspectives on Drug-induced Vascular Injury. *Toxicol Pathol* **42**, 633–634 (2014).

63. Sobota, J. T. Review of cardiovascular findings in humans treated with minoxidil. *Toxicol Pathol* **17**, 193–202 (1989).

64. Kavanaugh, A. *et al.* Treatment of psoriatic arthritis in a phase 3 randomised, placebo-controlled trial with apremilast, an oral phosphodiesterase 4 inhibitor. *Ann Rheum Dis* **73**, 1020–1026 (2014).

65. Johansson, S. Cardiovascular lesions in Sprague-Dawley rats induced by long-term treatment with caffeine. *Acta Pathol Microbiol Scand A* **89**, 185–91 (1981).

66. Dalmas, D. A. *et al.* Potential candidate genomic biomarkers of drug induced vascular injury in the rat. *Toxicol Appl Pharmacol* **257**, 284–300 (2011).

67. Pearson, T. A. *et al.* Markers of inflammation and cardiovascular disease: Application to clinical and public health practice: A statement for healthcare professionals from the centers for disease control and prevention and the American Heart Association. *Circulation* **107**, 499–511 (2003).

68. Al-Soudi, A., Kaaij, M. H. & Tas, S. W. Endothelial cells: From innocent bystanders to active participants in immune responses. *Autoimmun Rev* **16**, 951–962 (2017).

69. Mai, J., Virtue, A., Shen, J., Wang, H. & Yang, X. F. An evolving new paradigm: Endothelial cells - Conditional innate immune cells. *J Hematol Oncol* **6**, (2013).

70. Oka, K. *et al.* Lectin-like oxidized low-density lipoprotein receptor 1 mediates phagocytosis of aged/apoptotic cells in endothelial cells. *Proc Natl Acad Sci U S A* **95**, 9535–9540 (1998).

71. Opitz, B., Eitel, J., Meixenberger, K. & Suttorp, N. Role of Toll-like receptors, NOD-like receptors and RIG-I-like receptors in endothelial cells and systemic infections. *Thromb Haemost* **102**, 1103–1109 (2009).

72. Rothermel, A. L. *et al.* Endothelial cells present antigens in vivo. *BMC Immunol* **5**, (2004).

73. Rao, R. M., Yang, L., Garcia-Cardena, G. & Luscinskas, F. W. Endothelial-dependent mechanisms of leukocyte recruitment to the vascular wall. *Circ Res* **101**, 234–247 (2007).

74. Muñoz-Chápuli, R., Quesada, A. R. & Medina, M. Á. Angiogenesis and signal transduction in endothelial cells. *Cellular and Molecular Life Sciences* **61**, 2224–2243 (2004).

75. Kovacic, J. C. *et al.* Endothelial to Mesenchymal Transition in Cardiovascular Disease: JACC State-of-the-Art Review. *J Am Coll Cardiol* **73**, 190–209 (2019).

76. Davidson, S. M. Endothelial mitochondria and heart disease. *Cardiovasc Res* **88**, 58–66 (2010).

77. Frismantiene, A., Philippova, M., Erne, P. & Resink, T. J. Smooth muscle cell-driven vascular diseases and molecular mechanisms of VSMC plasticity. *Cell Signal* **52**, 48–64 (2018).

78. Joseph, E. C. Arterial lesions induced by phosphodiesterase III (PDE III) inhibitors and DA1 agonists. in *Toxicology Letters* vols 112–113 537–546 (Elsevier, 2000).

79. Onakpoya, I. J., Heneghan, C. J. & Aronson, J. K. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: A systematic review of the world literature. *BMC Med* **14**, 1–11 (2016).

80. Reuben, A., Koch, D. G. & Lee, W. M. Drug-induced acute liver failure: Results of a U.S. multicenter, prospective study. *Hepatology* **52**, 2065–2076 (2010).

81. Björnsson, E. S., Bergmann, O. M., Björnsson, H. K., Kvaran, R. B. & Olafsson, S. Incidence, presentation, and outcomes in patients with drug-induced liver injury in the general population of iceland. *Gastroenterology* **144**, 1419-1425.e3 (2013).

82. Katarey, D. & Verma, S. Drug-induced liver injury. *Clin Med (Lond)* **16**, s104–s109 (2016).

83. Ostapowicz, G. *et al.* Results of a Prospective Study of Acute Liver Failure at 17 Tertiary Care Centers in the United States. (2002).

84. Lauschke, V. M. Toxicogenomics of drug induced liver injury–from mechanistic understanding to early prediction. *Drug Metab Rev* **53**, 245–252 (2021).

85. Teschke, R. & Danan, G. Drug Induced Liver Injury: Mechanisms, Diagnosis, and Clinical Management. in *Liver Diseases* 95–105 (Springer International Publishing, 2020). doi:10.1007/978-3-030-24432-3_9.

86. Weaver, R. J. *et al.* Managing the challenge of drug-induced liver injury: a roadmap for the development and deployment of preclinical predictive models. *Nat Rev Drug Discov* **19**, 131–148 (2020).

87. Chen, M. *et al.* DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov Today* **21**, 648–653 (2016).

88. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res* **44**, D1075–D1079 (2016).

89. Liu, A. *et al.* Prediction and mechanistic analysis of drug-induced liver injury (DILI) based on chemical structure. *Biol Direct* **16**, 1–15 (2021).

90. Vall, A. *et al.* The Promise of AI for DILI Prediction. *Front Artif Intell* **4**, 15 (2021).

91. Mervin, L. H. *et al.* Target prediction utilising negative bioactivity data covering large chemical space. *J Cheminform* **7**, 1–16 (2015).

92. Canbay, A., Bechmann, L. & Gerken, G. Lipid metabolism in the liver. *Z Gastroenterol* **45**, 35–41 (2007).

93. Adeva-Andany, M. M., Pérez-Felpete, N., Fernández-Fernández, C., Donapetry-García, C. & Pazos-García, C. Liver glucose metabolism in humans. *Biosci Rep* **36**, (2016).

94. Hou, Y., Hu, S., Li, X., He, W. & Wu, G. Amino acid metabolism in the liver: Nutritional and physiological significance. in *Advances in Experimental Medicine and Biology* vol. 1265 21–37 (Adv Exp Med Biol, 2020).

95. Chiang, J. Y. L. & Ferrell, J. M. Bile acid metabolism in liver pathobiology. *Gene Expr* **18**, 71–87 (2018).

96. Almazroo, O. A., Miah, M. K. & Venkataramanan, R. Drug Metabolism in the Liver. *Clin Liver Dis* **21**, 1–20 (2017).

97. Chen, M., Suzuki, A., Borlak, J., Andrade, R. J. & Lucena, M. I. Drug-induced liver injury: Interactions between drug properties and host factors. *J Hepatol* **63**, 503–514 (2015).

98. Antoine, D. J., Williams, D. P. & Park, B. K. Understanding the role of reactive metabolites in drug-induced hepatotoxicity: State of the science. *Expert Opin Drug Metab Toxicol* **4**, 1415–1427 (2008).

99. Park, B. K. *et al.* Drug bioactivation and protein adduct formation in the pathogenesis of drug-induced toxicity. in *Chemico-Biological Interactions* vol. 192 30–36 (Elsevier, 2011).

100. Limban, C. *et al.* The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicol Rep* **5**, 943–953 (2018).

101. Morio, B., Panthu, B., Bassot, A. & Rieusset, J. Role of mitochondria in liver metabolic health and diseases. *Cell Calcium* **94**, (2021).

102. Jaeschke, H. Mitochondrial dysfunction as a mechanism of drug-induced hepatotoxicity: current understanding and future perspectives. *J Clin Transl Res* **4**, 75 (2018).

103. Woodhead, J. L. *et al.* Exploring BSEP inhibition-mediated toxicity with a mechanistic model of drug-induced liver injury. *Front Pharmacol* **5**, 240 (2014).

104. Xie, Z. *et al.* Targeted Metabolomics Analysis of Bile Acids in Patients with Idiosyncratic Drug-Induced Liver Injury. *Metabolites* **11**, (2021).

105. Michalopoulos, G. K. & Bhushan, B. Liver regeneration: biological and pathological mechanisms and implications. *Nat Rev Gastroenterol Hepatol* **18**, 40–55 (2021).

106. Walesky, C. M. *et al.* Functional compensation precedes recovery of tissue mass following acute liver injury. *Nature Communications 2020 11:1* **11**, 1–12 (2020).

107. Martinez, F. J. *et al.* Idiopathic pulmonary fibrosis. *Nat Rev Dis Primers* **3**, (2017).

108. Roth, G. J. *et al.* Nintedanib: From discovery to the clinic. *J Med Chem* **58**, 1053–1063 (2015).

109. Spagnolo, P. *et al.* Idiopathic pulmonary fibrosis: Disease mechanisms and drug development. *Pharmacol Ther* **222**, 107798 (2021).

110. Ogura, T. *et al.* Safety and pharmacokinetics of nintedanib and pirfenidone in idiopathic pulmonary fibrosis. *European Respiratory Journal* **45**, 1382–1392 (2015).

111. Galli, J. A. *et al.* Pirfenidone and nintedanib for pulmonary fibrosis in clinical practice: Tolerability and adverse drug reactions. *Respirology* **22**, 1171–1178 (2017).

112. George, P. M., Patterson, C. M., Reed, A. K. & Thillai, M. Lung transplantation for idiopathic pulmonary fibrosis. *Lancet Respir Med* **7**, 271–282 (2019).

113. Vreman, R. A. *et al.* Unmet Medical Need: An Introduction to Definitions and Stakeholder Perceptions. *Value in Health* **22**, 1275–1282 (2019).

114. Lee, A. S., Mira-Avendano, I., Ryu, J. H. & Daniels, C. E. The burden of idiopathic pulmonary fibrosis: An unmet public health need. *Respir Med* **108**, 955–967 (2014).

115. Taskar, V. S. & Coultas, D. B. Is idiopathic pulmonary fibrosis an environmental disease? *Proc Am Thorac Soc* **3**, 293–298 (2006).

116. Desai, T. J., Brownfield, D. G. & Krasnow, M. A. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature 2014 507:7491* **507**, 190–194 (2014).

117. Barkauskas, C. E. *et al.* Type 2 alveolar cells are stem cells in adult lung. *J Clin Invest* **123**, 3025–3036 (2013).

118. Nabhan, A. N., Brownfield, D. G., Harbury, P. B., Krasnow, M. A. & Desai, T. J. Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science (1979)* **359**, 1118–1123 (2018).

119. Sisson, T. H. *et al.* Targeted injury of type II alveolar epithelial cells induces pulmonary fibrosis. *Am J Respir Crit Care Med* **181**, 254–263 (2010).

120. Parimon, T., Yao, C., Stripp, B. R., Noble, P. W. & Chen, P. Alveolar epithelial type II cells as drivers of lung fibrosis in idiopathic pulmonary fibrosis. *Int J Mol Sci* **21**, (2020).

121. King, T. E. *et al.* Idiopathic pulmonary fibrosis: relationship between histopathologic features and mortality. *Am J Respir Crit Care Med* **164**, 1025–1032 (2001).

122. Ptasinski, V. A., Stegmayr, J., Belvisi, M. G., Wagner, D. E. & Murray, L. A. Targeting alveolar repair in idiopathic pulmonary fibrosis. *Am J Respir Cell Mol Biol* **65**, 347–365 (2021).

123. Kamp, D. W. Idiopathic Pulmonary Fibrosis: The Inflammation Hypothesis Revisited. *Chest* **124**, 1187–1190 (2003).

124. Scotton, C. J. & Chambers, R. C. Molecular targets in pulmonary fibrosis: The myofibroblast in focus. *Chest* **132**, 1311–1321 (2007).

125. Yuan, T. *et al.* FGF10-FGFR2B Signaling Generates Basal Cells and Drives Alveolar Epithelial Regeneration by Bronchial Epithelial Stem Cells after Lung Injury. *Stem Cell Reports* **12**, 1041–1055 (2019).

126. Shi, Y. *et al.* Distal airway stem cells ameliorate bleomycin-induced pulmonary fibrosis in mice. *Stem Cell Res Ther* **10**, 1–11 (2019).

127. Williamson, J. D., Sadofsky, L. R. & Hart, S. P. The pathogenesis of bleomycin-induced lung injury in animals and its applicability to human idiopathic pulmonary fibrosis. *Exp Lung Res* **41**, 57–73 (2015).

128. Kobayashi, Y. *et al.* Persistence of a regeneration-associated, transitional alveolar epithelial cell state in pulmonary fibrosis. *Nat Cell Biol* **22**, 934–946 (2020).

129. Habermann, A. C. *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* **6**, (2020).

130. Adams, T. S. *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* **6**, (2020).

131. Liu, T., de Los Santos, F. G. & Phan, S. H. The Bleomycin Model of Pulmonary Fibrosis. *Methods in Molecular Biology* **1627**, 27–42 (2017).

132. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat Commun* **11**, 1–20 (2020).

133. Choi, J. *et al.* Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated Transient Progenitors that Mediate Alveolar Regeneration. *Cell Stem Cell* **27**, 366-382.e7 (2020).

134. Herholt, A., Galinski, S., Geyer, P. E., Rossner, M. J. & Wehr, M. C. Multiparametric Assays for Accelerating Early Drug Discovery. *Trends Pharmacol Sci* **41**, 318–335 (2020).

135. Horgan, R. P. & Kenny, L. C. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist* **13**, 189–195 (2011).

136. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 1–15 (2017).

137. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights* **14**, (2020).

138. Nelson, N. J. Microarrays have arrived: Gene expression tool matures. *J Natl Cancer Inst* **93**, 492–493 (2001).

139. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLoS Comput Biol* **13**, (2017).

140. Bumgarner, R. DNA microarrays: Types, applications, and their future. *Curr Protoc Mol Biol* (2013) doi:10.1002/0471142727.mb2201s101.

141. Heller, M. J. DNA microarray technology: Devices, systems, and applications. *Annu Rev Biomed Eng* **4**, 129–153 (2002).

142. Ragoussis, J. & Elvidge, G. Affymetrix GeneChip® system: Moving from research to the clinic. *Expert Rev Mol Diagn* **6**, 145–152 (2006).

143. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

144. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).

145. Li, X. & Wang, C. Y. From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci* **13**, 1–6 (2021).

146. Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* **12**, 44–73 (2017).

147. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

148. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 1–12 (2017).

149. Zhang, X. *et al.* Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol Cell* **73**, 130-142.e5 (2019).

150. Igarashi, Y. *et al.* Open TG-GATEs: A large-scale toxicogenomics database. *Nucleic Acids Res* **43**, D921–D927 (2015).

151. Ganter, B., Snyder, R. D., Halbert, D. N. & Lee, M. D. Toxicogenomics in drug discovery and development: Mechanistic analysis of compound/class-dependent effects using the DrugMatrix® database. *Pharmacogenomics* **7**, 1025–1044 (2006).

152. Alexander-Dann, B. *et al.* Developments in toxicogenomics: Understanding and predicting compound-induced toxicity from gene expression data. *Mol Omics* **14**, 218–236 (2018).

153. Chen, M., Zhang, M., Borlak, J. & Tong, W. A decade of toxicogenomic research and its contribution to toxicological science. *Toxicological Sciences* **130**, 217–228 (2012).

154. Ye, C. *et al.* DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat Commun* **9**, 1–9 (2018).

155. Lamb, J. *et al.* The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science (1979)* **313**, 1929–1935 (2006).

156. Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution. *Science (1979)* **367**, 45–51 (2020).

157. de Wolf, H. *et al.* High-Throughput Gene Expression Profiles to Define Drug Similarity and Predict Compound Activity. *Assay Drug Dev Technol* **16**, 162–176 (2018).

158. Musa, A. *et al.* Systems Pharmacogenomic Landscape of Drug Similarities from LINCS data: Drug Association Networks. *Sci Rep* **9**, 1–16 (2019).

159. Baillif, B., Wichard, J., Méndez-Lucio, O. & Rouquié, D. Exploring the Use of Compound-Induced Transcriptomic Data Generated From Cell Lines to Predict Compound Activity Toward Molecular Targets. *Front Chem* **8**, (2020).

160. Wang, Z., Clark, N. R. & Ma'ayan, A. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* **32**, 2338–2345 (2016).

161. Verbist, B. *et al.* Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. *Drug Discov Today* **20**, 505–513 (2015).

162. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

163. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics* **10**, 946–963 (2016).

164. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

165. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 1–21 (2014).

166. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, 258D – 261 (2004).

167. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649–D655 (2018).

168. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361 (2017).

169. Slenter, D. N. *et al.* WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* **46**, D661–D667 (2018).

170. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**, 1363–1375 (2019).

171. Smith, I. *et al.* Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biol* **15**, e2003213 (2017).

172. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607–D613 (2019).

173. Väremo, L., Nielsen, J. & Nookaew, I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res* **41**, 4378–4391 (2013).

174. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).

175. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).

176. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a World Beyond “p<0.05”. *https://doi.org/10.1080/00031305.2019.1583913* **73**, 1–19 (2019).

177. Groenwold, R. H. H., Goeman, J. J. & le Cessie, S. Multiple testing: when is many too much? *Eur J Endocrinol* **184**, E11–E14 (2021).

178. Bland, j. M. & Altman, D. G. Multiple significance tests: the Bonferroni method. *BMJ: British Medical Journal* **310**, 170 (1995).

179. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

180. FDA-NIH Biomarker Working Group. BEST (Biomarkers , EndpointS , and other Tools). 55 https://www.ncbi.nlm.nih.gov/books/NBK326791/ (2017).

181. Califf, R. M. Biomarker definitions and their applications. *Exp Biol Med* **243**, 213–221 (2018).

182. Robinson, W. H., Lindstrom, T. M., Cheung, R. K. & Sokolove, J. Mechanistic biomarkers for clinical decision making in rheumatic diseases. *Nat Rev Rheumatol* **9**, 267–276 (2013).

183. Goodsaid, F. M., Frueh, F. W. & Mattes, W. The Predictive Safety Testing Consortium: A synthesis of the goals, challenges and accomplishments of the Critical Path. *Drug Discov Today Technol* **4**, 47–50 (2007).

184. Schofield, A. L. *et al.* Systems analysis of miRNA biomarkers to inform drug safety. *Arch Toxicol* **95**, 3475–3495 (2021).

185. OECD (Organisation for Economic Co-operation and Development). *USERS' HANDBOOK SUPPLEMENT TO THE GUIDANCE DOCUMENT FOR DEVELOPING AND ASSESSING AOPs*. *OECD Environment, Health and Safety Publications Series on Testing and Assessment* (2018).

186. Hill, A. B. The Environment and Disease: Association or Causation? *Proc R Soc Med* 295–300 (1965).

187. Becker, R. A. *et al.* Quantitative weight of evidence to assess confidence in potential modes of action. *Regulatory Toxicology and Pharmacology* **86**, 205–220 (2017).

188. Spinu, N., Cronin, M. T. D., Enoch, S. J., Madden, J. C. & Worth, A. P. Quantitative adverse outcome pathway (qAOP) models for toxicity prediction. *Arch Toxicol* **94**, 1497–1510 (2020).

189. Oki, N. O., Nelms, M. D., Bell, S. M., Mortensen, H. M. & Edwards, S. W. Accelerating Adverse Outcome Pathway Development Using Publicly Available Data Sources. *Curr Environ Health Rep* **3**, 53–63 (2016).

190. Oki, N. O. & Edwards, S. W. An integrative data mining approach to identifying adverse outcome pathway signatures. *Toxicology* **350–352**, 49–61 (2016).

191. Bell, S. M., Angrish, M. M., Wood, C. E. & Edwards, S. W. Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver. *Toxicological Sciences* **150**, 510–520 (2016).

192. Zgheib, E. *et al.* Application of three approaches for quantitative AOP development to renal toxicity. *Computational Toxicology* **11**, 1–13 (2019).

193. Hassan, I. *et al.* Neurodevelopment and thyroid hormone synthesis inhibition in the rat: Quantitative understanding within the adverse outcome pathway framework. *Toxicological Sciences* **160**, 57–73 (2017).

194. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery 2018 18:1* **18**, 41–58 (2018).

195. Nosengo, N. Can you teach old drugs new tricks? *Nature* **534**, 314–316 (2016).

196. Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med* **23**, 405 (2017).

197. Bray, M. A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat Protoc* **11**, 1757–1774 (2016).

198. Litichevskiy, L. *et al.* A Library of Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations. *Cell Syst* **6**, 424 (2018).

199. Duran-Frigola, M. *et al.* Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat Biotechnol* **38**, 1087–1096 (2020).

200. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* **18**, 41–58 (2018).

201. Iorio, F., Rittman, T., Ge, H., Menden, M. & Saez-Rodriguez, J. Transcriptional data: A new gateway to drug repositioning? *Drug Discov Today* **18**, 350–357 (2013).

202. Wei, G. *et al.* Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* **10**, 331–342 (2006).

203. KalantarMotamedi, Y. *et al.* Transcriptional drug repositioning and cheminformatics approach for differentiation therapy of leukaemia cells. *Sci Rep* **11**, 1–18 (2021).

204. Dudley, J. T. *et al.* Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med* **3**, (2011).

205. Liu, A., Munoz-Muriedas, J., Bender, A. & Dalmas, D. A. Identification of potential biomarker candidates of drug-induced vascular injury (DIVI) in rats using gene expression and histopathology data. *bioRxiv* 2022.08.24.505120 (2022) doi:10.1101/2022.08.24.505120.

206. Bendjama, K. *et al.* Translation Strategy for the Qualification of Drug-induced Vascular Injury Biomarkers. *Toxicol Pathol* **42**, 658–671 (2014).

207. Mikaelian, I. *et al.* Nonclinical Safety Biomarkers of Drug-induced Vascular Injury:Current Status and Blueprint for the Future. *Toxicol Pathol* **42**, 635–657 (2014).

208. Brott, D. *et al.* Biomarkers of drug-induced vascular injury. in *Toxicology and Applied Pharmacology* vol. 207 441–445 (Academic Press Inc., 2005).

209. Louden, C. *et al.* Biomarkers and mechanisms of drug-induced vascular injury in non-rodents. *Toxicol Pathol* **34**, 19–26 (2006).

210. Weaver, J. L. *et al.* Biomarkers in peripheral blood associated with vascular injury in Sprague-Dawley rats treated with the phosphodiesterase IV inhibitors SCH 351591 or SCH 534385. *Toxicol Pathol* **36**, 840–849 (2008).

211. Heydarkhan-Hagvall, S. *et al.* DNA microarray study on gene expression profiles in co-cultured endothelial and smooth muscle cells in response to 4- and 24-h shear stress. *Mol Cell Biochem* **281**, 1–15 (2006).

212. Slim, R. M., Yunling Song, Albassam, M. & Dethloff, L. A. Apoptosis and Nitrative Stress Associated with Phosphodiesterase Inhibitor-Induced Mesenteric Vasculitis in Rats. *Toxicol Pathol* **31**, 638–645 (2003).

213. Zhang, J. *et al.* Mechanisms and biomarkers of cardiovascular injury induced by phosphodiesterase inhibitor III SK&F 95654 in the spontaneously hypertensive rat. *Toxicol Pathol* **34**, 152–163 (2006).

214. Weaver, J. L. *et al.* Early events in vascular injury in the rat induced by the phosphodiesterase IV inhibitor SCH 351591. *Toxicol Pathol* **38**, 738–744 (2010).

215. Daguès, N. *et al.* Altered gene expression in rat mesenteric tissue following in vivo exposure to a phosphodiesterase 4 inhibitor. *Toxicol Appl Pharmacol* **218**, 52–63 (2007).

216. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).

217. Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics - A bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–416 (2009).

218. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

219. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).

220. Jafari, M. & Ansari-Pour, N. Why, When and How to Adjust Your P Values? *Cell Journal (Yakhteh)* **20**, 604 (2019).

221. R Core Team. *R: A Language and Environment for Statistical Computing*. (2020).

222. Wickham, H. *tidyverse: Easily Install and Load the 'Tidyverse'*. (2017).

223. Wickham, H. *et al.* Welcome to the {tidyverse}. *J Open Source Softw* **4**, 1686 (2019).

224. Kuhn, M. & Vaughan, D. *yardstick: Tidy Characterizations of Model Performance*. (2020).

225. Shimoyama, M. *et al.* The Rat Genome Database 2015: Genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res* **43**, D743–D750 (2015).

226. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

227. Smedley, D. *et al.* The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Res* **43**, W589–W598 (2015).

228. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

229. Imanaka-Yoshida, K., Yoshida, T. & Miyagawa-Tomita, S. Tenascin-C in development and disease of blood vessels. *Anatomical Record* **297**, 1747–1757 (2014).

230. Daguès, N. *et al.* Investigation of the molecular mechanisms preceding PDE4 inhibitor-induced vasculopathy in rats: Tissue inhibitor of metalloproteinase 1, a potential predictive biomarker. *Toxicological Sciences* **100**, 238–247 (2007).

231. Chen, D. *et al.* Fibronectin signals through integrin α5β1 to regulate cardiovascular development in a cell type-specific manner. *Dev Biol* **407**, 195–210 (2015).

232. Lok, Z. S. Y. & Lyle, A. N. Osteopontin in Vascular Disease. *Arterioscler Thromb Vasc Biol* **39**, 613–622 (2019).

233. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **1**, 417–425 (2015).

234. Fu, B. M. & Tarbell, J. M. Mechano-sensing and transduction by endothelial surface glycocalyx: Composition, structure, and function. *Wiley Interdiscip Rev Syst Biol Med* **5**, 381–390 (2013).

235. Tagliabracci, V. S. *et al.* A Single Kinase Generates the Majority of the Secreted Phosphoproteome. *Cell* **161**, 1619–1632 (2015).

236. Naba, A. *et al.* The matrisome: In silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Molecular and Cellular Proteomics* **11**, M111.014647 (2012).

237. Ponticos, M. & Smith, B. D. Extracellular matrix synthesis in vascular disease: hypertension, and atherosclerosis. *J Biomed Res* **28**, 25 (2014).

238. Groft, L. L. *et al.* Differential expression and localization of TIMP-1 and TIMP-4 in human gliomas. *British Journal of Cancer 2001 85:1* **85**, 55–63 (2001).

239. Murphy, G. Tissue inhibitors of metalloproteinases. *Genome Biol* **12**, 233 (2011).

240. Liu, A., Han, N., Munoz-Muriedas, J. & Bender, A. Deriving time-concordant event cascades from gene expression data: A case study for Drug-Induced Liver Injury (DILI). *PLoS Comput Biol* **18**, e1010148 (2022).

241. Kohonen, P. *et al.* A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. *Nat Commun* **8**, (2017).

242. Sutherland, J. J. *et al.* Toxicogenomic module associations with pathogenesis: A network-based approach to understanding drug toxicity. *Pharmacogenomics Journal* **18**, 377–390 (2018).

243. Souza, T. M., Kleinjans, J. C. S. & Jennen, D. G. J. Dose and time dependencies in stress pathway responses during chemical exposure: Novel insights from gene regulatory networks. *Front Genet* **8**, 142–142 (2017).

244. Rooney, J., Hill, T., Qin, C., Sistare, F. D. & Christopher Corton, J. Adverse outcome pathway-driven identification of rat liver tumorigens in short-term assays. *Toxicol Appl Pharmacol* **356**, 99–113 (2018).

245. Rooney, J. *et al.* Activation of Nrf2 in the liver is associated with stress resistance mediated by suppression of the growth hormone-regulated STAT5b transcription factor. *PLoS One* **13**, (2018).

246. Andrade, R. J. *et al.* Drug-induced liver injury. *Nat Rev Dis Primers* **5**, 1–22 (2019).

247. Regev, A. Drug-induced liver injury and drug development: Industry perspective. *Semin Liver Dis* **34**, 227–239 (2014).

248. Aguayo-Orozco, A., Bois, F. Y., Brunak, S. & Taboureau, O. Analysis of Time-Series Gene Expression Data to Explore Mechanisms of Chemical-Induced Hepatic Steatosis Toxicity. *Front Genet* **9**, 396 (2018).

249. Zhang, J. D., Berntenis, N., Roth, A. & Ebeling, M. Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity. *Pharmacogenomics Journal* **14**, 208–216 (2014).

250. Thakkar, S. *et al.* Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity. *Drug Discov Today* **25**, 201–208 (2020).

251. Dolgalev, I. *msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format*. (2020).

252. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* **13**, 966–967 (2016).

253. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol Syst Biol* **17**, e9923 (2021).

254. Makowski, D., Ben-Shachar, M., Patil, I. & Lüdecke, D. Methods and Algorithms for Correlation Analysis in R. *J Open Source Softw* **5**, 2306 (2020).

255. Dewey, M. *metap: meta-analysis of significance values*. (2020).

256. Tillander, V., Alexson, S. E. H. & Cohen, D. E. Deactivating Fatty Acids: Acyl-CoA Thioesterase-Mediated Control of Lipid Metabolism. *Trends in Endocrinology and Metabolism* **28**, 473–484 (2017).

257. Violante, S. *et al.* Substrate specificity of human carnitine acetyltransferase: Implications for fatty acid and branched-chain amino acid metabolism. *Biochim Biophys Acta Mol Basis Dis* **1832**, 773–779 (2013).

258. Hong, L. *et al.* New role and molecular mechanism of Gadd45a in hepatic Fibrosis. *World J Gastroenterol* **22**, 2779–2788 (2016).

259. Okazaki, H. *et al.* Identification of neutral cholesterol ester hydrolase, a key enzyme removing cholesterol from macrophages. *Journal of Biological Chemistry* **283**, 33357–33364 (2008).

260. Chiba, M. *et al.* Elevation and characteristics of Rab30 and S100a8/S100a9 expression in an early phase of liver regeneration in the mouse. *Int J Mol Med* **27**, 567–574 (2011).

261. Liu, A. *et al.* From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *NPJ Syst Biol Appl* **5**, 1–10 (2019).

262. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat Commun* **9**, 20 (2018).

263. Shi, Z. *et al.* Transcriptional factor ATF3 promotes liver fibrosis via activating hepatic stellate cells. *Cell Death Dis* **11**, 1–16 (2020).

264. Copple, I. M. *et al.* The hepatotoxic metabolite of acetaminophen directly activates the keap1-Nrf2 cell defense system. *Hepatology* **48**, 1292–1301 (2008).

265. Luedde, T. & Schwabe, R. F. NF-κB in the liver-linking injury, fibrosis and hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol* **8**, 108–118 (2011).

266. Schultz, J. R. *et al.* Role of LXRs in control of lipogenesis. *Genes Dev* **14**, 2831–2838 (2000).

267. Hetz, C., Zhang, K. & Kaufman, R. J. Mechanisms, regulation and functions of the unfolded protein response. *Nat Rev Mol Cell Biol* **21**, 421–438 (2020).

268. Wijaya, L. S. *et al.* Integration of temporal single cell cellular stress response activity with logic-ODE modeling reveals activation of ATF4-CHOP axis as a critical predictor of drug-induced liver injury. *Biochem Pharmacol* **190**, 114591 (2021).

269. Fredriksson, L. *et al.* Drug-induced endoplasmic reticulum and oxidative stress responses independently sensitize toward TNFα-mediated hepatotoxicity. *Toxicological Sciences* **140**, 144–159 (2014).

270. Seki, E., Brenner, D. A. & Karin, M. A liver full of JNK: Signaling in regulation of cell function and disease pathogenesis, and clinical approaches. *Gastroenterology* **143**, 307–320 (2012).

271. Win, S. *et al.* New insights into the role and mechanism of c-Jun-N-terminal kinase signaling in the pathobiology of liver diseases. *Hepatology* **67**, 2013–2024 (2018).

272. Simmons, S. O., Fan, C.-Y. & Ramabhadran, R. Cellular Stress Response Pathway System as a Sentinel Ensemble in Toxicological Screening. *Toxicological Sciences* **111**, 202–225 (2009).

273. Wong, M. M. K., Joyson, S. M., Hermeking, H. & Chiu, S. K. Transcription factor AP4 mediates cell fate decisions: To divide, age, or die. *Cancers (Basel)* **13**, 1–15 (2021).

274. Zuo, L. *et al.* HOXB13 expression is correlated with hepatic inflammatory activity of patients with hepatic fibrosis. *J Mol Histol* **51**, 183–189 (2020).

275. Delgado, I. *et al.* A role for transcription factor E2F2 in hepatocyte proliferation and timely liver regeneration. *Am J Physiol Gastrointest Liver Physiol* **301**, (2011).

276. B'Chir, W. *et al.* The eIF2α/ATF4 pathway is essential for stress-induced autophagy gene expression. *Nucleic Acids Res* **41**, 7683–7699 (2013).

277. Wijaya, L. S. *et al.* Integration of temporal single cell cellular stress response activity with logic-ODE modeling reveals activation of ATF4-CHOP axis as a critical predictor of drug-induced liver injury. *Biochem Pharmacol* **190**, 114591 (2021).

278. Pavel, M. *et al.* CCT complex restricts neuropathogenic protein aggregation via autophagy. *Nat Commun* **7**, 1–18 (2016).

279. Grantham, J. The Molecular Chaperone CCT/TRiC: An Essential Component of Proteostasis and a Potential Modulator of Protein Aggregation. *Front Genet* **11**, (2020).

280. Reebye, V. *et al.* Gene activation of CEBPA using saRNA: Preclinical studies of the first in human saRNA drug candidate for liver cancer. *Oncogene* **37**, 3216–3228 (2018).

281. Fusakio, M. E. *et al.* Transcription factor ATF4 directs basal and stress-induced gene expression in the unfolded protein response and cholesterol metabolism in the liver. *Mol Biol Cell* **27**, 1536–1551 (2016).

282. Hao, L. *et al.* ATF4 activation promotes hepatic mitochondrial dysfunction by repressing NRF1-TFAM signalling in alcoholic steatohepatitis. *Gut* **70**, 1933–1945 (2021).

283. Larigot, L., Juricek, L., Dairou, J. & Coumoul, X. AhR signaling pathways and regulatory functions. *Biochim Open* **7**, 1–9 (2018).

284. Kersten, S. & Stienstra, R. The role and regulation of the peroxisome proliferator activated receptor alpha in human liver. *Biochimie* **136**, 75–84 (2017).

285. Yoshikawa, T. *et al.* Cross-talk between peroxisome proliferator-activated receptor (PPAR) α and liver X receptor (LXR) in nutritional regulation of fatty acid metabolism. I. PPARS suppress sterol regulatory element binding protein-1c promoter through inhibition of LXR signali. *Molecular Endocrinology* **17**, 1240–1254 (2003).

286. Yoshikawa, T. *et al.* Identification of Liver X Receptor-Retinoid X Receptor as an Activator of the Sterol Regulatory Element-Binding Protein 1c Gene Promoter. *Mol Cell Biol* **21**, 2991–3000 (2001).

287. Boergesen, M. *et al.* Genome-Wide Profiling of Liver X Receptor, Retinoid X Receptor, and Peroxisome Proliferator-Activated Receptor in Mouse Liver Reveals Extensive Sharing of Binding Sites. *Mol Cell Biol* **32**, 852–867 (2012).

288. Kusumanchi, P. *et al.* Stress-Responsive Gene FK506-Binding Protein 51 Mediates Alcohol-Induced Liver Injury Through the Hippo Pathway and Chemokine (C-X-C Motif) Ligand 1 Signaling. *Hepatology* **74**, 1234–1250 (2021).

289. Manmadhan, S. & Ehmer, U. Hippo signaling in the liver - A long and ever-expanding story. *Front Cell Dev Biol* **7**, 33 (2019).

290. Kyrmizi, I. *et al.* Plasticity and expanding complexity of the hepatic transcription factor network during liver development. *Genes Dev* **20**, 2293–2305 (2006).

291. Nishikawa, T. *et al.* Resetting the transcription factor network reverses terminal chronic hepatic failure. *Journal of Clinical Investigation* **125**, 1533–1544 (2015).

292. Schulte, D. & Geerts, D. MEIS transcription factors in development and disease. *Development (Cambridge)* **146**, (2019).

293. Berenguer, M. & Duester, G. Role of retinoic acid signaling, FGF signaling and meis genes in control of limb development. *Biomolecules* **11**, 1–11 (2021).

294. Farooque, U. *et al.* The Pattern of Dyslipidemia in Chronic Liver Disease Patients. *Cureus* **13**, (2021).

295. Oliva, L., D'Incà, R., Medici, V. & Sturniolo, G. C. Metallothioneins and liver diseases. in *Metallothioneins in Biochemistry and Pathology* 289–316 (World Scientific Publishing Co., 2008). doi:10.1142/9789812778949_0014.

296. Huang, G. W. & Yang, L. Y. Metallothionein expression in hepatocellular carcinoma. *World J Gastroenterol* **8**, 650–653 (2002).

297. Devisscher, L. *et al.* Metallothioneins alter macrophage phenotype and represent novel therapeutic targets for acetaminophen-induced liver injury. *J Leukoc Biol* **111**, 123–133 (2022).

298. Lefebvre, V. The SoxD transcription factors - Sox5, Sox6, and Sox13 - are key cell fate modulators. *International Journal of Biochemistry and Cell Biology* **42**, 429–432 (2010).

299. Wang, Y., Ristevski, S. & Harley, V. R. SOX13 exhibits a distinct spatial and temporal expression pattern during chondrogenesis, neurogenesis, and limb development. *Journal of Histochemistry and Cytochemistry* **54**, 1327–1333 (2006).

300. Liu, L. *et al.* The microenvironment in hepatocyte regeneration and function in rats with advanced cirrhosis. *Hepatology* **55**, 1529–1539 (2012).

301. Guzman-Lepe, J. *et al.* Liver-enriched transcription factor expression relates to chronic hepatic failure in humans. *Hepatol Commun* **2**, 582–594 (2018).

302. Yamagishi, S. I. & Matsui, T. Role of receptor for advanced glycation end products (RAGE) in liver disease. *Eur J Med Res* **20**, (2015).

303. Zeng, M., Liu, W., Hu, Y. & Fu, N. Sumoylation in liver disease. *Clinica Chimica Acta* **510**, 347–353 (2020).

304. Perrier, V., Meyer, F. & Granjon, D. *shinyWidgets: Custom Inputs Widgets for Shiny*. (2022).

305. Bailey, E. *shinyBS: Twitter Bootstrap Components for Shiny*. (2022).

306. Sali, A. & Attali, D. *shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating*. (2020).

307. Sievert, C. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. (Chapman and Hall/CRC, 2020).

308. Gohel, D. & Skintzos, P. *ggiraph: Make 'ggplot2' Graphics Interactive*. (2022).

309. Sidi, J. & Galili, T. *shinyHeatmaply: Deploy 'heatmaply' using 'shiny'*. (2020).

310. Xie, Y., Cheng, J. & Tan, X. *DT: A Wrapper of the JavaScript Library 'DataTables'*. (2022).

311. Geng, L. & Hamilton, H. J. Interestingness measures for data mining. *ACM Computing Surveys (CSUR)* **38**, 3 (2006).

312. Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. Dynamic Itemset Counting and Implication Rules for Market Basket Data. *SIGMOD Record (ACM Special Interest Group on Management of Data)* **26**, 255–264 (1997).

313. Liu, A., Han, N., Munoz-Muriedas, J. & Bender, A. Deriving time-concordant event cascades from gene expression data: A case study for Drug-Induced Liver Injury (DILI). *PLoS Comput Biol* **18**, e1010148 (2022).

314. Ortiz, C. *et al.* Extracellular Matrix Remodeling in Chronic Liver Disease. *Curr Tissue Microenviron Rep* **2**, 41–52 (2021).

315. Shinde, A. v., Humeres, C. & Frangogiannis, N. G. The role of α-smooth muscle actin in fibroblast-mediated matrix contraction and remodeling. *Biochim Biophys Acta Mol Basis Dis* **1863**, 298–309 (2017).

316. Rockey, D. C., Weymouth, N. & Shi, Z. Smooth muscle α actin (Acta2) and myofibroblast function during hepatic wound healing. *PLoS One* **8**, e77166 (2013).

317. Iwaisako, K., Brenner, D. A. & Kisseleva, T. What's new in liver fibrosis? The origin of myofibroblasts in liver fibrosis. *Journal of Gastroenterology and Hepatology (Australia)* **27**, 65–68 (2012).

318. Charan, H. V., Dwivedi, D. K., Khan, S. & Jena, G. Mechanisms of NLRP3 inflammasome-mediated hepatic stellate cell activation: Therapeutic potential for liver fibrosis. *Genes Dis* (2022) doi:10.1016/j.gendis.2021.12.006.

319. Krishna, M. Patterns of necrosis in liver disease. *Clin Liver Dis (Hoboken)* **10**, 53 (2017).

320. Koyama, Y. & Brenner, D. A. Liver inflammation and fibrosis. *J Clin Invest* **127**, 55 (2017).

321. Liu, A., Lee, J.-H., Han, N. & Bender, A. scRNA-Seq-based drug repurposing targeting idiopathic pulmonary fibrosis (IPF). *bioRxiv* 2022.09.17.508360 (2022) doi:10.1101/2022.09.17.508360.

322. Nayak, R. & Hasija, Y. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics* **113**, 606–619 (2021).

323. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746 (2019).

324. Trump, S. *et al.* Hypertension delays viral clearance and exacerbates airway hyperinflammation in patients with COVID-19. *Nat Biotechnol* **39**, 705–716 (2021).

325. Wang, R. *et al.* Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. *Nat Med* **27**, 141–151 (2021).

326. Xu, N. *et al.* STING agonist promotes CAR T cell trafficking and persistence in breast cancer. *Journal of Experimental Medicine* **218**, (2021).

327. Srivastava, S. *et al.* Immunogenic Chemotherapy Enhances Recruitment of CAR-T Cells to Lung Tumors and Improves Antitumor Efficacy when Combined with Checkpoint Blockade. *Cancer Cell* **39**, 193-208.e10 (2021).

328. Wang, Z. *et al.* Repurposable drugs for SARS-CoV-2 and influenza sepsis with scRNA-seq data targeting post-transcription modifications. *Precis Clin Med* **4**, 215–230 (2021).

329. He, B. & Garmire, L. X. ASGARD: A Single-cell Guided pipeline to Aid Repurposing of Drugs. *ArXiv* (2021).

330. Alakwaa, F. M. Repurposing Didanosine as a Potential Treatment for COVID-19 Using Single-Cell RNA Sequencing Data. *mSystems* **5**, (2020).

331. Hemberger, M., Dean, W. & Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: Digging Waddington's canal. *Nat Rev Mol Cell Biol* **10**, 526–537 (2009).

332. Brum, A. M. *et al.* Connectivity Map-based discovery of parbendazole reveals targetable human osteogenic pathway. *Proc Natl Acad Sci U S A* **112**, 12711–12716 (2015).

333. Brum, A. M. *et al.* Using the Connectivity Map to discover compounds influencing human osteoblast differentiation. *J Cell Physiol* **233**, 4895–4906 (2018).

334. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945–D954 (2017).

335. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

336. Germain, P.-L., Lun, A., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res* **10**, 979 (2021).

337. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 1–15 (2019).

338. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

339. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).

340. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438 (2016).

341. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nature Communications 2021 12:1* **12**, 1–15 (2021).

342. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 Preprint at https://doi.org/10.1101/060012 (2021).

343. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics 2016 48:8* **48**, 838–847 (2016).

344. Cplex IBM ILOG. *V12. 1: User's Manual for CPLEX*. *International Business Machines Corporation* vol. 46 (2009).

345. Melas, I. N. *et al.* Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. *Integr. Biol.* **7**, 904–920 (2015).

346. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

347. Szalai, B. *et al.* Signatures of cell death and proliferation in perturbation transcriptomics data—from confounding factor to effective prediction. *Nucleic Acids Res* **47**, 10010–10026 (2019).

348. de Almeida, B., Saraiva-Agostinho, N. & Barbosa-Morais, N. *cTRAP: Identification of candidate causal perturbations from differential gene expression data*. (2020).

349. Davies, M. *et al.* ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* **43**, W612–W620 (2015).

350. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489 (2021).

351. Freissmuth, M., Offermanns, S. & Böhm, S. *Pharmakologie und Toxikologie*. (Springer Berlin Heidelberg, 2016). doi:10.1007/978-3-662-46689-6.

352. Papatheodorou, I. *et al.* Expression Atlas update: From tissues to single cells. *Nucleic Acids Res* **48**, D77–D83 (2020).

353. Jain, R. *et al.* Plasticity of Hopx+ type I alveolar cells to regenerate type II cells in the lung. *Nat Commun* **6**, 1–11 (2015).

354. Wang, Y. *et al.* Pulmonary alveolar type I cell population consists of two distinct subtypes that differ in cell fate. *Proc Natl Acad Sci U S A* **115**, 2407–2412 (2018).

355. Shi, Z. *et al.* Heterogeneous ribosomes preferentially translate distinct subpools of mRNAs genome-wide. *Mol Cell* **67**, 71 (2017).

356. el Khoury, W. & Nasr, Z. Deregulation of ribosomal proteins in human cancers. *Biosci Rep* **41**, (2021).

357. Han, Z., Zhang, Q., Zhu, Y., Chen, J. & Li, W. Ribosomes: An Exciting Avenue in Stem Cell Research. *Stem Cells Int* **2020**, (2020).

358. Vuoriluoto, K. *et al.* Syndecan-1 supports integrin alpha2beta1-mediated adhesion to collagen. *Exp Cell Res* **314**, 3369–3381 (2008).

359. McKleroy, W., Lee, T. H. & Atabai, K. Always cleave up your mess: Targeting collagen degradation to treat tissue fibrosis. *Am J Physiol Lung Cell Mol Physiol* **304**, L709 (2013).

360. Wang, L. *et al.* Differences between mice and humans in regulation and the molecular network of collagen, type III, alpha-1 at the gene expression level: Obstacles that translational research must overcome. *Int J Mol Sci* **16**, 15031–15056 (2015).

361. van Leer, C., Stutz, M., Haeberli, A. & Geiser, T. Urokinase plasminogen activator released by alveolar epithelial cells modulates alveolar epithelial repair in vitro. *Thromb Haemost* **94**, 1257–1264 (2005).

362. Seki, Y. PRDM14 Is a Unique Epigenetic Regulator Stabilizing Transcriptional Networks for Pluripotency. *Front Cell Dev Biol* **6**, 12 (2018).

363. Desai, T. J., Brownfield, D. G. & Krasnow, M. A. Alveolar progenitor and stem cells in lung development, renewal and cancer. **507**, 190–194 (2014).

364. Marjanovic, N. D. *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229-246.e13 (2020).

365. Balis, J. U., Bumgarner, S. D., Paciga, J. E., Paterson, J. F. & Shelley, S. A. Synthesis of lung surfactant-associated glycoproteins by A549 cells: Description of an in vitro model for human type II cell dysfunction. *Exp Lung Res* **6**, 197–213 (1984).

366. Lieber, M., Todaro, G., Smith, B., Szakal, A. & Nelson-Rees, W. A continuous tumor-cell line from a human lung carcinoma with properties of type II alveolar epithelial cells. *Int J Cancer* **17**, 62–70 (1976).

367. Garcia-De-Alba, C. Repurposing A549 adenocarcinoma cells: New options for drug discovery. *Am J Respir Cell Mol Biol* **64**, 405–406 (2021).

368. Swain, R. J., Kemp, S. J., Goldstraw, P., Tetley, T. D. & Stevens, M. M. Assessment of Cell Line Models of Primary Human Cells by Raman Spectral Phenotyping. *Biophys J* **98**, 1703–1711 (2010).

369. Coutinho, A. E. & Chapman, K. E. The anti-inflammatory and immunosuppressive effects of glucocorticoids, recent developments and mechanistic insights. *Mol Cell Endocrinol* **335**, 2 (2011).

370. Richeldi, L., Davies, H. R. H. R., Spagnolo, P. & Luppi, F. Corticosteroids for idiopathic pulmonary fibrosis. *Cochrane Database Syst Rev* **2003**, (2003).

371. Arai, T. & Inoue, Y. Is corticosteroid use truly not associated with improved outcomes in AE-IPF? *Respirology* **25**, 659 (2020).

372. Farrand, E., Vittinghoff, E., Ley, B., Butte, A. J. & Collard, H. R. Corticosteroid use is not associated with improved outcomes in acute exacerbation of IPF. *Respirology* **25**, 629–635 (2020).

373. Kost-Alimova, M. *et al.* A High-Content Screen for Mucin-1-Reducing Compounds Identifies Fostamatinib as a Candidate for Rapid Repurposing for Acute Lung Injury. *Cell Rep Med* **1**, 100137 (2020).

374. Strich, J. R. *et al.* Fostamatinib for the Treatment of Hospitalized Adults With Coronavirus Disease 2019: A Randomized Trial. *Clinical Infectious Diseases* **75**, e491–e498 (2021).

375. de Ruijter, A. J. M., van Gennip, A. H., Caron, H. N., Kemp, S. & van Kuilenburg, A. B. P. Histone deacetylases (HDACs): Characterization of the classical HDAC family. *Biochemical Journal* **370**, 737–749 (2003).

376. Bondarev, A. D. *et al.* Recent developments of HDAC inhibitors: Emerging indications and novel molecules. *Br J Clin Pharmacol* **87**, 4577–4597 (2021).

377. Lyu, X., Hu, M., Peng, J., Zhang, X. & Sanders, Y. Y. HDAC inhibitors as antifibrotic drugs in cardiac and pulmonary fibrosis. *Ther Adv Chronic Dis* **10**, 2040622319862697 (2019).

378. Ye, Q. *et al.* Prevention of pulmonary fibrosis via trichostatin A (TSA) in bleomycin induced rats. *Sarcoidosis Vasculitis and Diffuse Lung Diseases* **31**, 219–226 (2014).

379. Korfei, M. *et al.* Comparison of the antifibrotic effects of the pan-histone deacetylase-inhibitor panobinostat versus the IPF-drug pirfenidone in fibroblasts from patients with idiopathic pulmonary fibrosis. *PLoS One* **13**, e0207915 (2018).

380. Colunga Biancatelli, R. M. L., Solopov, P., Gregory, B. & Catravas, J. D. HSP90 Inhibition and Modulation of the Proteome: Therapeutical Implications for Idiopathic Pulmonary Fibrosis (IPF). *International Journal of Molecular Sciences 2020, Vol. 21, Page 5286* **21**, 5286 (2020).

381. Sibinska, Z. *et al.* Amplified canonical transforming growth factor-β signalling via heat shock protein 90 in pulmonary fibrosis. *European Respiratory Journal* **49**, (2017).

382. LaCanna, R. *et al.* Yap/Taz regulate alveolar regeneration and resolution of lung inflammation. *Journal of Clinical Investigation* **129**, 2107–2122 (2019).

383. Xi, Y. *et al.* Local lung hypoxia determines epithelial fate decisions during alveolar regeneration. *Nat Cell Biol* **19**, 904–914 (2017).

384. Habermehl, D. *et al.* Glucocorticoid Activity during Lung Maturation Is Essential in Mesenchymal and Less in Alveolar Epithelial Cells. *Molecular Endocrinology* **25**, 1280 (2011).

385. Knight, Z. A., Lin, H. & Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nat Rev Cancer* **10**, 130–137 (2010).

386. Jiang, C., Guo, Y., Yu, H., Lu, S. & Meng, L. Pleiotropic microRNA-21 in pulmonary remodeling: Novel insights for molecular mechanism and present advancements. *Allergy, Asthma and Clinical Immunology* **15**, 1–9 (2019).

387. Eid, S., Turk, S., Volkamer, A., Rippmann, F. & Fulle, S. Kinmap: A web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics* **18**, 1–6 (2017).

388. Daniels, C. E. *et al.* Imatinib treatment for idiopathic pulmonary fibrosis: Randomized placebo-controlled trial results. *Am J Respir Crit Care Med* **181**, 604–610 (2010).

389. Justice, J. N. *et al.* Senolytics in idiopathic pulmonary fibrosis: Results from a first-in-human, open-label, pilot study. *EBioMedicine* **40**, 554–563 (2019).

390. Aono, Y. *et al.* Imatinib as a novel antifibrotic agent in bleomycin-induced pulmonary fibrosis in mice. *Am J Respir Crit Care Med* **171**, 1279–1285 (2005).

391. Yilmaz, O., Oztay, F. & Kayalar, O. Dasatinib attenuated bleomycin-induced pulmonary fibrosis in mice. *Growth Factors* **33**, 366–375 (2015).

392. Khatri, A. *et al.* ABL kinase inhibition promotes lung regeneration through expansion of an SCGB1A1+ SPC+ cell population following bacterial pneumonia. *Proc Natl Acad Sci U S A* **116**, 1603–1612 (2019).

393. Moeller, A., Ask, K., Warburton, D., Gauldie, J. & Kolb, M. The bleomycin animal model: a useful tool to investigate treatment options for idiopathic pulmonary fibrosis? *Int J Biochem Cell Biol* **40**, 362 (2008).

394. Tashiro, J. *et al.* Exploring animal models that resemble idiopathic pulmonary fibrosis. *Front Med (Lausanne)* **4**, 118 (2017).

395. Deconinck, L., Cannoodt, R., Saelens, W., Deplancke, B. & Saeys, Y. Recent advances in trajectory inference from single-cell omics data. *Curr Opin Syst Biol* **27**, 100344 (2021).

396. la Manno, G. *et al.* RNA velocity of single cells. *Nature 2018 560:7719* **560**, 494–498 (2018).

397. Kim, E. C. & Kim, J. R. Senotherapeutics: emerging strategy for healthy aging and age-related disease. *BMB Rep* **52**, 47 (2019).

398. Lu, B. & Atala, A. Small molecules and small molecule drugs in regenerative medicine. *Drug Discov Today* **19**, 801–808 (2014).

Appendix A

Table A.1 Frequency of mesenteric arterial necrosis (MAN) across treatments.

For each treatment and dose, the number of animals with MAN is shown followed by the number of animals without MAN.

Treatment	Histopathology				Endothelium				Smooth muscle			
Dose index	0	1	2	3	0	1	2	3	0	1	2	3
Dopamine (1day)	0/5	0/6	1/5	1/3	0/5	0/6	1/5	1/3	0/5	0/6	1/5	1/3
SKF-82526 (1day)	0/10	0/11	0/11	2/10	0/10	0/8	0/9	2/9	0/10	0/10	0/11	1/10
Dopamine (4days)	0/10	1/11	2/10	7/1	0/10	1/10	2/9	4/1	0/9	0/10	2/7	6/1
SKF-82526 (4days)	0/9	0/9	0/8	4/4	0/9	0/8	0/8	4/3	0/9	0/9	0/8	3/4
Methoxamine (4days)	0/6	0/6	0/6	2/4	0/5	0/6	0/5	2/4	0/6	0/5	0/4	2/4
Midodrine (4days)	0/6	0/6	3/3	4/0	0/5	0/5	3/3	4/0	0/5	0/6	3/3	4/0
SKF-95654 (4days)	0/6	0/6	1/5	5/0	0/6	0/5	1/5	4/0	0/5	0/6	1/5	5/0
Amphetamine (4days)	0/6	0/6	0/6	0/0	0/5	0/6	0/6	0/0	0/5	0/6	0/6	0/0
S-Propranolol (4days)	0/6	0/5	0/6	0/6	0/5	0/3	0/5	0/6	0/6	0/5	0/5	0/4
Yohimbine (4days)	0/6	0/6	0/6	0/6	0/5	0/6	0/6	0/4	0/6	0/5	0/5	0/6
Hydralazine (4days)	0/5	0/5	0/6	0/6	0/5	0/5	0/6	0/6	0/5	0/5	0/6	0/6
Sodium Nitroprusside (4days)	0/5	0/5	0/6	0/5	0/5	0/5	0/6	0/5	0/5	0/5	0/6	0/5
Minoxidil (4days)	0/6	0/6	0/6	0/5	0/5	0/5	0/6	0/5	0/6	0/5	0/6	0/4
GW788388 (4days)	0/6	0/6	0/6	0/6	0/5	0/5	0/6	0/6	0/6	0/6	0/4	0/6

Table A.2: Number of samples by Medial Arterial Necrosis (MAN) group.

The number of samples with gene expression measured in the endothelium and smooth muscle is shown based on the histopathology observed in the respective animal and treatment groups. For DIVI conditions (>20% MAN), the number of samples is shown by severity of MAN in the respective animal. Other samples from animals without MAN were groups based on whether these animals were treated with vehicle control, a compound treated as negative control (no histological changes observed in the medial artery) or a compound which did not show MAN but other histological changes in the mesenteric artery, such as perivascular and/or fibrinoid necrosis, perivascular fibrosis, EC hypertrophy and/or inflammatory cell infiltration.

MAN severity	**0**	**0**	**0**	**0**	**1**	**2**	**3**	**4**
Subgroup	Control	NoDIVI condition	Other treatment	DIVI condition				
Endothelium	85	42	164	14	10	11	3	5
Smooth muscle	88	42	166	15	9	9	6	5

Table A.3: Functional protein-protein interactions between conserved proteins derived from STRING.

The STRING scores for all observed sources of evidence are shown for all associations with at least middle confidence requiring a combined probability that the interaction is true, or short combined score, above 0.4. The two proteins, Protein 1 and Protein 2, are thereby provided in alphabetical order.

Protein 1	Protein 2	Homology	Co-expression	Experiment	Database	Text-mining	Combined score
Cd44	Spp1	0	0.088	0.181	0.9	0.857	0.987
Fn1	Itga5	0	0.079	0.494	0.9	0.763	0.987
Fn1	Serpine1	0	0.12	0.055	0.9	0.79	0.98
Spp1	Timp1	0	0.257	0	0.9	0.757	0.98
Fn1	Spp1	0	0.064	0	0.9	0.801	0.979
Fn1	Timp1	0	0.178	0	0.9	0.717	0.974
Serpine1	Timp1	0	0.172	0	0.9	0.713	0.974
Fn1	Vcan	0	0.102	0	0.9	0.721	0.972
Cd44	Fn1	0	0.08	0.181	0.8	0.724	0.952
Timp1	Vcan	0	0.122	0	0.9	0.498	0.952
Fn1	Tnc	0.584	0.149	0.181	0.9	0.499	0.939
Spp1	Vcan	0	0.063	0	0.9	0.371	0.935
Tnc	Vcan	0	0.166	0.174	0.9	0.157	0.934
Itga5	Spp1	0	0.063	0.148	0.9	0.199	0.927
Itgav	Tnc	0	0.063	0.212	0.9	0.064	0.921
Timp1	Tnc	0	0.148	0	0.9	0.14	0.92
Fn1	Itgav	0	0.064	0.801	0	0.6	0.919
Penk	Timp1	0	0	0	0.9	0.144	0.91
Spp1	Tnc	0	0.063	0	0.9	0.112	0.909
Fn1	Penk	0	0	0	0.9	0.064	0.902
Penk	Vcan	0	0.063	0	0.9	0	0.902
Penk	Tnc	0	0.063	0	0.9	0	0.902
Penk	Spp1	0	0	0	0.9	0.059	0.901
Col4a1	Itga5	0	0.082	0.067	0.6	0.336	0.742
Col4a1	Fn1	0	0.323	0.152	0	0.552	0.72
Cd44	Vcan	0	0.066	0.181	0	0.646	0.706
Kif22	Prc1	0	0.554	0.18	0	0.235	0.695
Anln	Prc1	0	0.585	0	0	0.274	0.686
Itga5	Tnc	0	0.063	0.212	0.6	0.057	0.684
Anlnl1	Prc1	0	0.576	0	0	0.274	0.679
Cd44	Col4a1	0	0.061	0.09	0.6	0.121	0.659
Cd44	Timp1	0	0.223	0	0	0.577	0.658
Cenpn	Prc1	0	0.617	0	0	0	0.617
Cd44	Itgav	0	0.063	0.108	0	0.561	0.601
Fn1	S100a4	0	0.117	0	0	0.561	0.596
Kpna2	Prc1	0	0.58	0	0	0	0.58
Serpine1	Tnfrsf12a	0	0.482	0	0	0.222	0.58
Timp1	Tnfrsf12a	0	0.495	0	0	0.2	0.579
Serpine1	Spp1	0	0.09	0	0	0.542	0.565
Cenpn	Kif22	0	0.551	0	0	0	0.551
S100a4	Timp1	0	0.366	0	0	0.302	0.539
Itgav	Spp1	0	0.063	0.181	0	0.417	0.514
Plp2	Timp1	0	0.512	0	0	0	0.512
Cd44	Itga5	0	0.064	0.108	0	0.459	0.509
Prc1	Tubb4b	0	0.064	0.396	0	0.175	0.492
Praf2	S100a4	0	0.063	0	0	0.47	0.482
Col4a1	Timp1	0	0.149	0.086	0	0.371	0.468
Cd44	S100a4	0	0.203	0	0	0.321	0.436
S100a4	Spp1	0	0.13	0	0	0.345	0.406
Plp2	S100a4	0	0.402	0	0	0	0.402
Cenpn	Kpna2	0	0.4	0	0	0	0.4

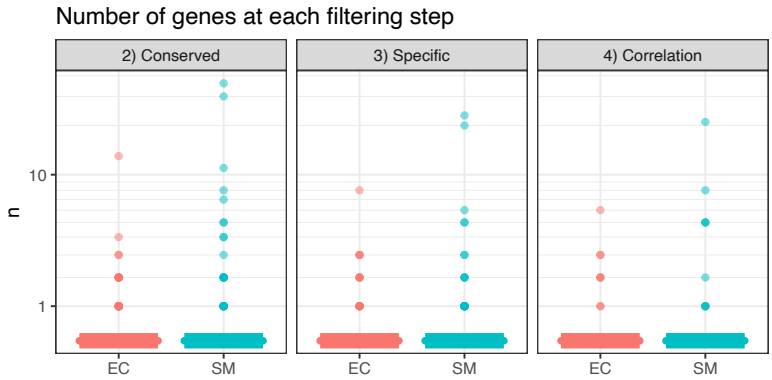


Figure A.1: Number of genes identified in filtering with permutated labels.

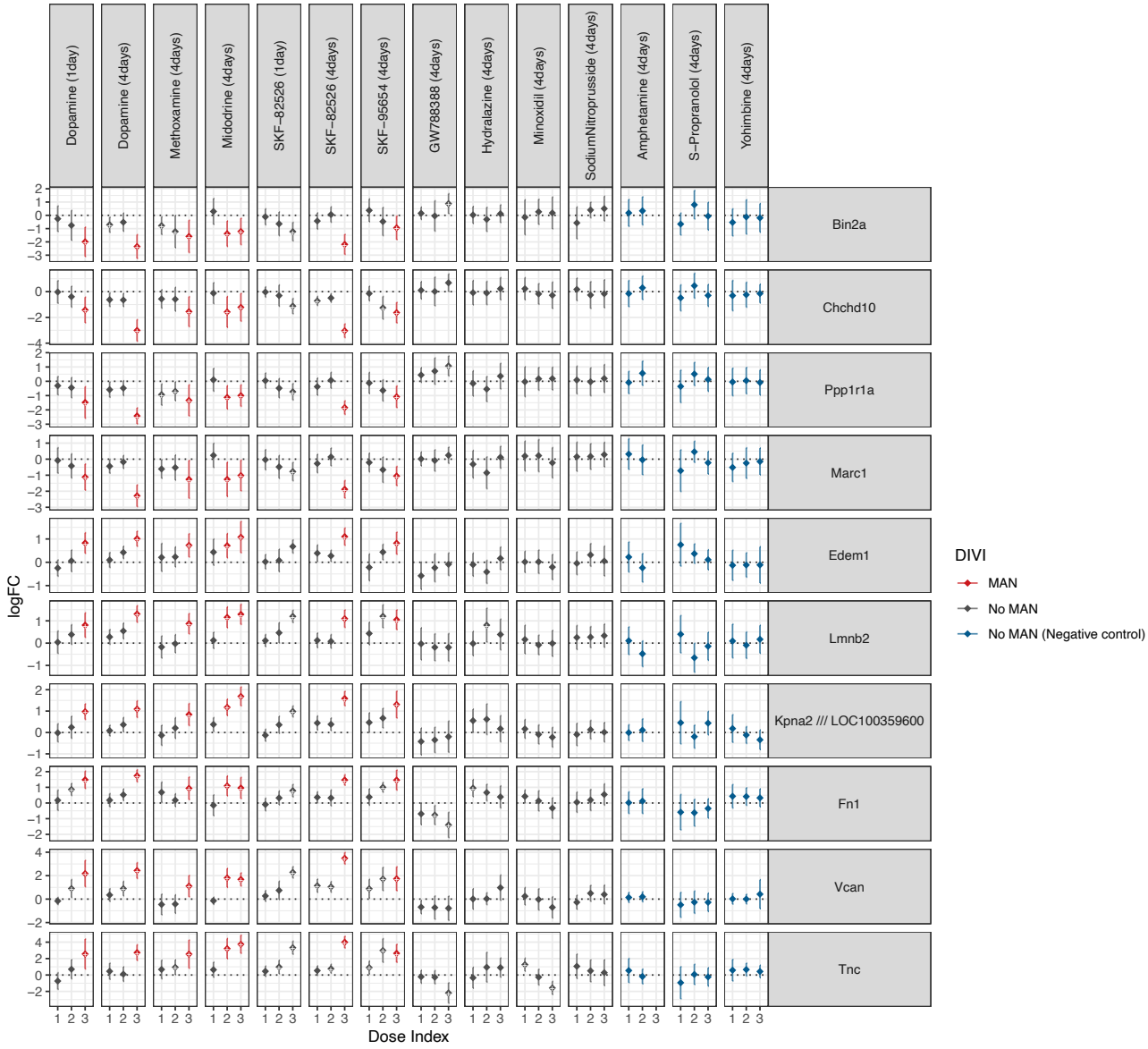
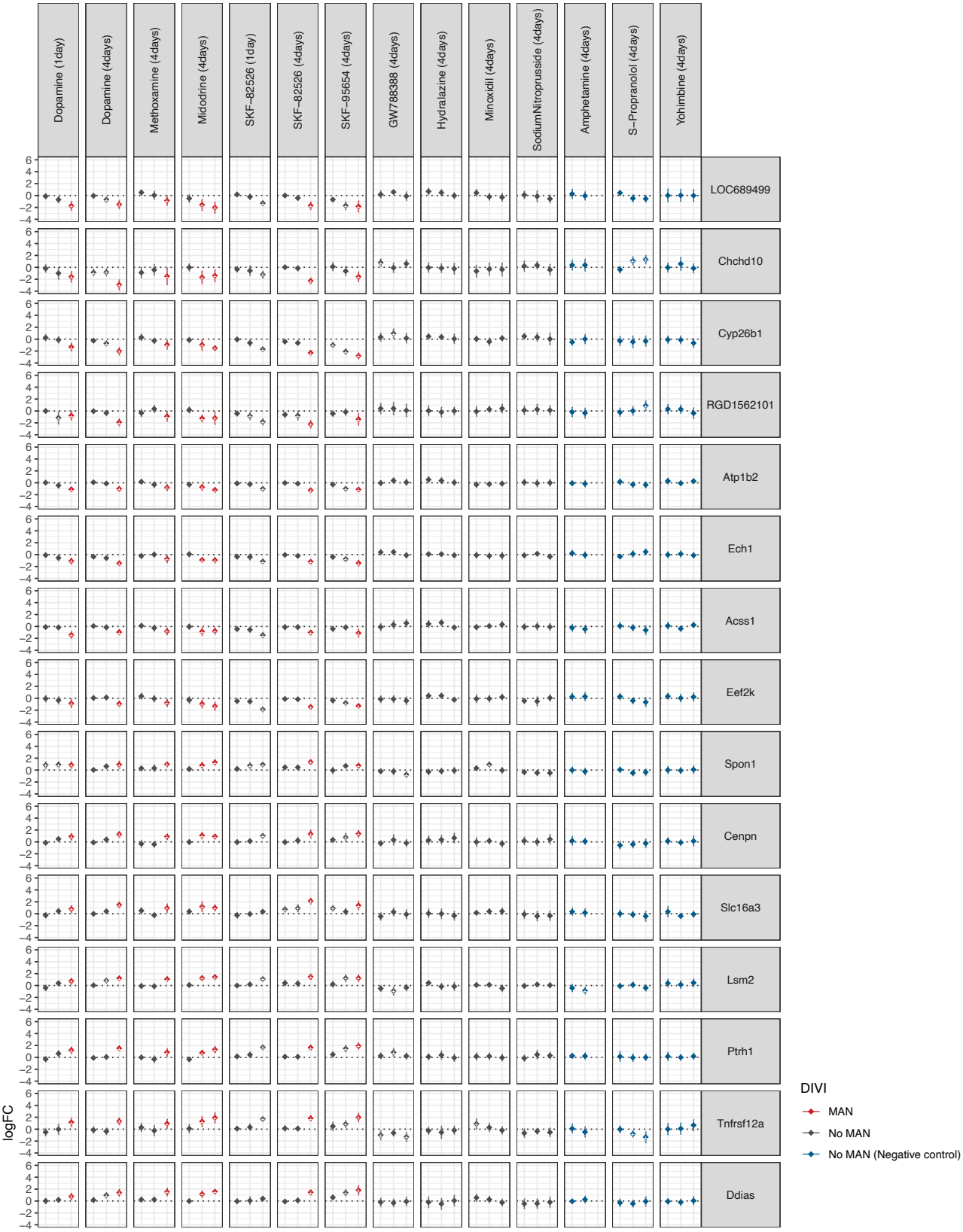


Figure A.2: Expression distribution across experiments for markers identified in the endothelium.


Dopamine (1day)
Dopamine (4days)
Methoxamine (4days)
Midodrine (4days)
SKF−82526 (1day)
SKF−82526 (4days)
SKF−95654 (4days)
GW788388 (4days)
Hydralazine (4days)
Minoxidil (4days)
SodiumNitroprusside (4days)
Amphetamine (4days)
S−Propranolol (4days)
Yohimbine (4days)
LOC689499
Chchd10
Cyp26b1
RGD1562101
Atp1b2
Ech1
Acss1
Eef2k
Spon1
Cenpn
Slc16a3
Lsm2
Ptrh1
Tnfrsf12a
Ddias
logFC
6
4
2
0
−2
−4
DIVI
MAN
No MAN
No MAN (Negative control)

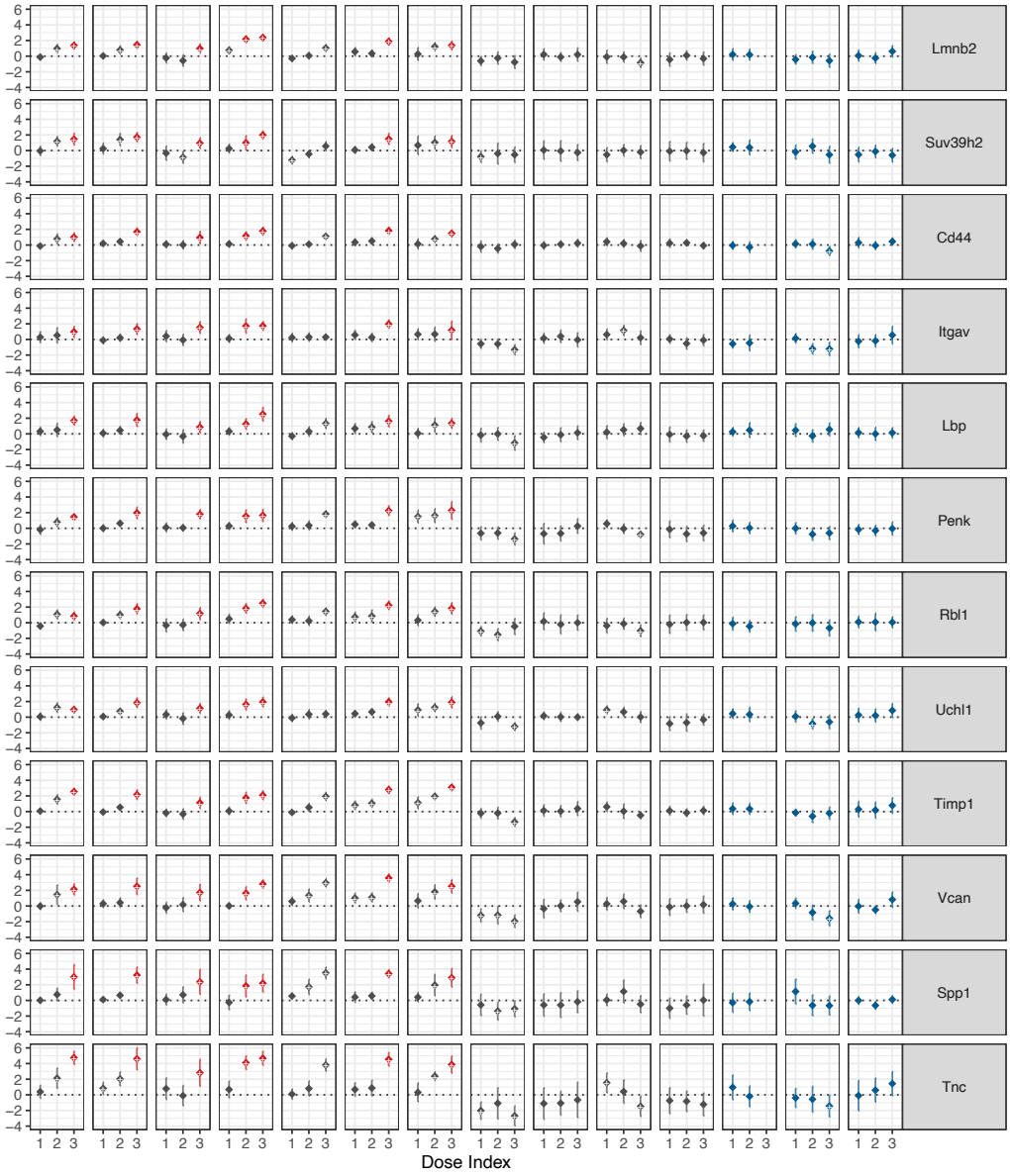


Figure A.3: Expression distribution across experiments for markers identified in the smooth muscle.

Appendix B

Table B.1: CEL files which were identified as outliers and removed.

Compound Name	Barcode
Amphotericin B	3017884018
Aspirin	3017659002
Caffeine	3017248021
Cephalothin	3017588011, 3017588012
Chlorpheniramine	3017296019
Cisplatin	3017472017
Coumarin	3017736025, 3017686002
Cyclosporine A	3017588014, 3017566019
Fluphenazine	3017050017
Haloperidol	3017001020, 3017001021
Ibuprofen	3017152020, 3017152021, 3017152030
Imipramine	3017153029
Iproniazid	3017140006
Lomustine	3017122006, 3017075012
Mefenamic Acid	3017175008,3017175010,3017182004
Metformin	3017044024
Methimazole	3017060009, 3017060014, 3017060011, 3017060012
Methyldopa	3017106011, 3017106012, 3017096030, 3017134026
Methylene Dianiline	3017875010, 3017875005
Methyltestosterone	3017014004,3017014018,3017014019,3017032018
Mexiletine	3017237028
Moxisylyte	3017108014, 3017075015, 3017075030, 3017014019, 3017032018
Naproxen	3017166017, 3017166022, 3017167015
Nifedipine	3017300003,3017300004, 3017300006
Nitrosodiethylamine	3017507001
Pemoline	3017062018,3017034026, 3017034029, 3017034030
Perhexiline	3017704003, 3017704010, 3017062025, 3017034022, 3017034030, 3017062018
Quinidine	3017151029, 3017151030, 3017152019
Rifampicin	3017666014
Rotenone	3017860029
Tacrine	3017089004, 3017122005
Tamoxifen	3017122002
Tetracycline	3017070015, 3017110008, 3017069028, 3017110015, 3017082005
Thioacetamide	3017663008
Thioridazine	3017049012, 3017049009
Tiopronin	3017261007
Tiopronin	3017238005, 3017238006, 3017239002, 3017243016, 3017243023
Tolbutamide	3017410002
Triamterene	3017417023, 3017417024, 3017352013
Trimethadione	3017477012, 3017503009
Vitamin A	3017072030
Wy-14643	3017708011, 3017709011

Table B.2: Time concordance metrics for Reactome pathway maps which map to known key events based on literature review.

Key event	Gene set description	Reactome ID	Direction	p-value	TPR	PPV
Bile acids	Recycling of bile acids and salts	R-HSA-159418	Down	5.655E-09	0.304 (17/56)	1 (17/17)
	Bile acid and bile salt metabolism	R-HSA-194068	Down	2.381E-06	0.25 (14/56)	0.933 (14/15)
	Synthesis of bile acids and bile salts	R-HSA-192105	Down	7.705E-04	0.179 (10/56)	0.833 (10/12)
Cell death	Programmed Cell Death	R-HSA-5357801	Up	4.870E-04	0.25 (14/56)	0.737 (14/19)
ER stress	Autophagy	R-HSA-9612973	Up	1.283E-02	0.161 (9/56)	0.692 (9/13)
	Unfolded Protein Response (UPR)	R-HSA-381119	Up	8.219E-02	0.143 (8/56)	0.571 (8/14)
	Unfolded Protein Response (UPR)	R-HSA-381119	Down	3.794E-01	0.054 (3/56)	0.5 (3/6)
JNK signaling	JNK (c-Jun kinases) phosphorylation and activation mediated by activated human TAK1	R-HSA-450321	Up	5.854E-03	0.179 (10/56)	0.714 (10/14)
	JNK (c-Jun kinases) phosphorylation and activation mediated by activated human TAK1	R-HSA-450321	Down	7.005E-01	0.054 (3/56)	0.333 (3/9)
LXR signaling	NR1H2 & NR1H3 regulate gene expression linked to lipogenesis	R-HSA-9029558	Down	1.483E-04	0.232 (13/56)	0.812 (13/16)
	NR1H2 & NR1H3 regulate gene expression linked to gluconeogenesis	R-HSA-9632974	Down	2.924E-04	0.196 (11/56)	0.846 (11/13)
	NR1H2 and NR1H3-mediated signaling	R-HSA-9024446	Down	9.618E-03	0.107 (6/56)	0.857 (6/7)
	NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux	R-HSA-9029569	Down	9.618E-03	0.107 (6/56)	0.857 (6/7)
	NR1H2 & NR1H3 regulate gene expression to control bile acid homeostasis	R-HSA-9623433	Down	1.273E-02	0.143 (8/56)	0.727 (8/11)
	NR1H2 & NR1H3 regulate gene expression to limit cholesterol uptake	R-HSA-9031525	Down	2.737E-02	0.107 (6/56)	0.75 (6/8)
	NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux	R-HSA-9029569	Up	1.307E-01	0.036 (2/56)	1 (2/2)
	NR1H2 & NR1H3 regulate gene expression linked to triglyceride lipolysis in adipose	R-HSA-9031528	Down	4.617E-01	0.036 (2/56)	0.5 (2/4)
	NR1H2 & NR1H3 regulate gene expression to limit cholesterol uptake	R-HSA-9031525	Up	5.966E-01	0.018 (1/56)	0.5 (1/2)
	NR1H2 & NR1H3 regulate gene expression linked to gluconeogenesis	R-HSA-9632974	Up	6.085E-01	0.054 (3/56)	0.375 (3/8)
	NR1H2 & NR1H3 regulate gene expression to control bile acid homeostasis	R-HSA-9623433	Up	7.108E-01	0.036 (2/56)	0.333 (2/6)

Mitochondrial beta oxidation	Mitochondrial Fatty Acid Beta-Oxidation	R-HSA-77289	Up	6.942E-05	0.268 (15/56)	0.789 (15/19)
	mitochondrial fatty acid beta-oxidation of unsaturated fatty acids	R-HSA-77288	Up	1.747E-04	0.357 (20/56)	0.667 (20/30)
	mitochondrial fatty acid beta-oxidation of saturated fatty acids	R-HSA-77286	Up	4.496E-04	0.232 (13/56)	0.765 (13/17)
	Mitochondrial Fatty Acid Beta-Oxidation	R-HSA-77289	Down	1.889E-01	0.089 (5/56)	0.556 (5/9)
	mitochondrial fatty acid beta-oxidation of saturated fatty acids	R-HSA-77286	Down	2.041E-01	0.125 (7/56)	0.5 (7/14)
	mitochondrial fatty acid beta-oxidation of unsaturated fatty acids	R-HSA-77288	Down	7.005E-01	0.054 (3/56)	0.333 (3/9)
Mitophagy	Mitophagy	R-HSA-5205647	Up	4.839E-06	0.411 (23/56)	0.719 (23/32)
	PINK1-PRKN Mediated Mitophagy	R-HSA-5205685	Up	5.874E-05	0.411 (23/56)	0.657 (23/35)
	Receptor Mediated Mitophagy	R-HSA-8934903	Up	5.609E-03	0.214 (12/56)	0.667 (12/18)
	Receptor Mediated Mitophagy	R-HSA-8934903	Down	4.617E-01	0.036 (2/56)	0.5 (2/4)
	Mitophagy	R-HSA-5205647	Down	8.995E-01	0.018 (1/56)	0.2 (1/5)
	PINK1-PRKN Mediated Mitophagy	R-HSA-5205685	Down	9.909E-01	0.018 (1/56)	0.1 (1/10)
Oxidative stress	Oxidative Stress Induced Senescence	R-HSA-2559580	Up	2.182E-01	0.071 (4/56)	0.571 (4/7)
	Oxidative Stress Induced Senescence	R-HSA-2559580	Down	8.397E-01	0.018 (1/56)	0.25 (1/4)
Peroxisome	Beta-oxidation of very long chain fatty acids	R-HSA-390247	Up	3.423E-05	0.339 (19/56)	0.731 (19/26)
	Peroxisomal lipid metabolism	R-HSA-390918	Up	1.483E-04	0.232 (13/56)	0.812 (13/16)
	Peroxisomal protein import	R-HSA-9033241	Up	1.483E-04	0.232 (13/56)	0.812 (13/16)
	Peroxisomal protein import	R-HSA-9033241	Down	2.790E-02	0.125 (7/56)	0.7 (7/10)
	Beta-oxidation of very long chain fatty acids	R-HSA-390247	Down	3.560E-02	0.196 (11/56)	0.579 (11/19)
	Peroxisomal lipid metabolism	R-HSA-390918	Down	4.487E-02	0.161 (9/56)	0.6 (9/15)
TNF signaling	TNF signaling	R-HSA-75893	Up	4.908E-03	0.143 (8/56)	0.8 (8/10)
	TNF signaling	R-HSA-75893	Down	2.993E-01	0.036 (2/56)	0.667 (2/3)

Table B.3: Time concordance metrics for top 10 ranking pathway events by True Positive Rate (TPR), significance and median max. |logFC|.

Event	Direction	p-value	TPR	PPV	logFC
Dectin-1 mediated noncanonical NF-kB signaling	Up	8.61E-05	0.446 (25/56)	0.625 (25/40)	0.79
Recycling of eIF2:GDP	Up	9.53E-06	0.446 (25/56)	0.676 (25/37)	0.94
RNA Polymerase I Promoter Escape	Up	1.67E-06	0.446 (25/56)	0.714 (25/35)	0.61
RNA Polymerase I Transcription Termination	Up	4.12E-06	0.446 (25/56)	0.694 (25/36)	0.67
rRNA modification in the nucleus and cytosol	Up	8.61E-05	0.446 (25/56)	0.625 (25/40)	0.86
SCF(Skp2)-mediated degradation of p27/p21	Up	8.61E-05	0.446 (25/56)	0.625 (25/40)	0.78
tRNA modification in the nucleus and cytosol	Up	3.65E-06	0.464 (26/56)	0.684 (26/38)	0.6
Folding of actin by CCT/TriC	Up	1.48E-05	0.482 (27/56)	0.643 (27/42)	0.96
Nonsense-Mediated Decay (NMD)	Up	6.98E-06	0.482 (27/56)	0.659 (27/41)	0.73
Response of EIF2AK4 (GCN2) to amino acid deficiency	Up	6.98E-06	0.482 (27/56)	0.659 (27/41)	0.74
rRNA processing	Up	2.99E-05	0.482 (27/56)	0.628 (27/43)	0.75
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	Up	4.69E-05	0.5 (28/56)	0.609 (28/46)	0.8
Cytosolic tRNA aminoacylation	Up	4.11E-07	0.518 (29/56)	0.69 (29/42)	0.74
Mitotic G1 phase and G1/S transition	Up	1.79E-06	0.393 (22/56)	0.759 (22/29)	0.5
Defects in biotin (Btn) metabolism	Down	1.67E-06	0.375 (21/56)	0.778 (21/27)	-0.73
HS-GAG biosynthesis	Down	1.58E-06	0.286 (16/56)	0.889 (16/18)	-0.56
DNA Replication	Up	1.46E-06	0.357 (20/56)	0.8 (20/25)	0.55
S Phase	Up	1.46E-06	0.357 (20/56)	0.8 (20/25)	0.52
Cell Cycle Checkpoints	Up	1.18E-06	0.339 (19/56)	0.826 (19/23)	0.5
Diseases associated with glycosaminoglycan metabolism	Down	7.44E-08	0.304 (17/56)	0.944 (17/18)	-0.45
Recycling of bile acids and salts	Down	5.66E-09	0.304 (17/56)	1 (17/17)	-0.57
Tyrosine catabolism	Down	7.23E-06	0.286 (16/56)	0.842 (16/19)	-0.91
Beta-oxidation of very long chain fatty acids	Up	3.42E-05	0.339 (19/56)	0.731 (19/26)	0.92
Removal of aminoterminal propeptides from gamma-carboxylated proteins	Down	3.42E-05	0.339 (19/56)	0.731 (19/26)	-0.92
Alpha-oxidation of phytanate	Up	1.99E-04	0.268 (15/56)	0.75 (15/20)	0.96
Cholesterol biosynthesis	Down	9.83E-03	0.304 (17/56)	0.567 (17/30)	-0.99
Mitochondrial fatty acid beta-oxidation of saturated fatty acids	Up	4.50E-04	0.232 (13/56)	0.765 (13/17)	1.02
Beta oxidation of decanoyl-CoA to octanoyl-CoA-CoA	Up	1.09E-03	0.214 (12/56)	0.75 (12/16)	1.08
Mitochondrial fatty acid beta-oxidation of unsaturated fatty acids	Up	1.75E-04	0.357 (20/56)	0.667 (20/30)	1.11

Table B.4: Time concordance metrics for top 10 ranking transcription factor events by True Positive Rate (TPR), significance and median max. |logFC|.

Event	Direction	p-value	TPR	PPV	logFC
Etv4	Down	7.51E-04	0.259 (15/58)	0.714 (15/21)	-0.5
Zfp217	Down	1.10E-02	0.259 (15/58)	0.6 (15/25)	-0.58
Kdm5b	Down	5.73E-03	0.276 (16/58)	0.615 (16/26)	-0.62
Zbtb11	Up	3.13E-03	0.276 (16/58)	0.64 (16/25)	0.57
Nfe2l2	Up	1.42E-02	0.293 (17/58)	0.567 (17/30)	0.79
Atf4	Up	1.41E-03	0.31 (18/58)	0.643 (18/28)	0.84
Srebf1	Down	1.41E-03	0.31 (18/58)	0.643 (18/28)	-0.74
Srebf2	Down	1.93E-02	0.31 (18/58)	0.545 (18/33)	-0.9
E2f2	Up	1.44E-04	0.328 (19/58)	0.704 (19/27)	0.69
Nr1h3	Down	6.84E-03	0.328 (19/58)	0.576 (19/33)	-0.66
Ebf1	Down	7.15E-04	0.241 (14/58)	0.737 (14/19)	-0.43
Foxl2	Up	6.62E-04	0.155 (9/58)	0.9 (9/10)	0.59
Hnf4a	Down	6.62E-04	0.155 (9/58)	0.9 (9/10)	-0.49
Mafb	Down	6.62E-04	0.155 (9/58)	0.9 (9/10)	-0.48
Tead1	Down	4.03E-04	0.19 (11/58)	0.846 (11/13)	-0.47
Sox13	Down	2.17E-04	0.224 (13/58)	0.812 (13/16)	-0.5
Hoxb13	Down	8.86E-05	0.19 (11/58)	0.917 (11/12)	-0.45
Tfap4	Down	5.77E-05	0.224 (13/58)	0.867 (13/15)	-0.53
Pou2f2	Down	4.03E-02	0.207 (12/58)	0.571 (12/21)	-0.65
Tfap2c	Down	3.13E-02	0.155 (9/58)	0.643 (9/14)	-0.7
Irf9	Up	4.03E-02	0.207 (12/58)	0.571 (12/21)	0.74
Klf4	Up	1.31E-03	0.19 (11/58)	0.786 (11/14)	0.79

Table B.5: TF-TF relations supported by known relations and time concordance.

For TF events which are significantly enriched before or at adverse histopathology, known interactions supported by time concordance are shown. With respect to the interaction, the absolute and relative frequency are shown for how often the source TF was observed "before" or "before or at" downstream TF activity. Additionally, the source of the interactions provided in Omnipath are shown for protein-protein interactions and the DoRothEA confidence level for TF-target gene interactions.

Class	Preceding event	Later event	TPR (Before or at)	TPR (Before)	Sources
PPI	Mef2c(Down)	Myod1(Down)	0.667 (4/6)	0.333 (2/6)	BioGRID;Lit-BM-17;SIGNOR;Wang
	Nr1h2(Down)	Ppara(Down)	0.3 (3/10)	0.1 (1/10)	SignaLink3
	Pax6(Down)	Maf(Down)	0.333 (3/9)	0.111 (1/9)	SPIKE
	Ppara(Down)	Nr1h2(Down)	0.444 (4/9)	0.222 (2/9)	SignaLink3
Regulon	Cebpa(Down)	Hnf4a(Down)	0.333 (3/9)	0 (0/9)	A
	Elf3(Down)	Meis1(Down)	0.364 (4/11)	0 (0/11)	C
	Hnf1a(Down)	Hnf4a(Down)	0.667 (6/9)	0 (0/9)	A
	Hnf4a(Down)	Cebpa(Down)	0.75 (3/4)	0 (0/4)	A
	Nfe2l1(Down)	Tead1(Down)	0.727 (8/11)	0 (0/11)	C
	Nr1h2(Down)	Srebf1(Down)	0.222 (4/18)	0 (0/18)	C
	Nr1h3(Down)	Srebf1(Down)	0.5 (9/18)	0.278 (5/18)	C
	Pbx2(Down)	Meis1(Down)	0.636 (7/11)	0 (0/11)	C
	Pbx3(Down)	Meis2(Down)	0.636 (7/11)	0 (0/11)	C
	Pdx1(Down)	Hnf4a(Down)	0.667 (6/9)	0.444 (4/9)	C
	Prdm1(Down)	Tead1(Down)	0.364 (4/11)	0 (0/11)	C
	Rara(Down)	Hnf4a(Down)	0.222 (2/9)	0 (0/9)	A
	Rela(Up)	Nfkb1(Up)	1 (4/4)	0 (0/4)	A
	Sox11(Down)	Tead1(Down)	0.273 (3/11)	0 (0/11)	C
	Tal1(Down)	Nfkb1(Up)	0.25 (1/4)	0 (0/4)	A
	Tcf12(Down)	Tead1(Down)	0.818 (9/11)	0 (0/11)	C
	Tcf4(Down)	Tead1(Down)	0.545 (6/11)	0 (0/11)	C
	Zfp384(Down)	Meis2(Down)	0.636 (7/11)	0 (0/11)	C
	Zfx(Up)	Zfx(Up)	1 (10/10)	0 (0/10)	E

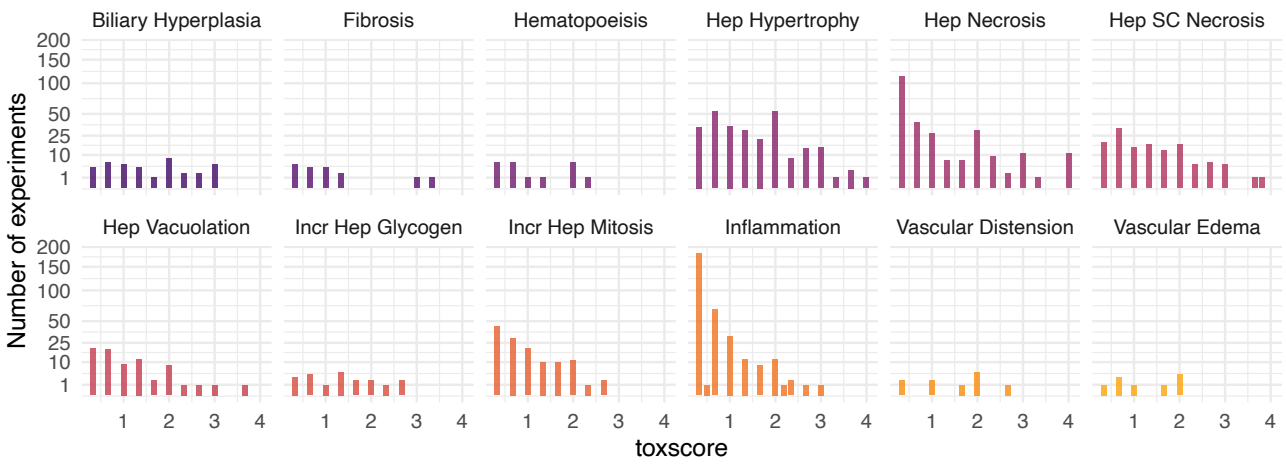


Figure B.1: Distribution of toxscores across histopathological findings.

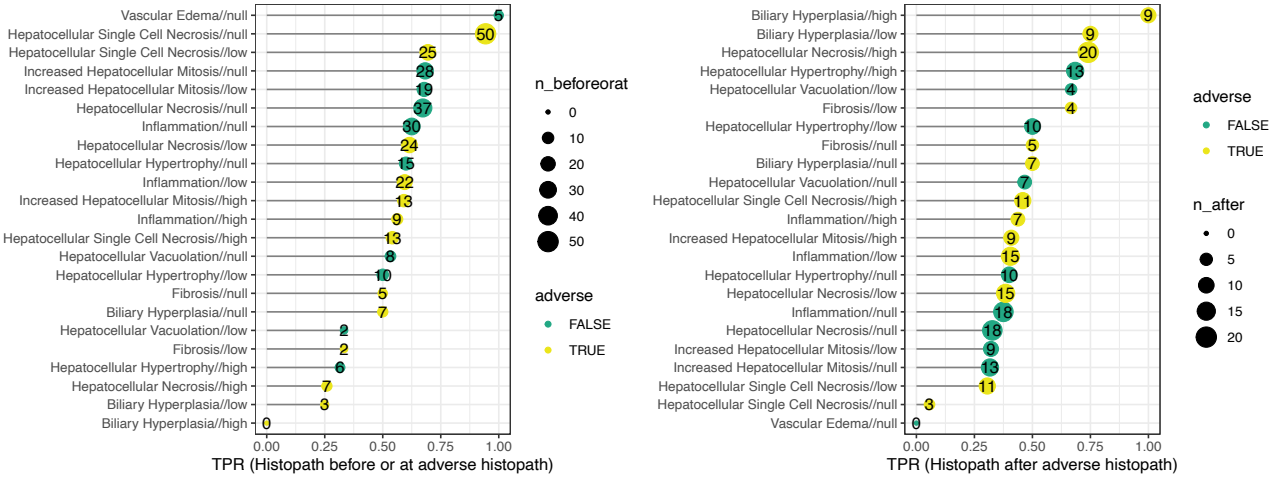


Figure B.2: Frequency of histopathological findings before and after first adverse histopathology.

For adverse and non-adverse histopathological findings, the frequency before or at first non-adverse histopathology is shown (left). For adverse findings, this indicates how frequently they were one of the first adverse histopathological findings given that they cannot occur before by definition. This identifies single-cell necrosis at any severity (“null”), as the most frequent finding, both in absolute and relative terms.

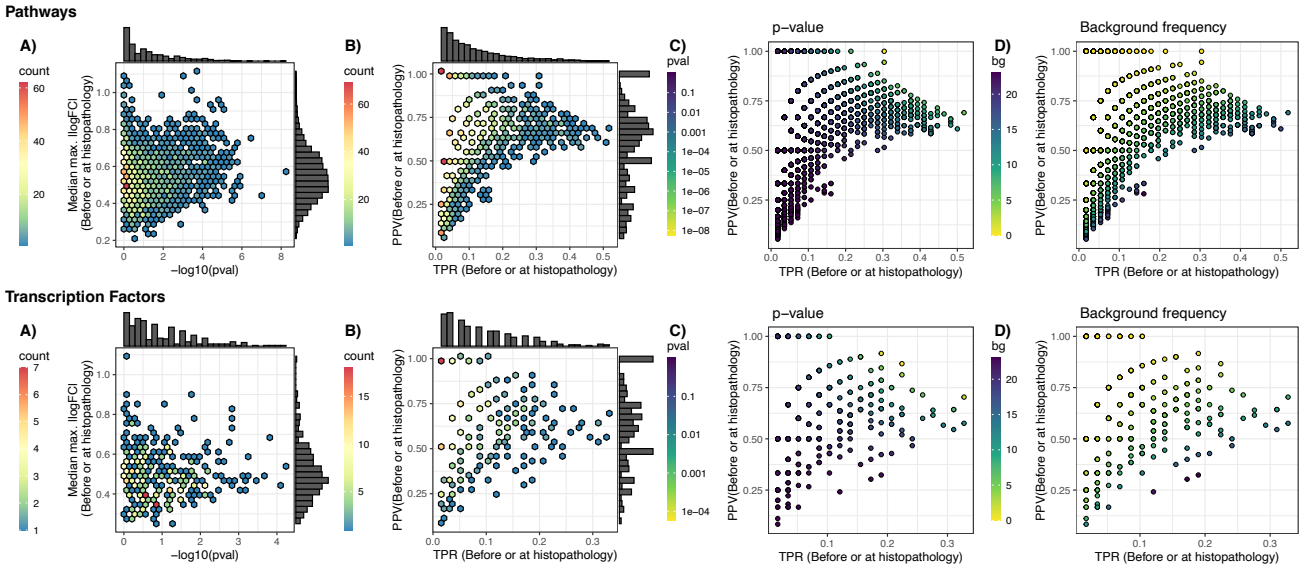


Figure B.3: Background distribution of temporal association metrics across pathway and Transcription Factor (TF) events.

The dependency between different metrics is shown. A) Frequency of events by median max. |logFC| before or at histopathology and enrichment p-value. B) Frequency of events by true positive rate (TPR) and positive predictive value (PPV) before or at adverse histopathology. C) Direct relation between TPR, PPV and enrichment p-value. D) Direct relation between TPR, PPV and frequency in background time-series.

File B.1: Time concordance metrics for all TFs, pathways as well as genes using both a minimal |logFC| of 0.5 and 1.

The file can be accessed via zenodo (DOI: 10.5281/zenodo.7017239).

Appendix C

Table C.1: Ten most significantly enriched transcription factor events before or at the time of liver fibrosis.

Event	TP	FP	OR	pval	Lift	Jaccard	TPR	PPV	logFC
Hinfp (Up)	3	1	52.2	9.94E-04	10.5	0.3	0.333	0.75	0.518
Srebf2 (Down)	6	19	10	1.95E-03	3.36	0.214	0.667	0.24	0.788
Hnf4a (Down)	3	2	26.6	2.39E-03	8.4	0.273	0.333	0.6	0.42
Batf (Up)	2	0	-	4.57E-03	14	0.222	0.222	1	0.459
Foxm1 (Up)	2	0	-	4.57E-03	14	0.222	0.222	1	0.513
Lef1 (Up)	2	0	-	4.57E-03	14	0.222	0.222	1	0.602
Smad1 (Up)	3	3	17.8	4.61E-03	7	0.25	0.333	0.5	0.518
Sp2 (Down)	3	3	17.8	4.61E-03	7	0.25	0.333	0.5	0.614
E2f2 (Up)	4	8	10.5	4.75E-03	4.67	0.235	0.444	0.333	0.681
Nr2f2 (Down)	4	8	10.5	4.75E-03	4.67	0.235	0.444	0.333	0.642

Appendix D

File D.1: Cell annotations provided by Choi et al. [133]

The anno_final_v3 cluster annotation provided directly by Choi et al. [133] was used to generate the cell type annotations used in Chapter 5. The file can be accessed via zenodo (DOI: 10.5281/zenodo.7017239).

Table D.1: Filtering steps implemented in the corresponding publications.

Dataset origin	Filter
Adams et al.	Number of UMI > 1000; Mitochondrial fraction < 20%
Habermann et al.	Number of UMI > 1000; Mitochondrial fraction < 25%
Choi et al.	7000> Number of genes > 500; Number of UMI > 2000
Strunz et al.	Number of genes > 200; Number of UMI <5000 (applied in the cell annotation step)

Table D.2: Differential expression of previously identified AT1 and intermediate progenitor (IP) markers in the IP→AT1 transition signature.

The most distinctive AT1 and IP markers were derived from previous work by Choi et al.[133] (Figure 1), mapped from rat to human gene names, and Strunz et al.[132] (Figure 3).

			Bleomycin			IPF		
class	**Gene**	**Source**	**logFC**	**p-value**	**FDR**	**logFC**	**p-value**	**FDR**
AT1	AGER	Choi	3.024	2.41E-16	1.48E-13	7.310	3.57E-29	2.00E-25
	CAV1	Choi	2.605	1.56E-15	7.70E-13	4.229	1.61E-17	9.86E-15
	CLIC5	Strunz	2.320	1.63E-11	4.01E-09	2.098	6.59E-07	3.09E-05
	HOPX	Choi, Strunz	3.177	9.00E-22	1.29E-18	1.841	1.20E-04	2.76E-03
	IGFBP2	Strunz	3.448	1.79E-10	3.50E-08	0.463	3.31E-01	5.05E-01
	PDPN	Choi, Strunz	2.621	2.57E-11	5.99E-09	3.325	6.78E-10	7.05E-08
	RTKN2	Strunz	5.237	1.42E-35	2.44E-31	5.550	1.54E-19	1.50E-16
	SPOCK2	Strunz	4.489	5.74E-27	4.95E-23	2.480	1.39E-07	7.89E-06
	VEGFA	Strunz	3.615	7.76E-26	3.34E-22	2.730	2.17E-09	1.95E-07
IP	AREG	Strunz	-2.693	4.13E-15	1.92E-12	-2.265	2.17E-04	4.45E-03
	CDKN1A	Choi, Strunz	-2.733	1.28E-12	4.07E-10	-1.716	9.10E-05	2.20E-03
	EDN1	Strunz	-3.820	1.71E-14	7.18E-12	-0.512	2.19E-01	4.24E-01
	HBEGF	Strunz	0.428	1.22E-01	3.86E-01	0.691	8.07E-02	2.81E-01
	ITGB6	Strunz	-0.016	9.61E-01	9.82E-01	-0.796	3.21E-02	1.73E-01
	KRT8	Choi, Strunz	-1.293	4.92E-05	1.54E-03	-0.548	1.41E-01	3.58E-01
	NDRG1	Choi	-5.884	8.36E-16	4.65E-13	1.600	3.25E-04	6.21E-03
	PLAUR	Strunz	-2.458	1.07E-07	9.39E-06	-2.331	9.23E-07	4.13E-05
	TNIP3	Choi, Strunz	-4.548	8.33E-24	1.79E-20	-0.508	4.55E-01	6.23E-01

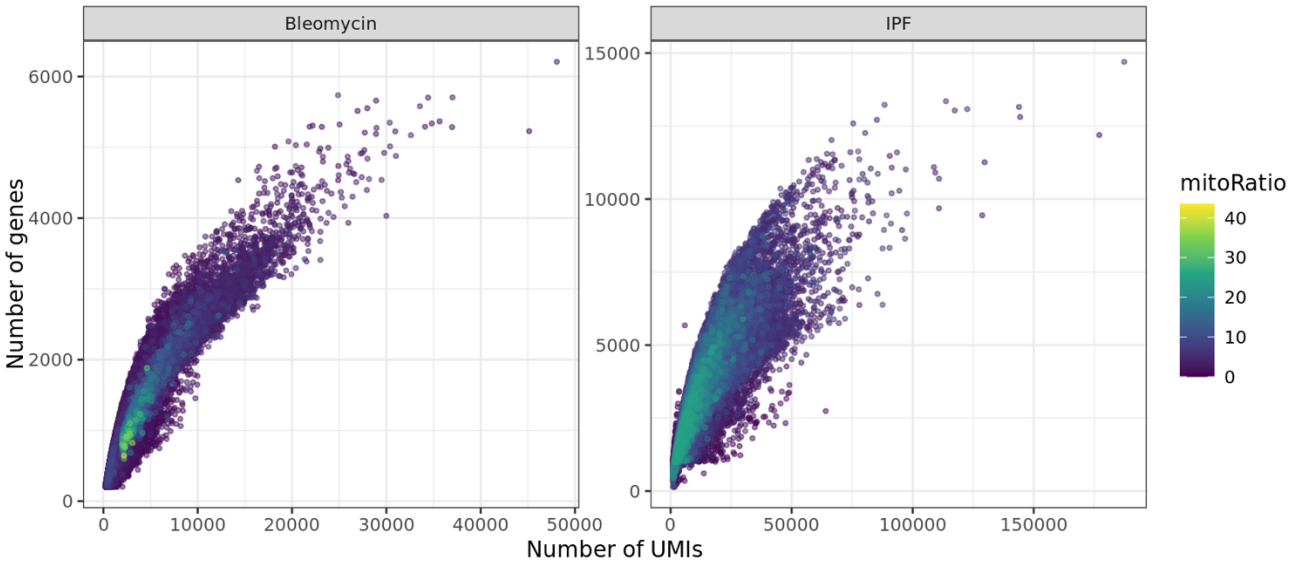


Figure D.1: Distribution of number of genes and Unique Molecular Identifiers (UMI) across all cells.

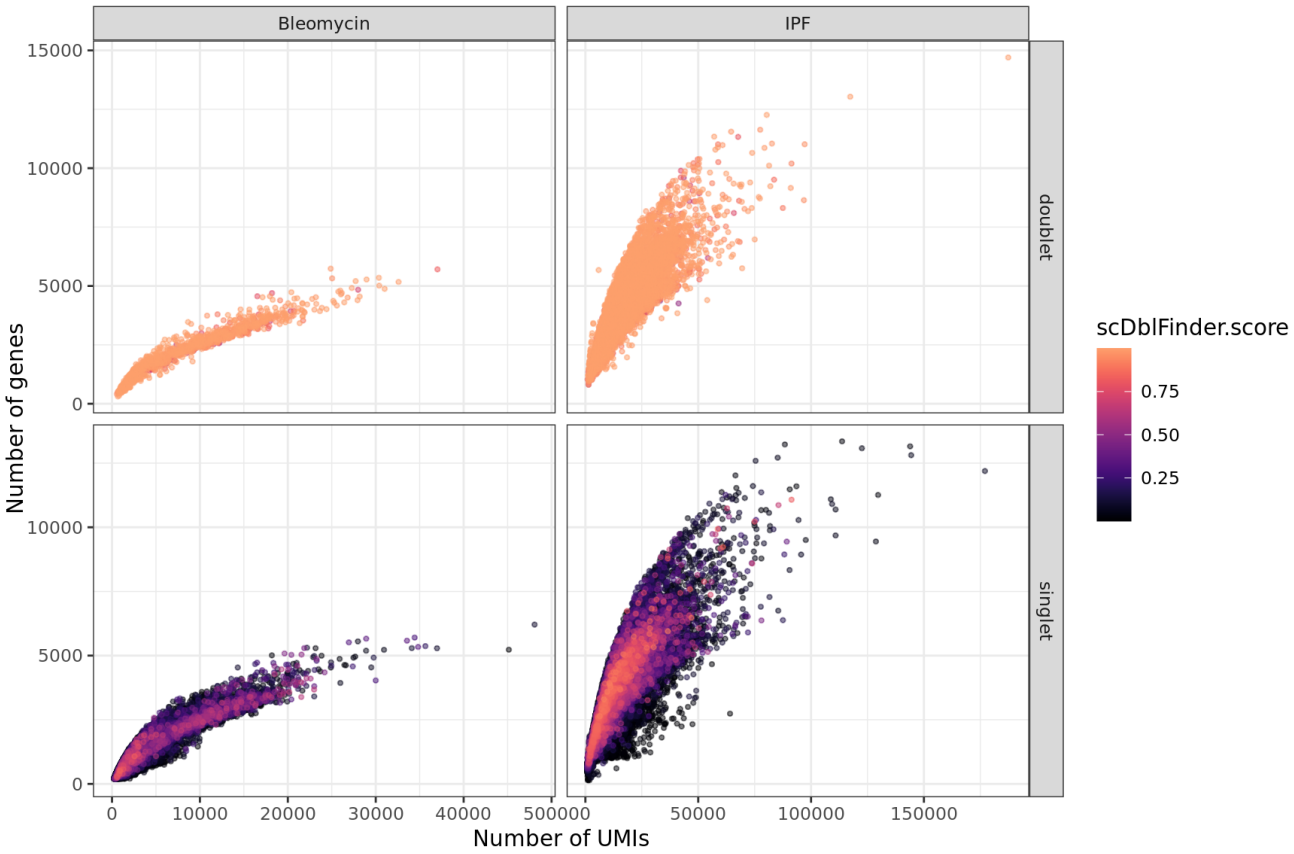


Figure D.2: Distribution of number of genes and Unique Molecular Identifiers (UMI) across all cells.

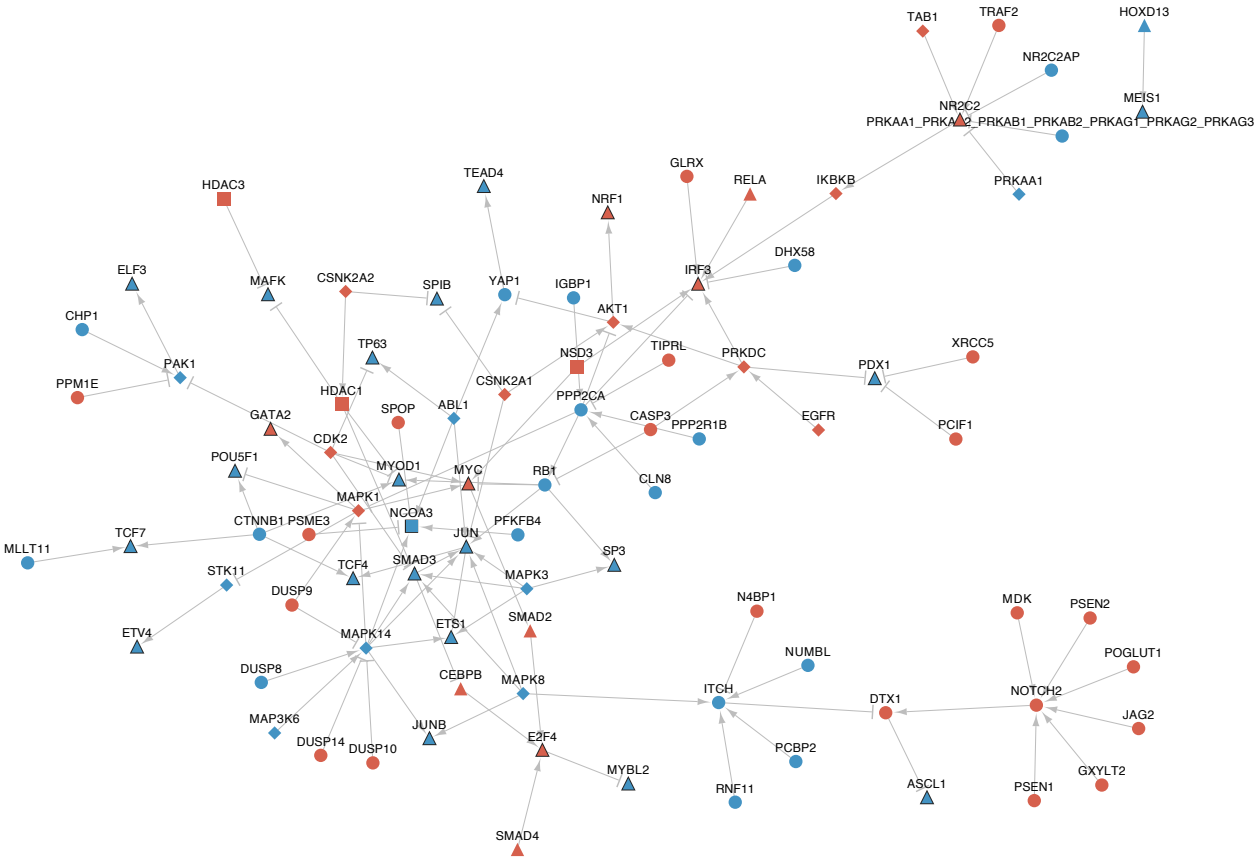


Figure D.3: CARNIVAL network for IPF signature.

Up- and down-regulation are indicated as red and blue, respectively. Transcription factors (▲), epigenetic regulators (■), and kinases (◆) are indicated through distinct node shapes given the specific interest in these targets in this study.

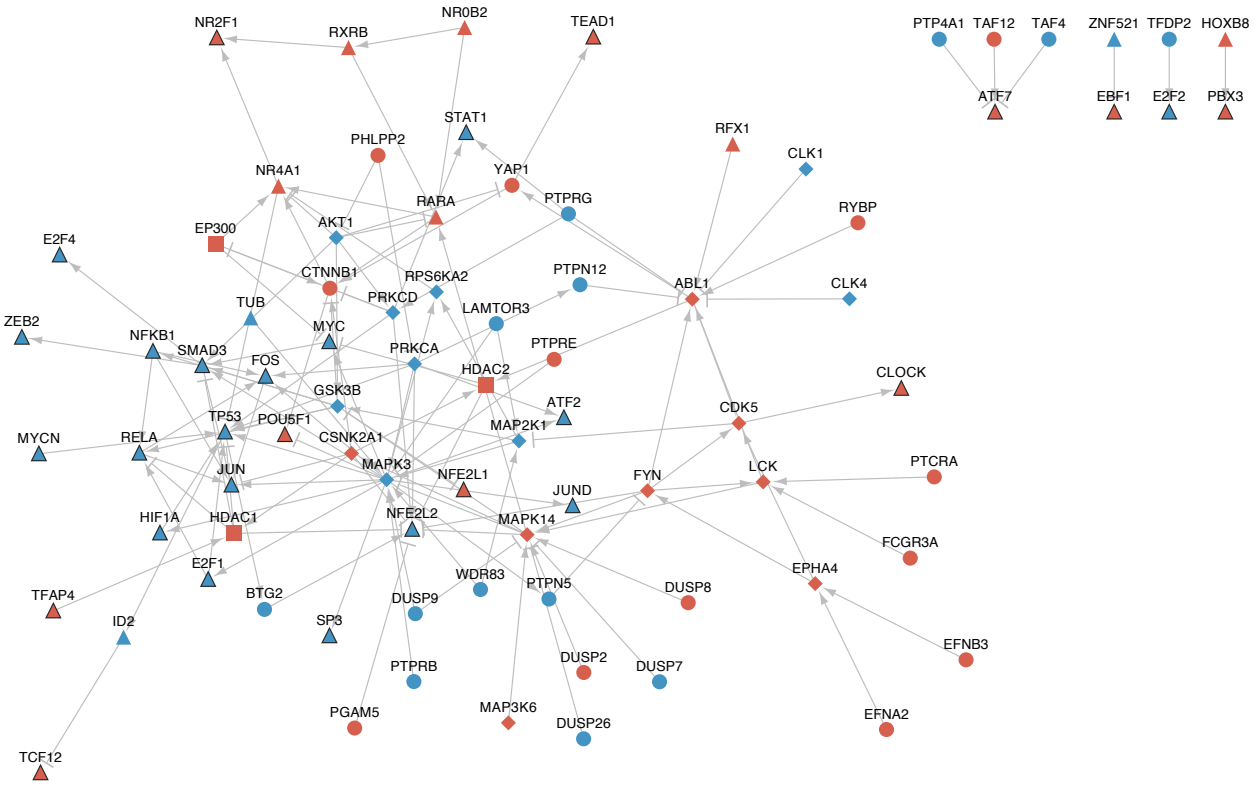


Figure D.4: CARNIVAL network for bleomycin injury signature.

Up- and down-regulation are indicated as red and blue, respectively. Transcription factors (▲), epigenetic regulators (■), and kinases (◆) are indicated through distinct node shapes given the specific interest in these targets in this study.

Figure D.5: Bioactivity data for correlated targets and matched compounds.

For all targets with positive correlations between bioactivity and signature matching for both signatures and significance for at least one (p-value < 0.05), the pCHEMBL values across matched compounds are shown and have been clustered using the Jaccard distance based on the presence or absence of a measured pCHEMBL value.

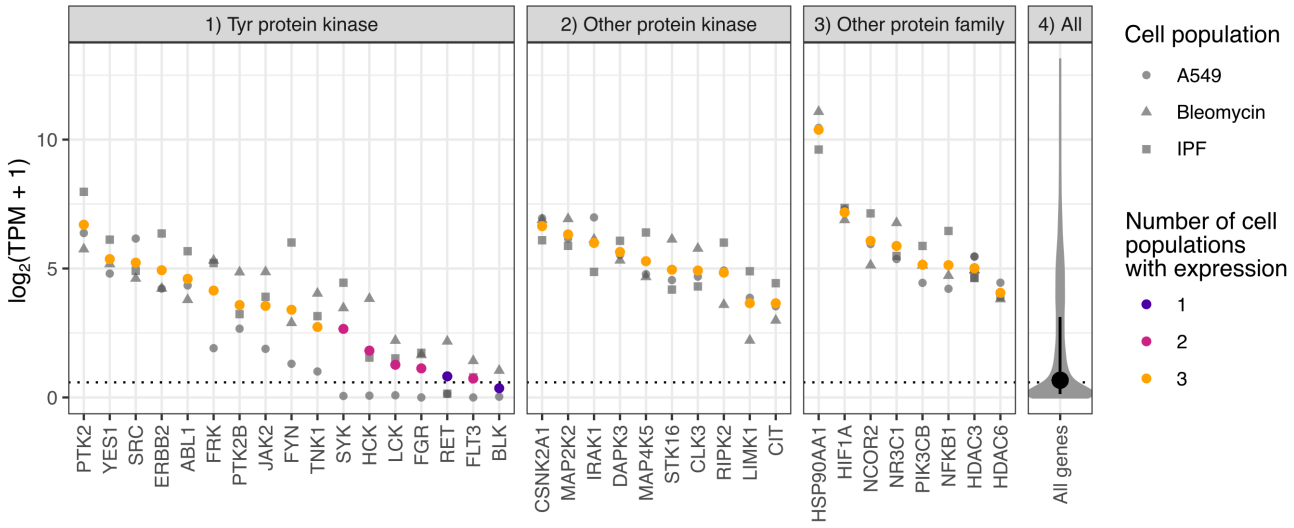


Figure D.6: Baseline expression of identified targets in intermediate progenitors and the A549 cell line.

The highest expression is found for *HSP90AA1*, while some of the tyrosine kinases are not expressed (Expression ≤ 0.5 TPM) across all of the cell populations indicating that target engagement may not be feasible *in vivo*.