Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
Christian H. Holland, M.Sc.
born in: Jülich, Germany

Oral examination: 03.12.2021

From gene expression to pathway and transcription factor activities to study chronic liver diseases

Referees: Prof. Dr. Ursula Klingmüller
Prof. Dr. Julio Saez-Rodriguez

Table of Contents

Abstract

High-throughput techniques such as microarrays and RNA-sequencing enable the relatively easy and inexpensive collection of bulk gene expression profiles from any biological condition. Recently, also the transcriptome of single cells can be efficiently captured via novel single-cell RNA-sequencing technologies. Functional analysis of bulk or single-cell gene expression data has been proven to be a powerful approach as they summarize the large and noisy gene expression space into a smaller number of biologically meaningful features such as pathway and transcription factor activities. In the first part of this thesis, I expanded the scope on the pathway analysis tool PROGENy and the transcription factor analysis tool DoRothEA through thorough benchmarking pipelines. First I transferred their regulatory knowledge from human to mouse to enable the functional characterization of gene expression profiles from mice. Moreover, I demonstrated the robustness and applicability of both tools on human single-cell RNA-sequencing data. In the second part of this thesis, I focussed on the analysis of gene expression profiles from mice and humans in the context of acute and chronic liver diseases. Finally, I identified and functionally characterized exclusively and commonly regulated genes of chronic and acute liver damage in mice and a set of genes that were consistently altered in a novel chronic mouse model and patients of chronic liver disease. Especially the latter demonstrates that, although major interspecies differences remain, there is a common and consistent transcriptomic response to chronic liver damage in mice and humans. This set of genes could be further investigated to study the pathophysiology of the liver in in-vitro and in-vivo studies.

Zusammenfassung

Hochdurchsatzmethoden, wie „Microarrays“ oder RNA-Sequenzierung erlauben die einfache und kostengünstige Generierung von „bulk“ Genexpressionsprofilen von beliebigen biologischen Zuständen. Ermöglicht durch neuartige Einzelzell RNA-Sequenzier-Technologien kann seit Kurzem auch das Transkriptom auf Einzelzellebene bestimmt werden. Funktionelle Analysen von Transkriptomdaten auf „bulk“ oder Einzelzelleben haben sich als geeignete Methode etabliert, da sie den großen und stark verrauschten Raum von Genexpressionswerten in eine kleinere Anzahl von biologisch relevanten Größen zusammenfässt, wie z.B. die Aktivität von Signalwegen oder Transkriptionsfaktoren. Im ersten Teil meiner Dissertation erweiterte ich die Funktionalität des Signalweg-Analyse Werkzeugs PROGENy und des Transkriptionsfakor-Analyse Werkzeugs DoRothEA durch systematische und sorgfältige Benchmark Analysen. Zunächst habe ich die regulatorischen Informationen auf denen PROGENy und DoRothEA basieren von Mensch auf Maus übertragen, um die funktionelle Analyse von Maus Genexpressionsprofilen zu ermöglichen und zu gewährleisten. Zusätzlich habe die Robustheit und Anwendbarkeit beider Werkzeuge auf humane Einzelzell RNA-Sequenzierung Daten nachgewiesen. In zweiten Teil meiner Dissertation habe ich mich auf akute und chronische Lebererkrankungen fokussiert und in diesem Zusammenhang human und maus-basierende Genexpressionsprofile analysiert. Schlussendlich konnte ich Gengruppen identifizieren und funktionell charakterisieren die entweder exklusiv im akuten oder chronischen oder in beiden Krankheitsbildern reguliert werden. Zusätzlich ergab meine Analyse eine Gruppe von Genen die konsistent in einem neuartigen chronischen Mausmodell und Patienten die an chronischen Lebererkrankung leiden reguliert sind. Insbesondere die zuletzt genannte Gengruppe zeigt auf, dass obwohl zwischen Mensch und Maus auf allen Ebenen große Unterschiede vorliegen, doch eine gemeinsame und konsistente Genexpressionssignatur als Antwort auf chronische Lebererkrankungen identifizieren werden kann. Diese Gene könnten in Zukunft mittels in-vivo und in-vitro Studien genauer untersucht werden, mit dem Ziel neue Erkenntnisse bezüglich der Pathophysiologie der Leber zu gewinnen.

Preface

This work is based on three published manuscripts, each forming a separate chapter. I am the sole lead author of all of them as I have performed the vast majority of all computational analyses and have written every manuscript myself.

The following manuscripts are included as part of this thesis

1. **Holland CH**, Szalai B, Saez-Rodriguez J. “Transfer of regulatory knowledge from human to mouse for functional genomics analysis.” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms.* (2020). DOI: 10.1016/j.bbagrm.2019.194431.

2. **Holland CH**, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, Joughin BA, Stegle O, Lauffenburger DA, Heyn H, Szalai B, Saez-Rodriguez, J. “Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data.” *Genome Biology.* (2020). DOI: 10.1186/s13059-020-1949-z.

3. **Holland CH**, Ramirez Flores RO, Myllys M, Hassan R, Edlund K, Hofmann U, Marchan R, Cadenas C, Reinders J, Hoehme S, Seddek A, Dooley S, Keitel V, Godoy P, Begher-Tibbe B, Trautwein C, Rupp C, Mueller S, Longerich T, Hengstler JG$^{\#}$, Saez-Rodriguez J$^{\#}$, Ghallab A$^{\#}$. “Transcriptomic cross-species analysis of chronic liver disease reveals consistent regulation between humans and mice.” *Hepatology Communications.* (2021). DOI: 10.1002/hep4.1797.

Other publications that I have contributed to but are not presented in this thesis

Next to these three main projects, I contributed during my candidature also to several other projects that are not directly related to the topic of this thesis, but still made up an essential part of my Ph.D.

1. Garcia-Alonso L, **Holland CH**, Ibrahim MM, Turei D, Saez-Rodriguez J. “Benchmark and integration of resources for the estimation of human transcription factor activities.” *Genome Research.* (2019). DOI: 10.1101/gr.240663.118.

2. Szalai B, Subramanian V, **Holland CH**, Alföldi R, Puskás LG, Saez-Rodriguez J. “Signatures of cell death and proliferation in perturbation transcriptomics data - from confounding factor to effective prediction.” *Nucleic Acids Research.* (2019). DOI: 10.1093/nar/gkz805.

3. Ghallab A, Myllys M, **Holland CH**, Zaza A, Murad W, Hassan R, Ahmed YA, Abbas T, Abdelrahim EA, Schneider KM, Matz-Soja M, Reinders J, Gebhardt R, Berres ML, Hatting M, Drasdo D, Saez-Rodriguez J, Trautwein C, Hengstler JG. “Influence of Liver Fibrosis on Lobular Zonation.” *Cells.* (2019). DOI: 10.3390/cells8121556.

4. Tajti F*, Kuppe C*, Antoranz A, Ibrahim MM, Kim H, Ceccarelli F, **Holland CH**, Olauson H, Floege J, Alexopoulos LG, Kramann R, Saez-Rodriguez J. “A

functional landscape of chronic kidney disease entities from public transcriptomic data." *Kidney International Reports.* (2020). DOI: 10.1016/j.ekir.2019.11.005.

5. Mohs A, Otto T, Schneider KM, Peltzer M, Boekschoten M, **Holland CH**, Hudert CA, Kalveram L, Wiegand S, Saez-Rodriguez J, Longerich T, Hengstler JG, Trautwein C. "Hepatocyte-specific NRF2 activation controls fibrogenesis and carcinogenesis in steatohepatitis." *Journal of Hepatology.* (2020). DOI: 10.1016/j.jhep.2020.09.037.

6. Ramirez Flores RO[*], Lanzer JD[*], **Holland CH**, Leuschner F, Most P, Schultz J-H, Levinson RT[#], Saez-Rodriguez J[#]. "Consensus Transcriptional Landscape of Human End-Stage Heart Failure." *Journal of the American Heart Association.* (2021). DOI: 10.1161/JAHA.120.019667.

7. Lopez-Dominguez R, Toro-Dominguez D, Martorell-Marugan J, Garcia-Moreno A, **Holland CH**, Saez-Rodriguez J, Goldman D, Petri M, Alarcón-Riquelme ME[#], Carmona-Sáez P[#].Transcription Factor Activity Inference in Systemic Lupus Erythematosus." *Life.* (2021). DOI: 10.3390/life11040299.

8. Robrahn L, Dupont A, Jumpertz S, Zhang K, **Holland CH**, Guillaume J, Rappold S, Cerovic V, Saez-Rodriguez J, Hornef MW, Cramer T. "Conditional deletion of HIF-1a provides new insight regarding the murine response to gastrointestinal infection with Salmonella Typhimurium." *bioRxiv.* (2021). DOI: 10.1101/2021.01.16.426940.

9. Schneider KM[*], Mohs A[*], Gui W, Galvez EJC, Candels LS, **Holland CH**, Elfers C, Kilic K, Schneider CV, Strnad P, Wirtz TH, Marschall HU, Latz E, Lelouvier B, Saez-Rodriguez J, de Vos W, Strowig T, Trebicka J, Trautwein C. "Imbalanced gut microbiota fuels HCC development by shaping the hepatic inflammatory microenvironment." *Under review at Nature Communications.* 2021.

10. Hernansaiz-Ballesteros R, **Holland CH**, Dugourd A, Saez-Rodriguez J. "FUNKI: Interactive functional footprint-based analysis of omics data." *arXiv.* (2021). ID: arXiv:2109.05796.

[*] *Shared first authorship* [#] *Shared senior authorship*

All of these fruitful and successful collaborations, which I have either led or contributed to, have made it possible for me to meet many fantastic scientists around the world, as illustrated in my collaboration network (Figure 1). This list of people is by no means exhaustive but just highlights the personal known collaborators with whom I have published joint scientific articles.

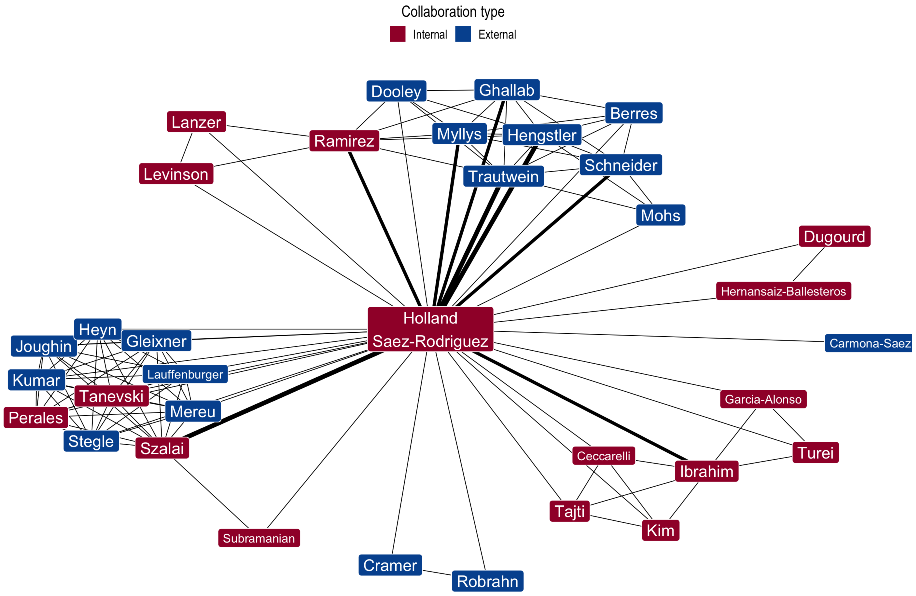


Figure 1: My Ph.D. collaboration network. Edge width corresponds to the number of joined publications. As my supervisor Julio Saez-Rodriguez was involved in all my collaborations we are represented in the network by a single node.

Acknowledgements

When I started my Ph.D. back in May 2017 in Aachen I would have never expected to finally submit my thesis four years later at Heidelberg University. Looking back, I am thankful for every experience and memory I was able to collect that allowed me to grow as an early career researcher but also as a human being. However, I would never have achieved anything presented in this thesis just on my own. Professionally and privately I was always fortunate to be accompanied by many patient, supportive, and friendly persons, I would like to thank here:

First of all, I wish to express my deepest gratitude to my doctoral supervisor Prof. Dr. Julio Saez-Rodriguez who gave me the chance to perform my studies in his lab. Throughout this entire time, I felt welcomed and respected. His liberal way of leading the lab promoted a great working atmosphere I have never experienced before to this extent. I truly enjoyed the freedom and always felt safe if things went badly or ended up in a dead-end. Moreover, through the open-door sessions, Julio was reachable on a daily basis which I definitely don't take for granted. In summary, Julio was the best supervisor I could have imagined and I deeply cherished the time I was allowed to spend in his lab.

Also, I am indebted to my faculty supervisor Prof. Dr. Ursula Klingmüller. Though it was not possible to share my entire Ph.D. process from the beginning with you due to the movement from Aachen to Heidelberg, I am nevertheless thankful for the lively discussions and feedback I received in the remaining time. Also, I would like to thank Prof. Dr. Robert Russell who kindly agreed to act as the chair of my thesis advisory committee. Lastly, I thank Prof. Dr. Karsten Niehaus who is my former supervisor of my bachelor's and master's studies at Bielefeld University for his willingness to be part of my committee.

Next, I am thankful for every past and current "saezlab" member who made this time very special and makes me now also kind of sad that this time has come to an end.

In particular, I wish to show my gratitude to the former postdoc Dr. Bence Szalai who especially during the first two years took care of me from my very first day. Just like Julio Bence has the talent to generate a psychologically safe and motivating environment. In the end, Bence's unwavering support and supervision were an existential decisive factor for my scientific and personal development. I am more than happy that our joyful collaboration was successfully crowned by several joint publications.

Also, I would like to thank other lab members who always felt much more like friends than colleagues: Alberto, Attila, Aurélien, Hyojin, Javier, Jovan, Luis, Nico, Mi, Olga, Rico, Rosa. All of you contributed in your own way to my Ph.D. Either with scientific discussions, valuable and highly appreciated contributions to my projects, or with fun activities outside of the office. Although I am no longer in contact with all of you, I am more than happy to have met you and will keep our time in the lab in memory.

I also wish to acknowledge the significant contributions of my many collaborators, without them the completion of my thesis would not have been possible. In particular, I would like to mention Prof. Dr. Jan Hengstler from Dortmund University who especially supported me in my liver disease-related project. With his outstanding knowledge of liver physiology, he complemented my bioinformatics analyses so we formed a successful and fruitful interdisciplinary collaboration.

Furthermore, I would like to thank the open-source software community who partially on a voluntary basis develop and maintain software. Without this community, my work would not have been possible.

Also I am especially grateful to my family who has supported me ever since.

Last but not least I wish to express my deepest gratitude to my girlfriend Laura: Laura, ich danke Dir von ganzem Herzen, dafür, dass du mir tagtäglich zur Seite standest und mir über all die Jahre den Rücken freigehalten hat, sodass ich mich, wenn es drauf ankam, voll und ganz auf meine Promotion fokussieren konnte. Auch hast du mich stets in all meinen, teils sehr weitreichenden, Entscheidungen unterstützt, wie z.B. der Entschluss, dass ich mit der Arbeitsgruppe nach Heidelberg umziehe und insgesamt für knapp 2 Jahre jede zweite Woche in Heidelberg verbracht habe, oder die Entscheidung für ein halbes Jahr in die Schweiz zu ziehen um Industrieerfahrung zu sammeln. Und selbst wenn ich dann zuhause war habe ich abends dann doch noch regelmäßig vor dem Laptop gesessen und “noch schnell eine Mail geschrieben” oder “nur kurz eine Idee ausprobiert.” Zum jetzigen Zeitpunkt steht es noch nicht fest wo unsere nächste gemeinsame Station sein wird, aber wo auch immer es ist, wir werden dort zusammen sein. Ich liebe dich.

Chapter 1

Introduction

1.1 The central dogma of molecular biology

All living organisms are based on a fundamental principle which is known as the central dogma of molecular biology (Crick, 1970). This dogma describes the information flow from a gene to a protein on the molecular level. Genes are encoded as the majority part of the deoxyribonucleic acid (DNA) and serve as a blueprint for transcripts. During the transcription, genes are copied from the DNA to a Ribonucleic acid (RNA) based transcript that is referred to as messenger RNA (mRNA). The transcription is followed by the process of translation. Thereby the mRNA is translated to a sequence of amino acids that are the building blocks of proteins. The amino acid sequence itself is linear but will form a complex three-dimensional structure.

Molecular Biology can be divided into several branches or disciplines each aiming to analyze a different layer of biological entities. They are based on technologies to quantitatively measure all involved biological molecules at each stage during the information flow from a gene to a protein. The branch genomics, for instance, aims to analyze the entire genome by deciphering the base sequence of the DNA via sequencing technologies. Transcriptomics is highly related to genomics but focuses on all RNA-based transcripts, the transcriptome. Proteomics detects proteins and quantifies their abundance and modifications via mass spectrometry (Altelaar, Munoz, & Heck, 2013). This list of omics technologies is by no means exhaustive as there exist many other branches, such as metabolomics (Patti, Yanes, & Siuzdak, 2012), lipidomics (Wenk, 2005), or epigenomics (Stricker, Köferle, & Beck, 2017) which each analyze their respective molecule class or layer of interest. The work described in this thesis focuses on transcriptomics.

1.2 Transcriptomics

Overview

Transcriptomics is the most widely studied field among the omics disciplines, which is most likely related to ever decreasing costs and to the good coverage of RNAs. The objective of transcriptomics is to quantify the entire transcriptome. Hence, this analysis is not limited to mRNA but comprises also other types of RNA such as ribosomal RNA or transfer RNA. The mRNA information alone is typically referred to as a gene expression profile. These profiles have been proven as a meaningful and interpretable data type as they can be considered as a blueprint of the status of the underlying cell or tissue. Over the years many technologies have been developed to measure the genome-wide expression profile. From the oldest to the most recent methods, all of them owe their existence to the advances in genome sequencing in the 90s and early noughties, particularly the sequencing of the human genome in 2001 (Lander et al., 2001).

Microarrays

One of the oldest but still reasonably popular method makes use of microarrays (Hoheisel, 2006). This technique is based on a chip with attached DNA fragments complementary to the DNA sequence of the genes of interest. Isolated RNA from the sample is reverse transcribed to complementary DNA (cDNA) and labeled with fluorescent molecules. Afterward, the cDNA library is transferred to the chip where cDNA molecules bind to their complement fragment that is attached to the chip. The cDNAs that do not bind are washed off. This setup shows clearly the caveat of microarrays as only the expression of genes for which there are attached complementary sequences on the chip can be quantified. Finally, a laser excites the fluorescence of the paired DNA sequences and considers their emission as a proxy for gene expression. Based on these principles the first samples were analyzed in 2003 with the arrays from Affymetrix. In 2015 the microarray technology reached its peak with over 15,000 samples analyzed and deposited on Gene Expression Omnibus (GEO) annually (Lachmann et al., 2018). Afterward, RNA-sequencing (RNA-seq) replaced microarrays as the most popular method for gene expression profiling.

RNA-sequencing

RNA-seq has a clear advantage over microarrays as theoretically nearly any RNA molecule in a sample can be quantified without prioritizing a priori which genes or transcripts are of interest (Zhong Wang, Gerstein, & Snyder, 2009). This implies that novel or different non-coding transcripts and also splice variants can be detected and quantified. Unlike microarrays, RNA-seq is not framed by background noise and signal saturation and thus has a much higher dynamic range to quantify transcripts (Wilhelm & Landry, 2009; Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014). Similar to the microarray technology a typical RNA-seq protocol starts with the generation of a cDNA library by RNA isolation and cDNA synthesis via reverse transcription. After amplification of the cDNA library via polymerase chain reaction (PCR), the cDNA molecules are fragmented into smaller so-called reads with a typical length of 50-100 base pairs. This step is crucial for the subsequent sequencing of the reads, as the standard sequencing machines cannot handle larger fragments, though this is changing recently with the emergence of long-read technologies such as Oxford Nanopore Technologies (Amarasinghe et al., 2020). After retrieving the base pair sequence for each read, the reads are mapped back to a representative genome of the respective species, a so-called reference genome. Since the number of reads can easily exceed 10 million per human sample (based on the sequencing depth), this step is highly computationally demanding. Hence, many computationally efficient alignment tools have been developed such as STAR (Dobin et al., 2013) or Kallisto (N. L. Bray, Pimentel, Melsted, & Pachter, 2016). Finally, the number of mapped reads per transcript is counted, which serves as a proxy for gene expression. Until 2019 more than 400,000 samples have been analyzed and deposited on GEO with different versions and protocols of the basic RNA-seq pipeline (Mahi, Najafabadi, Pilarczyk, Kouril, & Medvedovic, 2019). Despite the above-mentioned advantages of RNA-seq over microarrays, microarrays are still used and co-exist with RNA-seq.

While both methods made several important breakthroughs in biomedical research possible in the first place, they suffer from the same limitation. Their measured expression profile is the average of the expression profiles from many different cells or cell types. Therefore, their approach is referred to as bulk transcriptomics. While it is intuitive that highly distinct cell types such as parenchyma and immune cells have completely different transcription programs, it was also possible to show that even the gene expression of similar cell types is heterogeneous (Huang, Sherman, & Lempicki, 2009; Li & Clevers, 2010; Shalek et al., 2014). However, over the past decade, RNA-seq has evolved in such a way that nowadays the expression profile on a single-cell level can be captured.

Single-cell RNA-sequencing

First attempts with single-cell RNA-sequencing (scRNA-seq) were made in 2009 where the transcriptome of a single mouse blastomere was profiled (Tang et al., 2009). This technology promises to capture expression profiles at an unprecedented detail and was awarded as the technology of the year 2013 by Nature Methods ("Method of the year 2013." 2014). As the term scRNA-seq already indicates, RNA-sequencing is used across the majority of all technologies and protocols to profile the transcriptome. The different protocols vary how transcriptomic profiles are unambiguously mapped back to their origin cell, which is mostly achieved by cellular barcodes and in the construction of the cDNA library. Dependent on the experimental design either a plate or droplet-based approach would be more suitable (Baran-Gale, Chandra, & Kirschner, 2018). Inherently different protocols have different efficiency in capturing transcripts. This leads to a varying complexity of library composition and sensitivity to identify target genes. Recently, the human cell atlas consortium benchmarked

13 different protocols to identify the one with the greatest power of describing and distinguishing cell types and states (Mereu et al., 2020). Over the years the number of cells per study increased exponentially due to the rapid development of the underlying technology or protocol (Svensson, Vento-Tormo, & Teichmann, 2018). In 2017 it was possible to capture around 100,000 cells in a single run using in situ barcoding (Cao et al., 2017; Rosenberg et al., 2018). Nowadays, several million cells can be profiled as demonstrated in a recent study of human organ development where 4,000,000 single-cells have been sequenced (Cao et al., 2020).

Just like for bulk RNA-seq the transcripts must be reversely transcribed to cDNA. However, in a single cell, the number of available transcripts is very low in comparison to the number of transcripts in a bulk approach. Hence, some transcripts may be missed in the process of reverse transcription (Kharchenko, Silberstein, & Scadden, 2014). This can be due to several reasons and is still not fully understood. One essential factor is the gene expression level. Given that a gene is lowly expressed, also a low number of transcripts will be present in a cell which increases the chance that those transcripts are missed during reverse transcription (Kharchenko, Silberstein, & Scadden, 2014; Qiu, 2020). However, also the ratio of guanine-cytosine base pairs in the transcript or the enzyme named reverse transcriptase itself might influence whether certain transcripts are reversely transcribed. Accordingly, the missed genes are finally represented in the count matrix with zero counts even though they have been originally expressed in the cell and are thus referred to as drop-outs. Up to 90% of the final gene expression matrix can be zeros and it is not possible to distinguish whether a gene with a count of 0 is a drop-out or has truly not been expressed. Hence, scRNA-seq allows to profile the transcriptome of an enormous amount of cells but with limited gene coverage.

Selected flagship projects

Due to the affordable and continuously decreasing costs of transcriptomic studies, many large flagship projects have been established in the past two decades. The common core of these international and interdisciplinary efforts is to provide the scientific community with a comprehensive database of transcriptomic profiles of various human tissues or phenotypes measured at different resolutions. The following paragraphs briefly summarize selected flagship projects.

GTEx

GTEx stands for the Genotype-Tissue Expression Project and was launched in September 2010 by the National Institutes of Health (NIH) (Consortium, 2013). The main objective of GTEx is to provide tissue-specific gene expression profiles obtained from individual donors. In total GTEx provides these profiles for more than 30 distinct tissue types. Scientists worldwide query this database to improve the understanding of human diseases. A more concrete example of how this data is commonly used is the inference of tissue-specific gene regulatory networks via gene expression-based network reconstruction algorithms.

TCGA

TCGA stands for The Cancer Genome Atlas Program and was launched already in 2006 by the National Cancer Institute and the National Human Genome Research Institute (Network et al., 2013). Similar to the GTEx project TCGA focuses on individual tissue types, however, the objective is to study the transcriptomic profiles of their respective primary cancer (e.g. hepatocellular carcinoma or lung adenocarcinoma). Furthermore, TCGA also generates genomic, epigenomic, and proteomic data of primary cancers. This enormous data amount (2.5 petabytes) is interrogated to study the development and treatment of cancer either in specified or in multi-omic integration fashion.

CMAP

CMAP stands for connectivity map and was initially released in 2006 by the Broad Institute (Lamb et al., 2006). The objective of this project is to generate bulk gene expression signatures upon chemical or genetic perturbation across various human cell lines. Many of those perturbation experiments were also performed with different doses and perturbation times. In 2017 the next generation of CMAP was released which pushed the numbers of total perturbation signatures far beyond 1 million, perturbed by more than 20,000 perturbagenes including the majority of Food and Drug administration-approved drugs (Subramanian et al., 2017). This enormous effort was facilitated by the new high-throughput transcriptomic technology L1000 which lowered the sequencing costs drastically by only quantifying the expression of 978 landmark genes. The expression levels of the remaining genes are computationally inferred. The resulting large dataset enables scientists to systematically compare the signatures within CMAP or with custom gene signatures from e.g. a disease state. Identifying similar or dissimilar pairs and sets of signatures can help to identify novel drug targets or treatments for diseases such as cancer.

Human cell atlas

The human cell atlas is the most recent project and was launched in October 2016 (Regev et al., 2017). For a long time, there has been a wish to generate cellular maps of the human body. This idea is similar to GTEx efforts but with a much higher resolution. With the advent of fast-emerging single-cell RNA-seq technologies, this objective is now within reach. The human cell atlas project aims to profile the transcriptome of each cell type in the human body in unprecedented detail. From this dataset, we can learn how tissues are formed, to identify specific subpopulation cell types that drive the progression of a disease. This large-scale effort is still in its infancy, but the first single-cell datasets of various organs have been published which for sure will be a highly valuable resource for the entire scientific community.

1.3 Functional analyses

Overview

In general, there are many types of analyses that can be performed with transcriptomics data. Most commonly, the objective is to identify differences in gene expression levels between groups of samples via differential gene expression analysis. Bulk transcriptomic studies are often designed as perturbation studies to compare treated and untreated samples. In the clinical context, transcriptomic profiles of patients suffering from a certain disease are compared against the profiles of healthy individuals. In studies with animal models, the effect of a drug or a specific treatment can be tested by comparing treated and untreated animals. Since scRNA-seq is still in the early stages and thus expensive most studies do not follow a perturbation-based design, although this will increase in the future. Instead, individual cells of a tissue and organ are investigated. Still, comparisons can be made, e.g. by comparing the expression levels between different cell types of a tissue or organ.

Differential gene expression analysis typically leads to a large list with often more than 1000 significantly altered genes with associated p-value and effect size indicating the significance and magnitude of change in the expression level. Due to the vast number of potentially interesting genes, those lists can be hard to analyze and interpret looking at only a single gene at a time. Functional analysis of transcriptome data is a powerful downstream approach as it summarizes the large and noisy gene expression space into a smaller number of biological meaningful features. The concept behind this methodology is to analyze not the change in expression of individual genes but of groups of genes that are referred to as gene sets. This implies that each functional analysis tool couples a resource of gene sets with a statistical method that aims to analyze those sets.

Gene set types

Regarding gene sets, there is no limitation of how they can be constructed. Typically, gene set members are a collection of genes that share a common biological characteristic or function, such as the association to the same gene ontology term, position on the same chromosome, regulation by a common regulator, or encoding for members of a pathway. Especially the latter gene set type is widely used for classical pathway analysis. There exist many databases that provide those gene sets such as KEGG, REACTOME, PANTHER, or WikiPathways (Jassal et al., 2020; Kanehisa & Goto, 2000; Mi, Muruganujan, Ebert, Huang, & Thomas, 2019; Slenter et al., 2018). A common underlying assumption to summarize the expression of pathway members and then interpreted as pathway activity is that it is assumed that there is a positive correlation between gene expression, protein abundance, and protein activity. Based on those assumptions it follows that given that all genes of a pathway are highly expressed, those proteins are highly abundant and thus highly active. And if all individual proteins of a pathway are active, also the pathway itself is supposed to have high activity. This chain of assumptions violates several well-investigated biological principles. Indeed, several studies have shown that mRNA level can explain only ~40% of the variation in protein expression (Greenbaum, Colangelo, Williams, & Gerstein, 2003; Ideker et al., 2001; Sousa Abreu, Penalva, Marcotte, & Vogel, 2009; Washburn et al., 2003), though this correlation is higher for genes that are differentially expressed and thus under strong regulation (Koussounadis, Langdon, Um, Harrison, & Smith, 2015). Moreover, the activity of proteins is often rather determined by post-translational modifications than the abundance (Mann & Jensen, 2003). Regardless of those weaknesses and limitations, pathway analysis with gene sets of pathway members yields reasonable results and is widely used (Huang, Sherman, & Lempicki, 2009; Khatri, Sirota, & Butte, 2012; Krämer, Green, Pollard, & Tugendreich, 2014; Nguyen, Shafi, Nguyen, & Draghici, 2019; Tarca et al., 2009). A recent study indicates that this approach is effective because gene set members are regulated by a common regulator so that the pathway activity informs actually about the activity of the regulator (Szalai & Saez-Rodriguez, 2020). These common regulators are typically transcription factors, which serve as another class of biological meaning features, whose activity promises a valuable readout of the cellular state. Following the idea of classical pathway analysis, the activity of transcription factors could be inferred simply by their expression. Interestingly, this approach is rarely used, even though it violates the same principles. Instead, observing the expression of the transcriptional targets of a transcription factor yields a much more robust estimation of the transcription factor activity (Alvarez et al., 2016; Essaghir et al., 2010; Garcia-Alonso et al., 2018; Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019; Keenan et al., 2019; Kwon, Arenillas, Worsley Hunt, & Wasserman, 2012; Puente-Santamaria, Wasserman, & Del Peso, 2019; Roopra, 2020; Zhenjia Wang et al., 2018). Hence, the gene sets used to infer transcription factor activity is composed of downstream target genes, i.e. regulons. Those regulatory networks can be reconstructed in many ways, ranging from wet-lab techniques to pure in-silico generated networks and spanning multiple omics-technologies. In a recent study, networks derived from Chromatin Immunoprecipitation Sequencing (ChIP-seq) data, transcription factor binding sites, literature reviews, and gene expression data were integrated into a single consensus network referred to as DoRothEA (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019).

Observing the downstream effects of a biological process to gain functional and mechanistic insight into the upstream event is referred to as footprint analysis (Dugourd & Saez-Rodriguez, 2019). This concept is not exclusively limited to transcription factors. Intuitively, it can be easily transferred to estimate also kinase activity from phosphoproteomics data by exploiting the abundance of phosphorylated sites of kinase targets (Hernandez-Armenta, Ochoa, Gonçalves, Saez-Rodriguez, & Beltrao, 2017; Wiredja, Koyutürk, & Chance, 2017). However, this footprint concept can also be applied to biological processes that only have an indirect effect on e.g. gene expression, such as signaling pathways. This idea led to a novel way and perspective of predicting pathway activities from gene expression data. Instead of observing the expression of pathway members, the expression of the downstream affected genes is considered. The first large-scale tools that followed this principle are SPEED(2) and PROGENy (Parikh, Klinger, Xia, Marto, & Blüthgen, 2010; Rydenfelt, Klinger, Klünemann, & Blüthgen, 2020; Schubert et al., 2018). The limiting step of these methods is the number of pathways in the respective model. For each pathway separately, the downstream affected genes must be identified. The identification strategy from SPEED(2), as well as PROGENy, relies on the manual curation of pathway perturbation experiments with corresponding expression profiles. The footprint-based pathway analysis approach answers a different question than classical pathway analysis. The latter tries to explain the consequences of the measured expression pattern while footprint-based tools aim to identify the cause yielding the measured expression pattern (Szalai & Saez-Rodriguez, 2020).

Most gene sets are an unweighted collection of individual genes. However, it is also possible to assign weights to each gene set member, which opens up new avenues for how those gene sets could be analyzed. In terms of transcription factor analysis, the assigned weight could denote the mode of regulation, i.e. whether a transcription factor activates or suppresses the expression of its target gene. Similar to the footprint-based pathway analysis this weight could indicate the strength and direction of regulation upon pathway perturbation.

Different types of statistics to analyze gene sets

The available number of statistics to analyze gene sets together with transcriptomics data is comparable to the various types and sources of gene sets. The first generation of statistics tests whether gene set members are statistically over-represented in a list of differentially expressed genes, and is therefore referred to as over-representation analysis (ORA). Most commonly the test is based on the hypergeometric distribution known as the Fisher exact test. If gene set members are significantly over-represented in a list of differentially expressed genes it is assumed that the functional feature of the gene set is relevant for the underlying biological context. This strategy implies determining a cutoff classifying genes as differentially expressed. For instance, a gene can be considered differentially expressed if it passes a false discovery rate (FDR) $\leq$ 0.05 and an absolute log-fold change (logFC) ≥ 1. However, there is no objective legitimation for those exact values, so that any other combination could be used as well, and genes that just do not pass the chosen threshold won't be considered at all, even if they are just marginally below the thresholds. Consequently, the arbitrary selection of cutoffs directly impacts the results of ORA. Besides this limitation, ORA also treats the gene sets as an unweighted collection and thus equally, even though the degree and strength of regulation, depicted as significance and effect size, could be useful features to weight the individual gene set members.

The second generation of statistics tries to overcome those limitations and is referred to as functional class scoring (FCS) (Khatri, Sirota, & Butte, 2012). As opposed to ORA, where only the top differentially expressed genes are considered, FCS takes all genes (i.e. gene signature), irrespective of their strength of regulation, into account. Still, FCS builds on ORA as it acknowledges that if gene set members are strongly differentially expressed this has a significant functional effect. However, additionally, it is assumed that if gene set members are less strongly deregulated but in a coordinated manner this is still functionally relevant.

Gene set enrichment analysis (GSEA) is the most popular and widely used statistic out of the FCS generation. To detect whether gene sets are functionally relevant GSEA first ranks gene signatures derived from transcriptomic studies, based on a gene-level statistic, which can be any quantitative metric that is assigned per gene. Typically, log fold-changes, t-statistic, or even p-values serve as gene-level statistics. Subsequently, GSEA tests whether a gene set is significantly enriched at the top or the bottom of the list, indicating whether the functional feature of the gene set is increased or depleted in the given biological context. The original implementation is a rank-based approach, based on the Kolmogorov-Smirnov statistic. Next to GSEA and similar statistics, also general-purpose statistics as simple as z-score transformation, sum, or arithmetic mean could be applied to analyze gene sets operating on the chosen gene-level statistic. In the case of weighted gene sets also more complex approaches such as various types of linear models could be applied (Schubert et al., 2018; Trescher, Münchmeyer, & Leser, 2017).

So far those described statistics from the first and second generation operate on either a subset or entire gene signatures, which are typically the result of differential expression analysis of a case-control study. However, there are also methods that have been developed for the single-sample analysis such as ssGSEA, GSVA, PLAGE, or singscore (Barbie et al., 2009; Foroutan et al., 2018; Hänzelmann, Castelo, & Guinney, 2013; Lee, Chuang, Kim, Ideker, & Lee, 2008; Tomfohr, Lu, & Kepler, 2005). Those methods make gene set analysis also applicable to studies that do not follow the case-control design.

There have been attempts to combine a set of different statistics from the first and second generation to generate consensus functional analysis results (Väremo, Nielsen, & Nookaew, 2013).

Specifically for classical pathway analysis, an even third generation of statistics has been developed. As mentioned above classical pathway analysis is based on gene sets containing pathway members. However, both ORA and FCS ignore the functional relationship among pathway members. Especially the topology of a pathway has been neglected so far, even though this information is easily accessible in numerous databases. Henceforth, methods have been proposed that incorporate also pathway topology (Draghici et al., 2007; Hidalgo et al., 2017; Salviato, Djordjilović, Chiogna, & Romualdi, 2019; Tarca et al., 2009). Those methods assume that the position of a gene within a pathway is a meaningful feature, such as that upstream pathway members might have a larger influence on pathway activities than more downstream members, respective members without any downstream connection.

In summary, the suite of different approaches to functionally analyze transcriptome data can be applied to decipher key mechanisms of diseases and their progression. In my thesis, I focussed on liver-related diseases and disorders.

1.4 Chronic liver diseases

Structure of the liver

The liver is the largest solid organ in the human body comprising 2% of the body weight under healthy conditions. Among its primary functions is the metabolism of macromolecules such as fats, proteins, and carbohydrates to retain metabolic homeostasis. Accordingly, the liver also stores and redistributes nutrients. On a molecular level, the liver tissue is organized as hexagon-shaped hepatic lobules. Hepatocytes, which serve as the functional cells of a liver (“the liver cells”), constitute the largest part of those lobules and are circularly arranged around the lobule center. At each of the corners of the lobules, there is a distinctive structure consisting of branches of the portal vein, the hepatic artery, and the bile duct. Through the portal vein, hepatocytes are supplied with nutrients coming from the spleen, stomach, and intestines. This supply constitutes around 75% of the liver’s blood supply. The remaining 25% of the blood supply is delivered by the hepatic artery to serve hepatocytes with oxygen. The bile duct carries bile that is secreted by hepatocytes into the gallbladder (Boyer, 2013). The nutrient as well as the oxygen-rich blood flows to the center of the hepatic lobules and thereby distributes the nutrients and oxygens among the cells via the liver sinusoids. Finally, the nutrient and oxygen-poor blood reach the central vein from where it is transported to the hepatic vein that leads the blood back to the heart. Through the blood supply of the portal vein, the liver is continuously exposed to gut bacteria and associated endotoxins. Those particles are eliminated through phagocytosis by specialized macrophages, so-called Kupffer cells, which serve as another basic cell type of the liver. These Kupffer cells are part of the innate immune system and reside in the lumen of the sinusoids while being attached to the sinusoidal endothelial cells. Furthermore, the liver also contains hepatic stellate cells (HSC), which are liver-specific mesenchymal cells. They are located in the perisinusoidal space and store lipids. Under healthy conditions, they represent only 5-8% of all liver cells and are situated in a quiescent state (Blouin, Bolender, & Weibel, 1977).

Liver damage and repair

Like any other organ, the liver can take damage for various reasons. From a histological perspective, liver damage is reflected by necrotic and apoptotic hepatocytes. HSCs are pivotal for the wound healing response. Following liver damage, they get activated, proliferate, and start to synthesize extracellular matrix (ECM). In case of a minor or a single injury, ECM is deposited in and around the wound, which helps regenerate functional liver tissue by the proliferation of hepatocytes. However, if there is major damage ECM starts to accumulate which leads to scars on the liver. For repetitive damage, ECM continues to accumulate, and thus replaces functional liver tissue leading to the disruption of the tissue architecture. This scaring process is referred to as fibrosis but it is not exclusive to the liver. In fact, fibrosis can affect any organ in the body such as renal, pulmonary, or cardiac fibrosis (Henderson, Rieder, & Wynn, 2020).

It is estimated that fibrosis is responsible for 45% of all deaths in the industrialized world. If the underlying cause of the liver damage is not removed, over the years more and more functional tissue will be replaced by ECM. This process can take any time from 5 up to 50 years but ultimately leads to the loss of function (Pellicoro, Ramachandran, Iredale, & Fallowfield, 2014). This disease stage is referred to as cirrhosis and most patients suffering from it require liver transplantation. Otherwise, they will develop hepatocellular carcinoma (HCC), which is the third most common cause of cancer-related deaths worldwide and has an estimated incidence of more than 1,000,000 by 2025 (F. Bray et al., 2018; Llovet et al., 2021).

Etiologies of chronic liver diseases

Disorders that lead to repetitive liver damage are referred to as chronic liver diseases (CLDs) and can have manifold etiologies. In the past chronic liver injury was particularly induced by viral infections such as hepatitis C. In 1980 this virus was discovered and the first blood tests for the virus detected were established. These efforts were led by the scientists Harvey J. Alter, Charles M. Rice, and Michael Houghton who ultimately got awarded the medicine Nobel prize in 2020 for their research. Nowadays, there exist effective, yet expensive therapies for hepatitis. Therefore virus infections remain only a minor cause for CLD in the industrialized world, though, this is still a severe issue in developing countries.

Nevertheless, the number of chronic liver disease cases is increasing in the western world. This is partly due to the changing lifestyle with unlimited access to unhealthy food. Super nutrition leads finally to obesity, which goes along with several severe health risks. In terms of the liver, obesity leads to the massive accumulation of fat in the liver which is referred to as non-alcoholic fatty liver disease (NAFLD). Partially, NAFLD progresses to non-alcoholic steatohepatitis (NASH), which involves continuous damage of the liver tissue by inflammatory processes. Other etiologies are massive alcohol abuse, auto-immune disorders as well as metabolic diseases such as diabetes. If the underlying cause of CLD is removed even a cirrhotic liver has the capability to repair itself (Pellicoro, Ramachandran, Iredale, & Fallowfield, 2014).

1.5 Thesis overview and aims

The incidence of chronic liver diseases and hepatocellular carcinoma is continuously increasing. Therefore, scientists around the world are trying to decipher the molecular mechanisms to ultimately develop therapeutic options. Is it obvious that a single branch of biology or medicine cannot accomplish this goal alone. Instead, multiple disciplines must come together. During my Ph.D., I aimed to contribute to these efforts by analyzing transcriptomics data of liver diseases. Next to the classical analyses on gene level, the focus was in particular on the further development and application of the transcription factor and pathway analysis tools DoRothEA and PROGENy. On my journey I completed the following milestones:

1. Benchmarking the transcription factor and pathway analysis tools DoRothEA and PROGENy for their application in mice (Chapter 2).
2. Testing the robustness and applicability of the transcription factor and pathway analysis tools DoRothEA and PROGENy in single-cell RNA-sequencing data (Chapter 3).
3. Analysis and functional characterization of acute and chronic liver disease transcriptomic data in mice and humans (Chapter 4).

Chapter 2

Transfer of regulatory knowledge from human to mouse for functional genomics analysis

2.1 Preface

The text of the following chapter is largely taken from the publication "Transfer of regulatory knowledge from human to mouse for functional genomics analysis" (Christian H. Holland, Szalai, & Saez-Rodriguez, 2020) that was originally written by myself. The only changes were made to clarify my contribution to this project. Unless otherwise stated I performed all analyses myself. As the first author of this study the publishing house Elsevier grants me the right to include this work in my dissertation.

2.2 Background

The typical framework of functional genomics studies comprises the analysis of expression changes of groups of genes. These groups are referred to as gene sets and typically consist of genes sharing common functions (e.g. Gene Ontology analysis) or genes encoding for pathway members (Subramanian et al., 2005). The latter are used for classical pathway analysis studies, which assume that the transcript level is a proxy for protein abundance and thus the pathway activity. The framework of estimating transcription factor (TF) activity based on its gene expression follows the same principle (Figure 2.1A). However, mapping the transcript level to proteins neglects the effects of post-transcriptional and post-translational modifications, even though they are essential for the function of many proteins (Mann & Jensen, 2003).

To overcome this limitation, alternative approaches have been developed which are based on newly derived gene sets containing gene signatures obtained from genetic or chemical perturbations of pathways or TFs. These signatures are the footprint on gene expression of the corresponding pathway or TF (Figure 2.1A). Recent studies indicate that footprints outperform mapping gene sets (Cantini et al., 2018; Schubert et al., 2018). Since most of these footprints are generated for the application in humans, their usability in model organisms is uncertain. This question is of importance since the study of human diseases is limited by the availability of patient data and ethical concerns, and are thus often complemented with experimental work in model organisms, in particular mice (Mus musculus) (Fox, Barthold, Newcomer, Smith, & Quimby, 2006).

Perturbation of gene expression in humans can be estimated from mouse transcriptomic data (Brubaker, Proctor, Haigis, & Lauffenburger, 2019; Normand et al., 2018). Furthermore, previous studies suggest that pathway and TF footprints are evolutionarily conserved between mice and humans: pathway footprints derived from mouse B cells can provide valuable insights into human cancer (Tenenbaum, Walker, Utz, & Butte, 2008), and inferred prostate-specific gene regulatory networks of mice and humans overlap by over 70% (Aytes et al., 2014). This suggests that human functional genomics tools, which consider footprints as gene sets, could be applied on mice data. However, as of today there is no comprehensive study to prove this.

To validate whether pathway and TF footprints are evolutionarily conserved between mice and humans I performed a comprehensive benchmark study. I exploited two state of the art functional genomics approaches covering both aspects of gene regulatory networks: signaling pathways and transcriptional regulation. The first approach is PROGENy, a tool that estimated the activity of, originally, 11 signaling pathways from gene expression data (Schubert et al., 2018). It is based on consensus transcriptomic perturbation signatures, I refer to as footprints, of signaling pathways on gene expression. In this work Bence Szalai extended PROGENy with novel footprints of the signaling pathways Androgen, Estrogen, and WNT. The second approach is DoRothEA, a resource matching TFs with their transcriptional targets (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019), which allows me to estimate TF activity from gene expression data in humans by enriched regulon analysis (Alvarez et al., 2016). I consider the targets of a TF also as footprints of a TF on gene expression. I validated that both PROGENy and DoRothEA can recover mice perturbations, supporting my hypothesis about the conserved nature of pathway and TF footprints. To demonstrate the usability of PROGENy and DoRothEA I estimated pathway and TF activities for a large collection of mice and human diseases as well as chemical and genetic perturbation experiments. Based on the activities of the disease experiments I were able to recover known pathway/TF disease associations. For this, I constructed 738 novel disease sets matching 186 diseases with 467 disease experiments.

2.3 Methods

Benchmark dataset collection

I collected transcriptomic perturbation experiments in human and mouse profiled by single-channel microarrays from the database CREEDS (Zichen Wang et al., 2016), which contains among others resources single drug and single gene perturbation experiments from Gene Expression Omnibus (GEO). I extended this collection with manually curated perturbation experiments from ArrayExpress using the data collection and curation pipeline described previously (Schubert et al., 2018). Arrays with no raw data available or with no corresponding annotation package were discarded. I translated GEO accession ids to ArrayExpress accession ids and downloaded CEL files for all experiments using the function ArrayExpress from the BioConductor package ArrayExpress (version 1.40.0; Kauffmann et al. (2009)).

New PROGENy pathway

PROGENy is based on footprints which are consensus gene signatures delivered from pathway-related perturbation experiments. Bence Szalai added 3 new pathways footprints (Androgen, Estrogen and WNT) to the existing 11 pathways of PROGENy in this study. To collect the corresponding perturbation experiments, he queried ArrayExpress (Kauffmann et al., 2009) with keywords {'androgen,' 'DHT', 'testosterone'}, {'estrogen,' 'SERM,' 'tamoxifen'} and {'APC,' 'axin,' 'catenin,' 'Frizzled,' 'GSK3,' 'WNT'} for Androgen, Estrogen and WNT pathways, respectively. For further curation and PROGENy model fitting he used the pipeline described previously (Schubert et al., 2018).

Microarray processing

The processing steps from raw data to annotated probe levels, comprising quality control, background correction, normalization and annotation is described in the original PROGENy paper (Schubert et al., 2018).

Experiments with less than two control replicates remaining after the processing step were discarded. I used the BioConductor package limma (version 3.36.2; Phipson, Lee, Majewski, Alexander, & Smyth (2016)) to perform differential expression analysis calculating the contrast between perturbed and control replicates. Instead of log-fold changes I considered the moderated t-value as gene-level statistic.

Construction of mouse-PROGENy and calculation of pathway activities

The original PROGENy model is a matrix with footprint genes in rows and pathways in columns. The entries denote a measure accounting for how genes respond to pathway perturbation (up- or downregulation). Each pathway is limited to the top 100 most responsive genes.

To construct mouse-PROGENy I mapped the HGNC symbols of the original PROGENy matrix to their ortholog MGI-symbol using the BioConductor package biomaRt (version 2.36.1, ensembl release 96 April 2019; Durinck, Spellman, Birney, & Huber (2009)). The mapping can lead to duplicated genes. Either a single HGNC symbol is mapped to several MGI symbols or several HGNC symbols are mapped to a single MGI symbol. In the first case, the weight of the HGNC symbol is divided by the number of mapping MGI genes. In the second case, the weight of the new MGI symbol is the arithmetic mean value of all mapping HGNC symbols. Following this strategy I introduce a bias to human genes which have multiple orthologs in mice (e.g. the gene SERPINA3 that belongs to the footprint of JAK-STAT maps to 10 mouse genes: Serpina3a/b/c/f/g/i/j/k/m/n. Hence the JAK-STAT pathway is biased to SERPINA3). In order to compensate for this bias I extend the pathway footprints with a number of additional genes equal to the number of ortholog genes per human gene - 1. Finally, a mouse specific PROGENy matrix is retrieved (mouse-PROGENy) so that I can estimate pathway activity scores from mice gene expression data for the original 11 (EGFR, Hypoxia, JAK-STAT, MAPK, NFkB, p53, PI3K, TGFb, TNFa, Trail, VEGF) and the 3 newly added (Androgen, Estrogen, WNT) pathways. The pathway activity scores are calculated for each contrast using moderated t-values as gene-level statistic. Activity scores are pathway-wise z-score normalized.

Construction of mouse-DoRothEA and calculation of TF activities

DoRothEA is a resource linking human TFs to their direct transcriptional targets. Each TF is accompanied with a summary confidence level from A (high confidence) to E (low confidence) based on the amount of supporting TF's regulatory evidence. I inferred mouse-DoRothEA by mapping HGNC-symbols to their ortholog MGI-symbol using the BioConductor package BiomaRt (version 2.36.1, ensembl release 96 April 2019; Durinck, Spellman, Birney, & Huber (2009)). The mapping can lead to TFs with multiple confidence levels. To be more conservative, the lowest confidence level is chosen as TF-confidence level. The BioConductor package viper (version 1.14.0; Alvarez et al. (2016)) considers the regulons as gene sets and, thus, estimates TF activities from gene expression data using enriched regulon analysis. TF activity scores are computed for each contrast using moderated t-values as gene-level statistic. Only regulons comprised of at least 4 targets are tested. I consider the normalized enrichment score (NES) provided by viper as a measure for TF activity.

Quality control of single gene perturbation experiments

Single gene perturbation experiments provide the possibility for an intuitive quality control of the effect of the perturbation. If the gene-level statistic (t-value) sign of the perturbation target is not in agreement with the underlying perturbation ((+) for overexpression, (-) for knockdown/knockout) the perturbation experiment is considered unsuccessful and is thus discarded.

Computing ROC and PR curves

To transform the benchmark into a binary setup, all activity scores of experiments with negative perturbation effect (inhibition/knockdown) are multiplied by -1. This guarantees, that TFs/pathways belong to a binary class either deregulated or not regulated.

I computed the ROC-curves and associated statistics using the R package pROC (version 1.12.1; Robin et al. (2011)). For PR-curves I used the R package PRROC (version 1.3.1; Grau, Grosse, & Keilwagen (2015)). For the construction of ROC and PR curves I calculated for each perturbation experiment pathway (or TF) activities using PROGENy (or DoRothEA). As each perturbation experiment targets either a single pathway (or TF) only the activity score of the perturbed pathway (or TF) is associated with the positive class (e.g. EGFR pathway activity score in an experiment where EGFR was perturbed). Accordingly the activity scores of all non-perturbed pathways (or TFs) belong to the negative class (e.g. EGFR pathway activity score in an experiment where JAK-STAT pathway was perturbed). Using these positive and negative classes Sensitivity/(1-Specificity) or Precision/Recall values were calculated at different thresholds of activity, producing the ROC/PRC curves.

Downsampling true negatives

ROC curves are recommended when the numbers of true positives and true negatives are balanced (J. Davis & Goadrich, 2006). In order to balance my benchmark dataset I downsampled the number of true negatives to equal the number of true positives 3000 times and computed AUROC for each run.

Inference of disease sets using disease ontology network

To create disease sets I determined all related parent diseases of the diseases studied in CREEDs by using the function ancestors from the BioConductor package rols (version 2.9.1; `https://github.com/lgatto/rols/`) which provides an R interface to the Ontology Lookup service (J, B, L, & Parkinson, 2015). Each possible parent disease serves as a distinct disease set. CREEDs disease experiments which matches a child disease of a given disease set is considered as a set member.

Disease set enrichment analysis

To explore TF/Pathway-disease associations I downloaded all human and mouse disease signatures from the CREEDs database (Zichen Wang et al., 2016). I followed the processing steps described before resulting in an expression vector of moderated t-values for each disease experiment. I computed pathway and TF activity scores for each vector. To apply the Gene Set Enrichment Analysis framework (Subramanian et al., 2005) with my disease sets I used the BioConductor package fgsea (version 1.6.0; Sergushichev (2016)). Separately for each pathway/TF (e.g. only for EGFR), disease experiments (with associated diseases) are ranked based on the pathway/TF activity score (e.g. the activity score of EGFR). Subsequently, it is tested whether experiments belonging to the same disease set are enriched either at the top or at the bottom of the list. Disease sets with less than 5 and more than 45 members were discarded. P-values were adjusted for multiple comparisons using false discovery rate (FDR) (Benjamini & Hochberg, 1995).

2.4 Results

Benchmark pipeline

I established a benchmark pipeline to discover whether both PROGENy and DoRothEA human footprint methods could be applied to functionally characterize mice data (Figure 2.1B). Pathway/TF perturbation gene expression studies provide the opportunity to benchmark both tools: I can compare the predicted pathway and TF activities with the 'ground truth,' denoted as the original perturbed target. The database CREEDS (CRowd Extracted Expression of Differential Signatures) provides manually curated single drug and single gene perturbation experiments in human and mice (Zichen Wang et al., 2016). Additionally, I manually curated single drug perturbation experiments (see Methods). I included both perturbation directions - either activation/overexpression or inhibition/knockdown.

For the PROGENy validation I exploited both single drug and single gene perturbation studies. Experiments are considered to be relevant for my study if the perturbation target is a member or a gene encoding for a member of a PROGENy signaling pathway. I identified 347 experiments (123 single gene and 224 single drug perturbation; Figure 2.1C). These experiments cover 11 and 13 out of 14 possible pathways for human and mouse, respectively. These 14 pathways include Androgen, Estrogen, and WNT besides the 11 in the original PROGENy publication (see Methods) (Schubert et al., 2018). For DoRothEA I extracted only those single gene perturbation experiments where the target gene encodes for a TF which is defined by the human TF census from TFClass (Wingender, Schoeps, Haubrock, Krull, & Dönitz, 2018). In total I collected 302 single gene perturbation experiments covering 144 mouse TFs (Figure 2.1D).

To evaluate if PROGENy is applicable on mice data in the fairest way I would need to compare mouse-PROGENy vs a PROGENy version which was originally developed for application in mice. Since, to my knowledge, this resource does not currently exist I compared the performance of the newly derived mouse-PROGENy versus the original human-PROGENy tool. Note that this procedure introduces a bias towards the benchmark data as I benchmark mouse-PROGENy and human-PROGENy on independent data sets: one with human and the other with mouse pathway perturbation experiments. Regarding DoRothEA, there are resources that provide TF-target interactions for mice. Hence, I compared the newly derived mouse-DoRothEA versus dedicated mouse regulons from the TRRUST database (Han et al., 2018). Both mouse-PROGENy and mouse-DoRothEA were constructed by mapping human genes to their orthologs in mice (see Methods). To assess the model's prediction power I utilized the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves (see Methods).

Benchmarking PROGENy

To compare mouse-PROGENy and human-PROGENy unbiasedly, I included only pathways perturbed in both benchmark datasets. Moreover, I evaluated PROGENy's global performance across all pathways. Both models clearly performed better than random (AUROC of 0.71 with 95% confidence interval of 0.662–0.757 and AUROC of 0.656 with 95% confidence interval of 0.610–0.703 for human and mouse-PROGENy respectively; Figure 2.2A; ROC-curves for each pathway in Supplementary Figure A.1). AUROC was not significantly different between mouse and human (DeLong-test, $p = 0.113$). As my benchmark dataset is imbalanced (10% belong to the positive class) I also computed AUROC's upon multiple downsampling true negatives to equal the number of true positives with a resulting median AUROC equal to the AUROC of the unbalanced dataset (see Methods; Supplementary Figure A.2A and B). With precision-recall analysis, I obtained consistent results: human-PROGENy performed comparably to mouse-PROGENy (AUPRC of 0.254 and 0.246, respectively; Figure 2.2B; PR-curves for each pathway are provided in Supplementary Figure A.3. In addition, both performed better than a random model which would result in an AUPRC of 0.1. In summary, mouse-PROGENy performed comparably to human-PROGENy and better than a random model, regardless of the metric used. Thus, I conclude that PROGENy can recover pathway perturbations in mice.

Benchmarking DoRothEA

To evaluate if DoRothEA's regulons can functionally characterize mice data, I next compared the performance of mouse-DoRothEA to the performance of dedicated mouse regulons from the TRRUST database (Han et al., 2018). Human-DoRothEA was reconstructed by integrating different resources spanning from literature-curated databases to predictions of TF-target interactions (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019). Thereby, each TF is accompanied with a summary confidence level from A (high confidence) to E (low confidence) based on the amount of supporting TF's regulatory evidence. My novel mouse-DoRothEA regulons comprise in total 1170 TFs, targeting 17,512 unique targets with 402,937 interactions distributed across all confidence levels (see Methods; Supplementary Figure A.4A). In contrast, TRRUST covers 828 TFs, which overlap 553 TFs from mouse-DoRothEA (Supplementary Figure A.4B). Comparing similarity of overlapping regulons between TRRUST and mouse-DoRothEA revealed, for most regulons, substantial differences (Supplementary Figure A.4C). To benchmark the performance of mouse-DoRothEA and TRRUST unbiasedly, I consider only the 553 TFs which are available in both resources. I cover 34–76 TFs of those 553 TFs (dependent on mouse-DoRothEA confidence level) with my benchmark data (Supplementary Figure A.4D). Moreover, I evaluated DoRothEA's global performance across all TFs since there were not enough

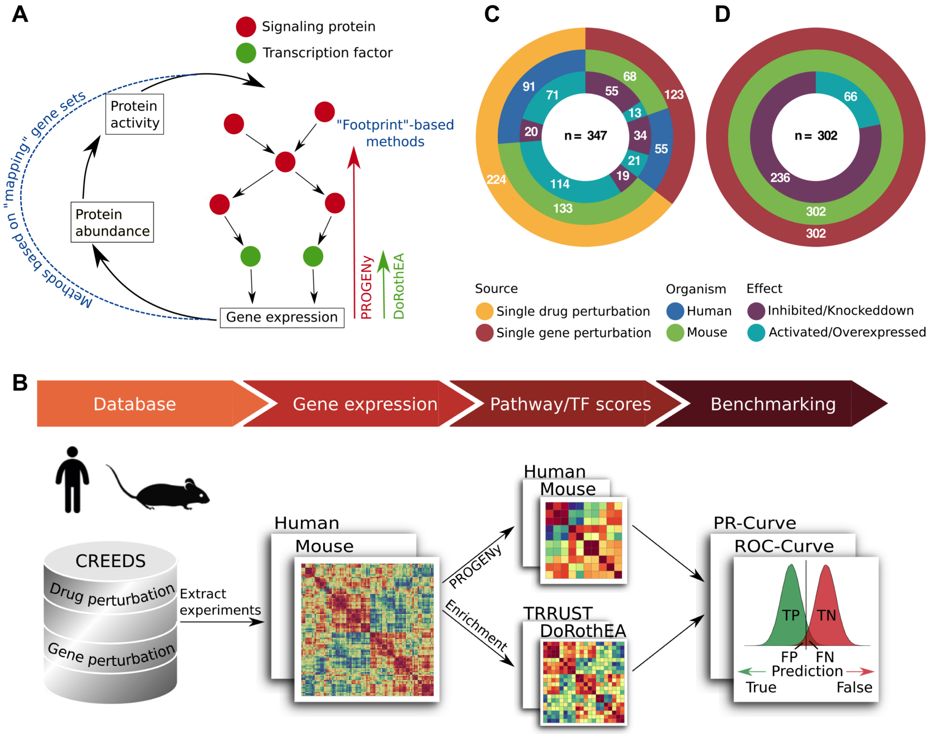


Figure 2.1: Benchmark overview. (A) Visualization of the classical 'mapping' framework, where transcript level is mapped to protein level and thus to protein activity in contrast to the footprint based methods PROGENy and DoRothEA. (B) Benchmark pipeline starting with the extraction of mouse and human single gene and single drug perturbation experiments from the CREEDs database. Pathway and TF activities are computed for each experiment separately based on their differential expression signature. For the PROGENy benchmark I compared human-PROGENy vs mouse-PROGENy. For DoRothEA benchmark I compared mouse-DoRothEA vs dedicated mouse regulons from TRRUST. I evaluate the performance of both approaches using ROC and PR-metrics. (CD) Overview of benchmark datasets for PROGENy (C) and DoRothEA (D), including the perturbation type, organism, and perturbation effect. Numbers indicate the amount of experiments in each group.

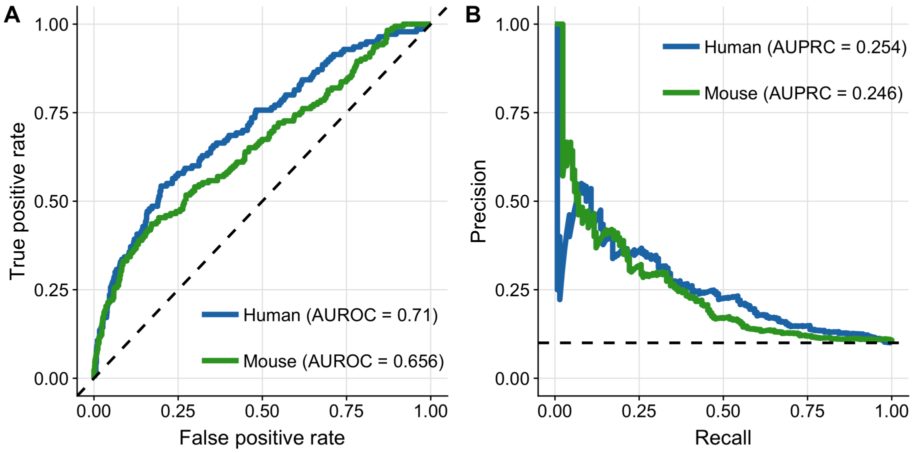


Figure 2.2: Benchmark of mouse-PROGENy. ROC-curve (A) and PR-curve (B) analysis comparing human-PROGENy vs. mouse-PROGENy. Dashed lines indicate the performance of a random models.

public data sets available to evaluate the performance at the TF level. In ROC space mouse-DoRothEA outperformed TRRUST at any confidence level combination (Figure 2.3A). However, in PR space I found that that TRRUST has an advantage throughout all confidence level combinations except AB where DoRothEA is slightly better (Figure 2.3B). All model subtypes performed better than a corresponding random model. In both metrics, I saw a peak at combined confidence level of A and B. Therefore, I decided to consider only TFs accompanied with the confidence levels A and B in the following analysis.

While both regulons resources performed better than random, mouse-DoRothEA (AUROC: 0.711, 95% confidence interval: 0.649–0.772) performed better than TRRUST (AUROC: 0.671, 95% confidence interval: 0.604–0.738; Figure 2.3C), but without a significant difference (DeLong-test, $p = 0.249$). As the DoRothEA benchmark dataset is even more imbalanced (2.63% belong to the positive class) I downsampled again true negatives to equal the number of true positives, showing that the median of downsampled AUROC is equal to the AUROC of the imbalanced dataset (see Methods; Supplementary Figure A.5A and B). In PR-space mouse-DoRothEA performed just as well as TRRUST (AUPRC of 0.108 for both; Figure 2.3D). Also both performed better than a random model with a corresponding AUPRC of 0.026. Considering the aforementioned results, I conclude that mouse-DoRothEA performs comparably to TRRUST and can thus recover transcriptional regulation in mice.

Pathway/TF-disease associations

Once shown that PROGENy and DoRothEA can also be applied to mice data, I investigated whether I can recover known associations between pathways/TFs and human diseases based on transcriptomic disease signatures of both mice and humans. I downloaded 469 disease signatures from the CREEDs database (Zichen Wang et al.,

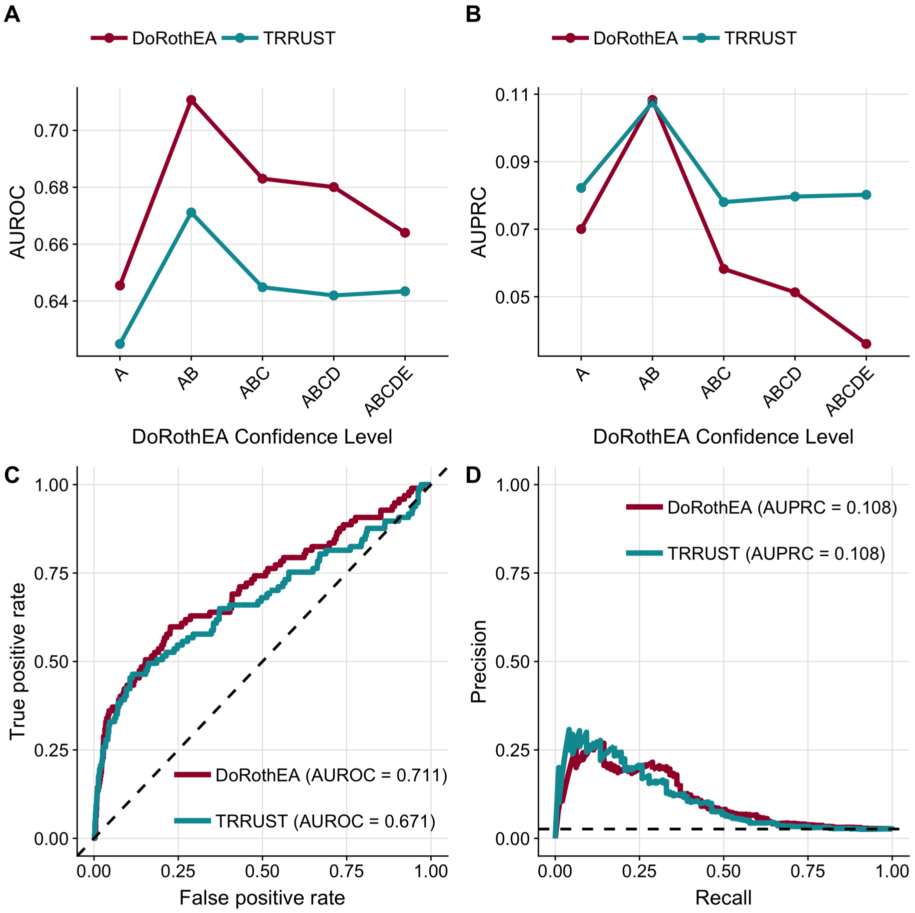


Figure 2.3: Benchmark of mouse-DoRothEA. DoRothEA result of ROC-curve (A) and PR-curve (B) analysis summarized in AUROC and AUPRC, respectively, for different confidence level cutoffs. ROC-curve (C) and PR-curve (D) analysis comparing mouse-DoRothEA filtered for TFs with confidence level A or B vs. mouse-TRRUST. Dashed lines indicate the performance of a random model.

2016) and computed pathway and TF activity scores for each experiment. To find associations I developed individual disease sets based on a disease ontology network from EBI Ontology Lookup Service (J, B, L, & Parkinson, 2015). Each node in this network represents a distinct disease set. If a node itself or a descendant of a node matched a CREEDs signature disease, the corresponding CREEDs experiment is considered as a member of the disease set (Figure 2.4A). In total I tested 732 distinct disease sets. Using these disease sets I found 434 significant (Gene set Enrichment Analysis (Subramanian et al., 2005); FDR ≤ 0.1 & $|\text{NES}| \geq 1$; see Methods) pathway-disease associations and 3586 significant (FDR ≤ 0.1 $|\text{NES}| \geq 1$) TF-disease associations covering 156 and 281 disease sets, respectively.

The results were, in general, dominated by upregulated activity of two TFs, ETS2 (100 associations) and E2F1 (88 associations; Figure 2.4B). Both are well-known oncogenes driving tumorigenesis (Fry & Inoue, 2018; Johnson, 2000). Accordingly, most of their associations I found were related to different forms of cancer. Similarly, I found the activity of the tumor suppressor TP53 to be downregulated in cancer (7 associations). Pathway specific associations were dominated by the PI3K pathway (86 associations; Figure 2.4C). Additionally, almost half of them were associations with different forms of cancer. My approach revealed for the majority of all cancer associations an elevated activity level of PI3K. This finding is in agreement with the literature that describes PI3K as having control over important hallmarks of cancer, i.e. cell cycle, survival, and metabolism (Fruman & Rommel, 2014). Also I found a strong upregulation of VEGF pathway in pancreatic cancer. Overexpressed VEGF (Vascular endothelial growth factor) is involved in angiogenesis and is considered as a diagnostic marker for pancreatic cancer (Costache et al., 2015). These examples emphasize the importance of signaling pathways and transcriptional regulation in the context of cancer diseases.

However, next to cancer related diseases I also recovered strong associations with other disease types, e.g. upregulated Hypoxia pathway activity in rheumatoid arthritis (Quiñonez-Flores, González-Chávez, & Pacheco-Tena, 2016). Also, NFKB1 and JAK-STAT showed elevated activity in immune and, therefore, leukocyte related diseases, such as inflammation of the lung, bowel, mucous membrane, or skin (Banerjee, Biehl, Gadina, Hasni, & Schwartz, 2017; Tak & Firestein, 2001).

In the context of chronic liver disease I recovered the role of PPARA. It's expression is reduced in hepatic stellate cells during liver fibrosis (Zardi et al., 2013). This finding is in agreement with my study as I found down regulated PPARA activity associated with the set 'liver disease.' Moreover, reduced PPARA activity was also significantly depleted within the disease sets 'hepatocellular carcinoma' and 'liver carcinoma.'

Altogether, I showed that PROGENy and DoRothEA are capable to recover known signaling pathway/TF disease association based on mice and human data.

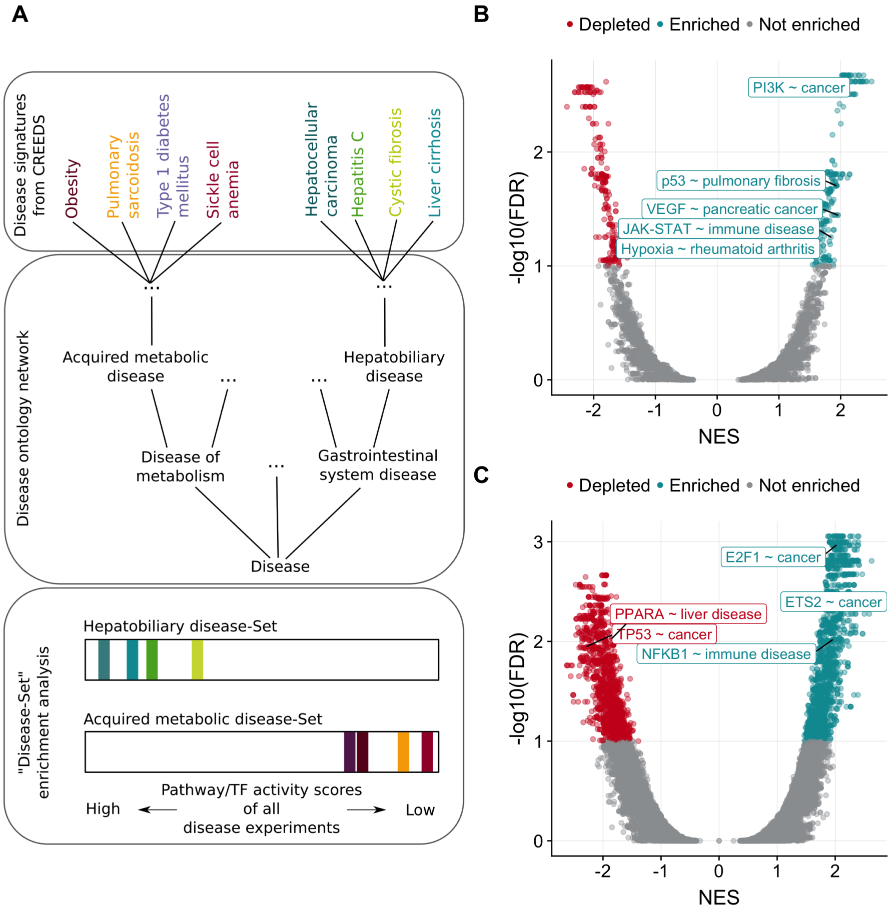


Figure 2.4: Pathway/TF-disease associations. (A) Framework of gene set enrichment adapted for 'disease set' enrichment. The disease sets are created based on a disease ontology network. Each node in the network represents a disease set. CREEDs diseases which are the node itself or descendants of a node are considered as corresponding disease set members. To perform the enrichment PROGENy and DoRothEA activity levels are ranked separately and checked whether a disease set is enriched either at the top or at the bottom of the ranked list. (B, C) Volcano plots showing separately for pathways and TFs the outcome of disease set enrichment. Selected associations are labelled.

2.5 Discussion

The evolutionarily conserved gene regulatory system between mouse and human suggests that the footprints of a pathway or TF on gene expression are evolutionarily conserved as well. This hypothesis has a direct impact on footprint methods developed for human application, such as PROGENy and DoRothEA. Both rely on gene sets comprising footprints and given that my assumption is true, they can be applied to mice data, which is an important resource for the study of human diseases. I addressed this question by establishing a benchmark pipeline to validate if DoRothEA and an extended version of PROGENy (footprints for Androgen, Estrogen and WNT were added by Bence Szalai) can be applied to functionally characterize mice data (Figure 2.1B).

I found that mouse-PROGENy is globally effective in inferring pathway activity on mouse data. However, the pathway-wise benchmark showed that the prediction power varies across pathways (Supplementary Figures A.1 and A.3). Especially for JAK-STAT, I saw a highly significant difference between mice and humans in ROC space for the benefit of humans (DeLong-test, p = 5e-5). Interestingly, I observed the inverse case for the pathway NFkB. Here, mouse-PROGENy tends to outperform human-PROGENy (DeLong-test, p = 0.057), while NFkB still performed well in human (Schubert et al., 2018). This difference emphasizes the importance of the quality of the benchmark data. The benchmark data in (Schubert et al., 2018) was curated very carefully by reviewing each perturbation experiment separately. My analysis is based on a broad collection of curated experiments via crowdsourcing. By their own nature, crowdsourcing projects cannot be fully controlled, and miss annotations can occur, which could contribute to the low performance I found for some pathways. Other pathways, with a low number of positive cases such as VEGF, must be interpreted with caution as reliability of ROC/PR tests decreases with decreasing positive cases.

Regarding mouse-DoRothEA, I found it's performance comparable to dedicated mouse regulons from TRRUST. However, I recommend the use of mouse-DoRothEA instead of TRRUST as it provides a better coverage at similar performance. Regulons with confidence levels A and B have been shown to perform the best for both resources. Including confidence level C almost doubled the TF coverage from 34 to 59 TFs (Supplementary Figure A.4B) but caused a performance drop. By including TFs labelled with confidence level C, I introduce regulons in my benchmark data that have not been thoroughly studied (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019), hence the drop in performance is expected.

Our above stated findings about the performance of PROGENy and DoRothEA support my initial hypothesis that footprints are evolutionarily conserved between mice and humans, however only indirectly. Comparative transcriptomic analysis of single drug and single gene perturbation experiments in mice and humans would be required to show this fact in a direct manner. Thus I conclude that it is reasonable to think that the footprints of a pathway or TF are evolutionarily conserved, at least at the level of my current footprint methods which rely on lists of genes.

Once shown that PROGENy and DoRothEA can also be applied to mouse data, I computed TF and pathway activities for a large collection of chemical and genetic perturbations and disease experiments. The results are provided as an interactive web application to browse corresponding pathway and TF activities. I envision that this resource can have broad applications including the study of diseases and therapeutics. Moreover, I demonstrated the usability of PROGENy and DoRothEA by recovering known pathway/TF disease associations using the aforementioned disease experiments. I found 4020 significant associations in total, where most were related to different forms of cancer, but I also recovered well-known associations of other disease types, such as liver disease (Figure 2.4B).

Finally, I believe that my finding of the conserved nature of footprints is especially interesting for further development of footprint methods. Integrating data from mice and humans will provide a much stronger data background for future model construction. Lastly, I speculate that the conserved nature of footprints will not hold to be exclusively true for mouse and human but will also extend to other mammals often used as model organisms.

2.6 Availability of data and materials

All source code is deposited at GitHub. Pathway and TF activities of perturbation and disease experiments can be browsed in a user friendly web application.

Chapter 3

Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data

3.1 Preface

The text of the following chapter is taken largely from the publication “Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data” (Christian H. Holland et al., 2020) that was originally written by myself. Changes were made to focus on my contribution to this project. Unless otherwise stated I performed all analyses myself. As the first author of this study the publishing house Springer Nature grants me the right to include this work in my dissertation.

3.2 Background

Gene expression profiles provide a blueprint of the status of cells. Thanks to diverse high-throughput techniques, such as microarrays and RNA-seq, expression profiles can be collected relatively easily and are hence very common. To extract functional and mechanistic information from these profiles, many tools have been developed that can, for example, estimate the status of molecular processes such as the activity of pathways or transcription factors (TFs). These functional analysis tools are broadly used and belong to the standard toolkit to analyze expression data (Essaghir et al., 2010; Hung, Yang, Hu, Weng, & DeLisi, 2012; Khatri, Sirota, & Butte, 2012; Nguyen, Shafi, Nguyen, & Draghici, 2019).

Functional analysis tools typically combine prior knowledge with a statistical method to gain functional and mechanistic insights from omics data. In the case of transcriptomics, prior knowledge is typically rendered as gene sets containing genes belonging to, e.g., the same biological process or to the same Gene Ontology (GO) annotation. The Molecular Signature Database (MSigDB) is one of the largest collections of curated and annotated gene sets (Liberzon et al., 2011). Statistical methods are as abundant as the different types of gene sets. Among them, the most commonly used are over-representation analysis (ORA) (Fisher, 1992) and Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). Still, there is a growing number of statistical methods spanning from simple linear models to advanced machine learning methods (Pang et al., 2006; Trescher, Münchmeyer, & Leser, 2017).

Recent technological advances in single-cell RNA-seq (scRNA-seq) enable the profiling of gene expression at the individual cell level (Tang et al., 2009). Multiple technologies and protocols have been developed, and they have experienced a dramatic improvement over recent years. However, single-cell data sets have a number of limitations and biases, including low library size and drop-outs. Bulk RNA-seq tools that focus on cell type identification and characterization as well as on inferring regulatory networks can be readily applied to scRNA-seq data (Stegle, Teichmann, & Marioni, 2015). This suggests that functional analysis tools should in principle be applicable to scRNA-seq data as well. However, it has not been investigated yet whether these limitations could distort and confound the results, rendering the tools not applicable to single-cell data.

In this paper, I benchmarked the robustness and applicability of various TF and pathway analysis tools on simulated and real scRNA-seq data. I focused on three tools for bulk and three tools for scRNA-seq data. The bulk tools were PROGENy (Schubert et al., 2018), DoRothEA (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019), and classical GO enrichment analysis, combining GO gene sets (Ashburner et al., 2000) with GSEA. PROGENy estimates the activity of 14 signaling pathways by combining corresponding gene sets with a linear model. DoRothEA is a collection of resources of TF's targets (regulons) that can serve as gene sets for TF activity inference. For this study, I coupled DoRothEA with the method VIPER (Alvarez et al., 2016) as it incorporates the mode of regulation of each TF-target interaction. Both PROGENy's and DoRothEA's gene sets are based on observing the transcriptomic consequences (the "footprint") of the processes of interest rather than the genes composing the process as gene sets (Dugourd & Saez-Rodriguez, 2019). This approach has been shown to be more accurate and informative in inferring the process's activity (Cantini et al., 2018; Schubert et al., 2018). The tools specifically designed for application on scRNA-seq data that I considered are SCENIC/AUCell (Aibar et al., 2017) and metaVIPER (Ding et al., 2018). SCENIC is a computational workflow that comprises the construction of gene regulatory networks (GRNs) from scRNA-seq data that are subsequently interrogated to infer TF activity with the statistical method AUCell. In addition, I coupled AUCell with the footprint-based gene sets from DoRothEA and PROGENy that I hereafter refer to as D-AUCell and P-AUCell. Using DoRothEA with both VIPER and AUCell on scRNA-seq for TF activity inference allowed me to compare the underlying statistical methods more objectively. metaVIPER is an extension of VIPER which is based on the same statistical method but relies on multiple GRNs such as tissue-specific networks.

I first benchmarked the tools on simulated single-cell transcriptome profiles. I found that on this in silico data the footprint-based gene sets from DoRothEA and PROGENy can functionally characterize simulated single cells. I observed that the performance of the different tools is dependent on the used statistical method and properties of the data, such as library size. I then used real scRNA-seq data upon CRISPR-mediated knock-out/knock-down of TFs (Dixit et al., 2016; Genga et al., 2019) to assess the performance of TF analysis tools. The results of this benchmark further supported my finding that TF analysis tools can provide accurate mechanistic insights into single cells. Finally, I demonstrated the utility of the tools for pathway and TF activity estimation on recently published data profiling a complex sample with 13 different scRNA-seq technologies (Mereu et al., 2020). Here, I showed that summarizing gene expression into TF and pathway activities preserves cell-type-specific information and leads to biologically interpretable results. Collectively, my results suggest that the bulk- and footprint-based TF and pathway analysis tools DoRothEA and PROGENy partially outperform the single-cell tools SCENIC, AUCell, and metaVIPER. Although on scRNA-seq data DoRothEA and PROGENy were less accurate than on bulk RNA-seq, I were still able to extract relevant functional insight from scRNA-seq data.

3.3 Methods

Functional analysis tools, gene set resources, and statistical methods

PROGENy

PROGENy is a tool that infers pathway activity for 14 signaling pathways (Androgen, Estrogen, EGFR, Hypoxia, JAK-STAT, MAPK, NFkB, PI3K, p53, TGFb, TNFa, Trail, VEGF, and WNT) from gene expression data (Christian H. Holland, Szalai, & Saez-Rodriguez, 2020; Schubert et al., 2018). By default pathway activity inference is based on gene sets comprising the top 100 most responsive genes upon corresponding pathway perturbation, which I refer to as footprint genes of a pathway. Each footprint gene is assigned a weight denoting the strength and direction of regulation upon pathway perturbation. Pathway scores are computed by a weighted sum of the product from expression and the weight of footprint genes.

DoRothEA

DoRothEA is a gene set resource containing signed transcription factor (TF)-target interactions (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019). Those interactions were curated and collected from different types of evidence such as literature curated resources, ChIP-seq peaks, TF binding site motifs, and interactions inferred directly from gene expression. Based on the number of supporting evidence, each interaction is accompanied by an interaction confidence level ranging from A to E, with A being the most confidence interactions and E the least. In addition, a summary TF confidence level is assigned (also from A to E) which is derived from the leading confidence level of its interactions (e.g., a TF is assigned confidence level A if at least ten targets have confidence level A as well). DoRothEA contains in total 470,711 interactions covering 1396 TFs targeting 20,238 unique genes. I use VIPER in combination with DoRothEA to estimate TF activities from gene expression data, as described in (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019).

GO-GSEA

I define GO-GSEA as an analysis tool that couples GO-terms from MsigDB with the GSEA framework (Subramanian et al., 2005).

VIPER

VIPER is a statistical framework that was developed to estimate protein activity from gene expression data using enriched regulon analysis performed by the algorithm aREA (Alvarez et al., 2016). It requires information about interactions (if possible signed) between a protein and its transcriptional targets and the likelihood of their interaction. If not further specified, this likelihood is set to 1. In the original workflow, this regulatory network was inferred from gene expression by the algorithm ARACNe providing mode of regulation and likelihood for each interaction (Margolin et al., 2006). However, it can be replaced by any other data resource reporting protein target interactions.

metaVIPER

metaVIPER is an extension of VIPER that uses multiple gene regulatory networks (Ding et al., 2018). TF activities predicted with each individual gene regulatory network are finally integrated to a consensus TF activity score.

SCENIC

SCENIC is a computational workflow that predicts TF activities from scRNA-seq data (Aibar et al., 2017). Instead of interrogating predefined regulons, individual regulons are constructed from the scRNA-seq data. First TF-gene co-expression modules are defined in a data-driven manner with GENIE3. Subsequently, those modules are refined via RcisTarget by keeping only those genes than contain the respective transcription factor binding motif. Once the regulons are constructed, the method AUCell scores individual cells by assessing for each TF separately whether target genes are enriched in the top quantile of the cell signature.

D-AUCell/P-AUCell

The statistical method AUCell is not limited to SCENIC regulons. In principle, it can be combined with any gene set resources. Thus, I coupled AUCell with gene sets from DoRothEA (D-AUCell) and PROGENy (P-AUCell). In comparison to other statistical methods, AUCell does not include weights of the gene set members. Thus, the mode of regulation or the likelihood of TF-target interactions or weights of the PROGENy gene sets are not considered for the computation of TF and pathway activities.

Application of PROGENy on single samples/cells and contrasts

I applied PROGENy on matrices of single samples (genes in rows and either bulk samples or single cells in columns) containing normalized gene expression scores or on contrast matrices (genes in rows and summarized perturbation experiments into contrasts in columns) containing logFCs. In the case of single sample analysis, the contrasts were built based on pathway activity matrices yielding the change in pathway activity (perturbed samples - control sample) summarized as logFC. Independent of the input matrix, I scaled each pathway to have a mean activity of 0 and a standard deviation of 1. I build different PROGENy versions by varying the number of footprint genes per pathway (100, 200, 300, 500, 1000 or all which corresponds to ~ 29,000 genes).

Application of DoRothEA on single samples/cells and contrasts

I applied DoRothEA in combination with the statistical method VIPER on matrices of single samples (genes in rows and either bulk samples or single cells in columns) containing normalized gene expression scores scaled gene-wise to a mean value of 0 and standard deviation of 1 or on contrast matrices (genes in rows and summarized perturbation experiments into contrasts in columns) containing logFCs. In the case of single sample analysis, the contrasts were built based on TF activity matrices yielding the change in TF activity (perturbed samples - control sample) summarized as logFC. TFs with less than four targets listed in the corresponding gene expression matrix were discarded from the analysis. VIPER provides a normalized enrichment score (NES) for each TF which I consider as a metric for the activity. I used the R package viper (version 1.17.0; Alvarez et al. (2016)) to run VIPER in combination with DoRothEA.

Application of GO-GSEA sets on contrasts

I applied GSEA with GO gene sets on contrast matrices (genes in rows and summarized perturbation experiments into contrasts in columns) containing logFCs that serve also as gene-level statistic. I selected only those GO terms which map to PROGENy pathways in order to guarantee a fair comparison between both tools. For the enrichment analysis, I used the R package fgsea (version 1.10.0; Sergushichev (2016)) with 1000 permutations per gene signature.

Application of metaVIPER on single samples

I ran metaVIPER with 27 tissue-specific gene regulatory networks which I constructed before for one of my previous studies (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019). Those tissue-specific gene regulatory networks were derived using ARACNe (Margolin et al., 2006) taking the database GTEx (Carithers et al., 2015) as tissue-specific gene expression sample resource. I applied metaVIPER on matrices of single samples (genes in rows and single cells in columns) containing normalized gene expression scores scaled gene-wise to a mean value of 0 and a standard deviation of 1. If required, contrasts were built based on TF activity matrices yielding the change in TF activity (perturbed samples - control sample) summarized as logFC. TFs with less than four targets listed in the corresponding input matrix were discarded from the analysis. metaVIPER provides a NES integrated across all regulatory networks for each TF which I consider as a metric for the activity. I used the R package viper (version 1.17.0; Alvarez et al. (2016)) to run metaVIPER.

Application of AUCell with either SCENIC, DoRothEA, or PROGENy gene sets on single samples

AUCell is a statistical method to determine specifically for single cells whether a given gene set is enriched at the top quantile of a ranked gene signature. Therefore, AUCell determines the area under the recovery curve to compute the enrichment score. I defined the top quantile as the top 5% of the ranked gene signature. I applied this method coupled with SCENIC, PROGENy, and DoRothEA gene sets. Before applying this method with PROGENy gene sets, I subsetted the footprint gene sets to contain only genes available in the provided gene signature. This guarantees a fair comparison as for the original PROGENy framework with a linear model, the intersection of footprint (gene set) members and signature genes are considered. I applied AUCell with SCENIC, PROGENy, and DoRothEA gene sets on matrices of single samples (genes in rows and single cells in columns) containing raw gene counts. Contrasts were built based on respective TF/pathway activity matrices yielding the change in TF/pathway activity (perturbed samples - control sample) summarized as logFC. For the AUCell analysis, I used the R package AUCell (version 1.5.5; Aibar et al. (2017)).

Induction of artificial low gene coverage in bulk microarray data

I induce the reduction of gene coverage by inserting zeros on the contrast level. In detail, I insert for each contrast separately randomly zeros until I obtained a predefined number of genes with a logFC unequal zero which I consider as “covered”/“measured” genes. I perform this analysis for a gene coverage of 500, 1000, 2000, 3000, 5000, 7000, 8000 and as reference all available genes. To account for stochasticity effects during inserting randomly zero, I repeat this analysis 25 times for each gene coverage value.

Simulation of single cells

Let C be a vector representing counts per gene for a single bulk sample. C is normalized for gene length and library size resulting in vector B containing TPM values per gene. I assume that samples are obtained from homogenous cell populations and that the probability of a dropout event is inversely proportional to the relative TPM of each measured gene in the bulk sample. Therefore, I define a discrete cumulative distribution function from the vector of gene frequencies $P = \frac{B}{|B|}$. To simulate a single cell from this distribution, I draw and aggregate L samples by inverse transform sampling. L corresponds to the library size for the count vector of the simulated single cell. I draw L from a normal distribution $N(\mu, \mu/2)$.

To benchmark the robustness of the methods, I vary the number of cells sampled from a single bulk sample (1, 10, 20, 30, 50, 100) and the value of μ (1000, 2000, 5000, 10,000, 20,000). To account for stochasticity effects during sampling, I repeat this analysis 25 times for each parameter combination.

Prior to normalization, I discarded cells with a library size lower than 100. I normalized the count matrices of the simulated cells by using the R package scran (version 1.11.27; Lun, McCarthy, & Marioni (2016)). Contrast matrices were constructed by comparing cells originating from one of the perturbation bulk samples vs cells originating from one of the control bulk samples.

Gene regulatory network (GRN) reconstruction using SCENIC

Javier Perales-Patón in close collaboration with me inferred GRNs on individual sub-datasets using the SCENIC workflow (version 1.1.2-2; Aibar et al. (2017)). In brief, gene expression was filtered using default parameters and log2-transformed for co-expression analysis following the recommendations by the authors. I identified potential targets of transcription factors (TFs) based on their co-expression to TFs using GENIE3 (v. 1.6.0, Random Forest with 1000 trees). I pruned co-expression modules to retrieve only putative direct-binding interactions using RcisTarget (v. 1.4.0) and the cis-regulatory DNA-motif databases for hg38 human genome assembly (Version 9 - mc9nr, with distances TSS+/- 10kbp and 500bpUp100Dw, from `https://resources.aertslab.org/cistarget/`) with default parameters. Only modules with a significant motif enrichment of the TF upstream were kept for the final GRN. While I were running the workflow, 75 genes out of 27,091 from the first DNA-motif database (TSS+/- 10kbp) were inconsistent, i.e., were not described in the second one (500bpUp100Dw), leading to an error of the workflow execution. Thus, these 75 genes were discarded from the database to complete the workflow.

Benchmarking process with ROC and PR metrics

To transform the benchmark into a binary setup, all activity scores of experiments with negative perturbation effect (inhibition/knockdown) are multiplied by -1. This guarantees that TFs/pathways belong to a binary class either deregulated or not regulated and that the perturbed pathway/TF has in the ideal case the highest activity.

I performed the ROC and PR analysis with the R package yardstick (version 0.0.3; `https://github.com/tidymodels/yardstick`). For the construction of ROC and PR curves, I calculated for each perturbation experiment pathway (or TF) activities. As each perturbation experiment targets either a single pathway (or TF), only the activity score of the perturbed pathway (or TF) is associated with the positive class (e.g., EGFR pathway activity score in an experiment where EGFR was perturbed). Accordingly, the activity scores of all non-perturbed pathways (or TFs) belong to the negative class (e.g., EGFR pathway activity score in an experiment where the JAK-STAT pathway was perturbed). Using these positive and negative classes, Sensitivity/(1-Specificity) or Precision/Recall values were calculated at different thresholds of activity, producing the ROC/PR curves.

Collecting, curating, and processing of transcriptomic data

General robustness study

I extracted single-pathway and single-TF perturbation data profiled with microarrays from a previous study conducted by me (Christian H. Holland, Szalai, & Saez-Rodriguez, 2020). I followed the same procedure of collection, curating, and processing the data as described in the previous study.

In silico benchmark

For the simulation of single cells, I collected, curated, and processed single TF and single pathway perturbation data profiled with bulk RNA-seq. I downloaded basic metadata of single TF perturbation experiments from the ChEA3 web-server (Keenan et al., 2019) and refined the experiment and sample annotation. Metadata of single pathway perturbation experiments were manually extracted by me from Gene Expression Omnibus (GEO) (Edgar, Domrachev, & Lash, 2002). Count matrices for all those experiments were downloaded from ARCHS4 (Lachmann et al., 2018).

I normalized count matrices by first calculating normalization factors and second transforming count data to log2 counts per million (CPM) using the R packages edgeR (version 3.25.8; Robinson, McCarthy, & Smyth (2010)) and limma (version 3.39.18; Ritchie et al. (2015)), respectively.

In vitro benchmark

To benchmark VIPER on real single-cell data, Jan Gleixner and I inspected related literature and identified two publications which systematically measure the effects of transcription factors on gene expression in single cells:

Dixit et al. introduced Perturb-seq and measured the knockout-effects of ten transcription factors on K562 cells 7 and 13 days after transduction (Dixit et al., 2016). Jan Gleixner downloaded the expression data from GEO (GSM2396858 and GSM2396859) and sgRNA-cell mappings made available by the author upon request in the files promoters_concat_all.csv (for GSM2396858) and pt2_concat_all.csv (for GSM2396859) on github.com/asncd/MIMOSCA. He did not consider the High MOI dataset due to the expected high number of duplicate sgRNA assignments. Cells were quality filtered based on expression, keeping the upper half of cells for each dataset. Only sgRNAs detected in at least 30 cells were used. For the day 7 dataset, 16,507, and for day 13 dataset, 9634 cells remained for benchmarking.

Ryan et al. measured knockdown effects of 50 transcription factors implicated in human definitive endoderm differentiation using a CRISPRi variant of CROPseq in human embryonic stem cells 6 days after transduction (Genga et al., 2019). Jan Gleixner obtained data of both replicates from GEO (GSM3630200, GSM3630201), which include sgRNA counts next to the rest of the transcription. He refrained from using the targeted sequencing of the sgRNA in GSM3630202, GSM3630203 as it contained less clear mappings due to amplification noise. Expression data lacked information on mitochondrial genes, and therefore, no further quality filtering of cells was performed. From this dataset, only sgRNAs detected in at least 100 cells were used. A combined 5282 cells remained for benchmarking.

Analysis was limited to the 10,000 most expressed genes for all three datasets.

I normalized the count matrices for each individual dataset (Perturb-Seq (7d), Perturb-Seq (13d), and CRISPRi) separately by using the R package scran (version 1.11.27; Lun, McCarthy, & Marioni (2016)).

Human Cell Atlas study

This scRNA-seq dataset originates from a benchmark study of the Human Cell Atlas project and is available on GEO (GSE133549) (Mereu et al., 2020). The dataset consists of PBMCs and a HEK293T sample which was analyzed with 13 different scRNA-seq technologies (CEL-Seq2, MARS-Seq, Quartz-Seq2, gmcSCRB-Seq, ddSEQ, ICELL8, C1HT-Small, C1HT-Medium, Chromium, Chromium(sn), Drop-seq, inDrop). Most cells are annotated with a specific cell type/cell line (CD4 T cells, CD8 T cells, NK cells, B cells, CD14+ monocytes, FCGR3A+ monocytes, dendritic cells, megakaryocytes, HEK cells). Megakaryocytes (due to their low abundance) and cells without annotation were discarded from this analysis.

I normalized the count matrices for each technology separately by using the R package scran (version 1.11.27; Lun, McCarthy, & Marioni (2016)).

Dimensionality reduction with UMAP and assessment of cluster purity

I used the R package umap (version 0.2.0.0) calling the Python implementation of Uniform Manifold Approximation and Projection (UMAP) with the argument "method = 'umap-learn' " to perform dimensionality reduction on various input matrices (gene expression matrix, pathway/TF activity matrix, etc.). I assume that the dimensionality reduction will result in clustering of cells that corresponds well to the cell type/cell type family. To assess the validity of this assumption, I assigned a cell-type/cell family-specific cluster-id to each point in the low-dimensional space. I then defined a global cluster purity measure based on silhouette widths (Rousseeuw, 1987), which is a well-known clustering quality measure.

Given the cluster assignments, in the low-dimensional space, for each cell, the average distance (a) to the cells that belong to the same cluster is calculated. Then, the smallest average distance (b) to all cells belonging to the newest foreign cluster is calculated. The difference, between the latter and the former, indicates the width of the silhouette for that cell, i.e., how well the cell is embedded in the assigned cluster. To make the silhouette widths comparable, they are normalized by dividing the difference with the larger of the two average distances.

Therefore, the possible values for the silhouette widths lie in the range -1 to 1, where higher values indicate good cluster assignment, while lower values close to 0 indicate poor cluster assignment. Finally, the average silhouette width for every cluster is calculated, and averages are aggregated to obtain a measure of the global purity of clusters. For the silhouette analysis, I used the R package cluster (version 2.0.8).

For statistical analysis of cluster quality, I fitted a linear model score=f(scRNA-seq protocol+input matrix), where score corresponds to average silhouette width for a given scRNA-seq protocol - input matrix pair. Protocol and input matrix are factors, with reference level Quartz-Seq2 and positive control, respectively. I fitted two separate linear models for transcription factor and pathway activity inference methods. I report the estimates and p values for the different coefficients of these linear models. Based on these linear models, I performed a two-way ANOVA and pairwise comparisons using TukeyHSD post hoc test.

3.4 Results

Robustness of bulk-based TF and pathway analysis tools against low gene coverage

Single-cell RNA-seq profiling is hampered by low gene coverage due to drop-out events (Kharchenko, Silberstein, & Scadden, 2014). In my first analysis, I focused solely on the low gene coverage aspect and whether tools designed for bulk RNA-seq can deal with it. Specifically, I aimed to explore how DoRothEA, PROGENy, and GO gene sets combined with GSEA (GO-GSEA) can handle low gene coverage in general, independently of other technical artifacts and characteristics from scRNA-seq protocols. Thus, I conducted this benchmark using bulk transcriptome benchmark data. In these studies, single TFs and pathways are perturbed experimentally, and the transcriptome profile is measured before and after the perturbation. These experiments can be used to benchmark tools for TF/pathway activity estimation, as they should estimate correctly the change in the perturbed TF or pathway. The use of these datasets allowed me to systematically control the gene coverage (see the "Methods" section). The workflow consisted of four steps (Supplementary Figure B.1a). In the first step, I summarized all perturbation experiments into a matrix of contrasts (with genes in rows and contrasts in columns) by differential gene expression analysis. Subsequently, I randomly replaced, independently for each contrast, logFC values with 0 so that I obtain a predefined number of "covered" genes with a logFC unequal to zero. Accordingly, a gene with a logFC equal to 0 was considered as missing/not covered. Then, I applied DoRothEA, PROGENy, and GO-GSEA to the contrast matrix, subsetted only to those experiments which are suitable for the corresponding tool: TF perturbation for DoRothEA and pathway perturbation for PROGENy and GO-GSEA. I finally evaluate the global performance of the methods with receiver operating characteristic (ROC) and precision-recall (PR) curves (see the "Methods" section). This process was repeated 25 times to account for stochasticity effects during inserting zeros in the contrast matrix (see the "Methods" section).

DoRothEA's TFs are accompanied by an empirical confidence level indicating the confidence in their regulons, ranging from A (most confident) to E (less confident; see the "Methods" section). For this benchmark, I included only TFs with confidence levels A and B (denoted as DoRothEA (AB)) as this combination has a reasonable tradeoff between TF coverage and performance (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019). In general, the performance of DoRothEA dropped as gene coverage decreased. While it showed reasonable prediction power with all available genes (AUROC of 0.690), it approached almost the performance of a random model (AUROC of 0.5) when only 500 genes were covered (mean AUROC of 0.547, Figure 3.1a, and similar trend with AUPRC, Supplementary Figure B.1b). I next benchmarked pathway activities estimated by PROGENy and GO-GSEA. In the original PROGENy framework, 100 footprint genes are used per pathway to compute pathway activities by default, as it has been shown that this leads to the best performance on bulk samples (Schubert et al., 2018). However, one can extend the footprint size to cover more genes of the expression profiles. I reasoned that this might counteract low gene coverage and implemented accordingly different PROGENy versions (see the "Methods" section). With the default PROGENy version (100 footprint genes per pathway), I observed a clear drop in the global performance with decreasing gene coverage, even though less drastic than for DoRothEA (from AUROC of 0.724 to 0.636, Figure 3.1b, similar trends with AUPRC, Supplementary Figure B.1c). As expected, PROGENy performed the best with 100 footprint genes per pathway when there is complete gene coverage. The performance differences between the various PROGENy versions shrank with decreasing gene coverage. This suggests that increasing the number of footprint genes can help to counteract low gene coverage. To provide a

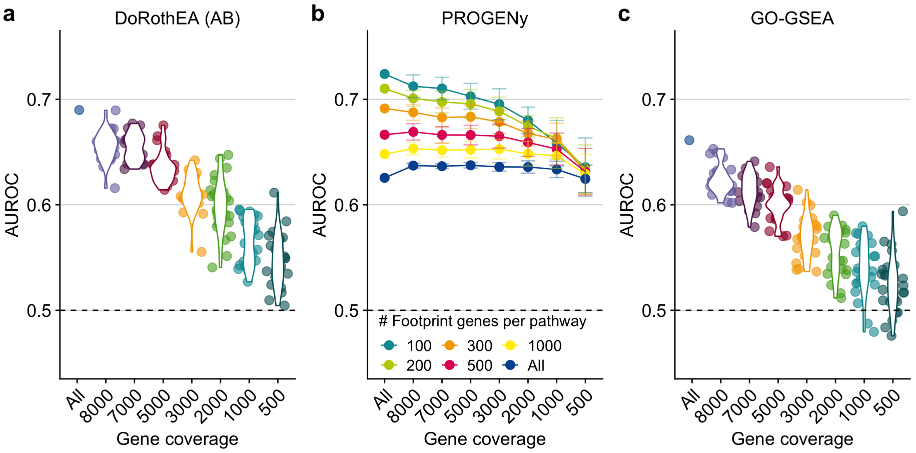


Figure 3.1: Testing the robustness of DoRothEA (AB), PROGENy, and GO-GSEA against low gene coverage. a DoRothEA (AB) performance (area under ROC curve, AUROC) versus gene coverage. b PROGENy performance (AUROC) for different number of footprint genes per pathway versus gene coverage. c Performance (AUROC) of GO-GSEA versus gene coverage. The dashed line indicates the performance of a random model. The colors in a and c are meant only as a visual support to distinguish between the individual violin plots and jittered point.

fair comparison between PROGENy and GO-GSEA, I used only those 14 GO terms that match the 14 PROGENy pathways (Supplementary Figure B.1d). In general, GO-GSEA showed weaker performance than PROGENy. The decrease in performance was more prominent as gene coverage decreased (from AUROC of 0.662 to 0.525, Figure 3.1c, and similar trend with AUPRC, Supplementary Figure B.1e). With a gene coverage of less than 2000 genes, GO-GSEA performance was no better than random.

As my benchmark data set comprises multiple perturbation experiments per pathway, I also evaluated the performance of PROGENy and GO-GSEA at the pathway level (Supplementary Figure B.2a and b). The pathway-wise evaluation supported my finding that PROGENy outperforms GO-GSEA across all gene coverages, but the performance between pathways is variable.

In summary, this first benchmark provided insight into the general robustness of the bulk-based tools DoRothEA, PROGENy, and GO-GSEA with respect to low gene coverage. DoRothEA performed reasonably well down to a gene coverage of 2000 genes. The performance of all different PROGENy versions was robust across the entire gene coverage range tested. GO-GSEA showed a worse performance than PROGENy, especially in the low gene coverage range. Since DoRothEA and PROGENy showed promising performance in low gene coverage ranges, I decided to explore them on scRNA-seq data. Due to its poor performance, I did not include GO-GSEA in the subsequent analyses.

Benchmark on simulated single-cell RNA-seq data

For the following analyses, I expanded the set of tools with the statistical methods AUCell that I decoupled from the SCENIC workflow (Aibar et al., 2017) and metaVIPER (Ding et al., 2018). Both methods were developed specifically for scRNA-seq analysis and thus allow the comparison of bulk vs single-cell based tools on scRNA-seq data. AUCell is a statistical method that is originally used with GRNs constructed by SCENIC and assesses whether gene sets are enriched in the top quantile of a ranked gene signature (see the "Methods" section). In this study, I combined AUCell with DoRothEA's and PROGENy's gene sets (referred to as D-AUCell and P-AUCell, respectively). metaVIPER is an extension of VIPER and requires multiple gene regulatory networks instead of a single network. In my study, I coupled 27 tissue-specific gene regulatory networks with metaVIPER, which provides a single TF consensus activity score estimated across all networks (see the "Methods" section). To benchmark all these methods on single cells, ideally, I would have scRNA-seq datasets after perturbations of TFs and pathways. However, these datasets, especially for pathways, are currently very rare. To perform a comprehensive benchmark study, I developed a strategy to simulate samples of single cells using bulk RNA-seq samples from TF and pathway perturbation experiments.

A major cause of drop-outs in single-cell experiments is the abundance of transcripts in the process of reverse-transcription of mRNA to cDNA (Kharchenko, Silberstein, & Scadden, 2014). Thus, my simulation strategy was based on the assumption that genes with low expression are more likely to result in drop-out events.

The simulation workflow started by transforming read counts of a single bulk RNA-seq sample to transcripts per million (TPM), normalizing for gene length and library size. Subsequently, for each gene, I assigned a sampling probability by dividing the individual TPM values with the sum of all TPM values. These probabilities are proportional to the likelihood for a given gene not to "drop-out" when simulating a single cell from the bulk sample. I determined the total number of gene counts for a simulated single cell by sampling from a normal distribution with a mean equal to the desired library size which is specified as the first parameter of the simulation. I refer hereafter to this number as the library size. For every single cell, I then sampled with replacement genes from the gene probability vector up to the determined library size. The frequency of occurrence of individual genes becomes the new gene count in the single cell. The number of simulated single cells from a single bulk sample can be specified as the second parameter of the simulation. Of note, this parameter is not meant to reflect a realistic number of cells, but it is rather used to investigate the loss of information: the lower the number of simulated cells, the more information is lost from the original bulk sample (Figure 3.2a; see the "Methods" section). This simple workflow guaranteed that the information of the original bulk perturbation is preserved and scRNA-seq characteristics, such as drop-outs, low library size, and a high number of samples/cells are introduced. Our bulk RNA-seq samples comprised 97 single TF perturbation experiments targeting 52 distinct TFs and 15 single pathway perturbation experiments targeting 7 distinct pathways (Supplementary Figure B.3a

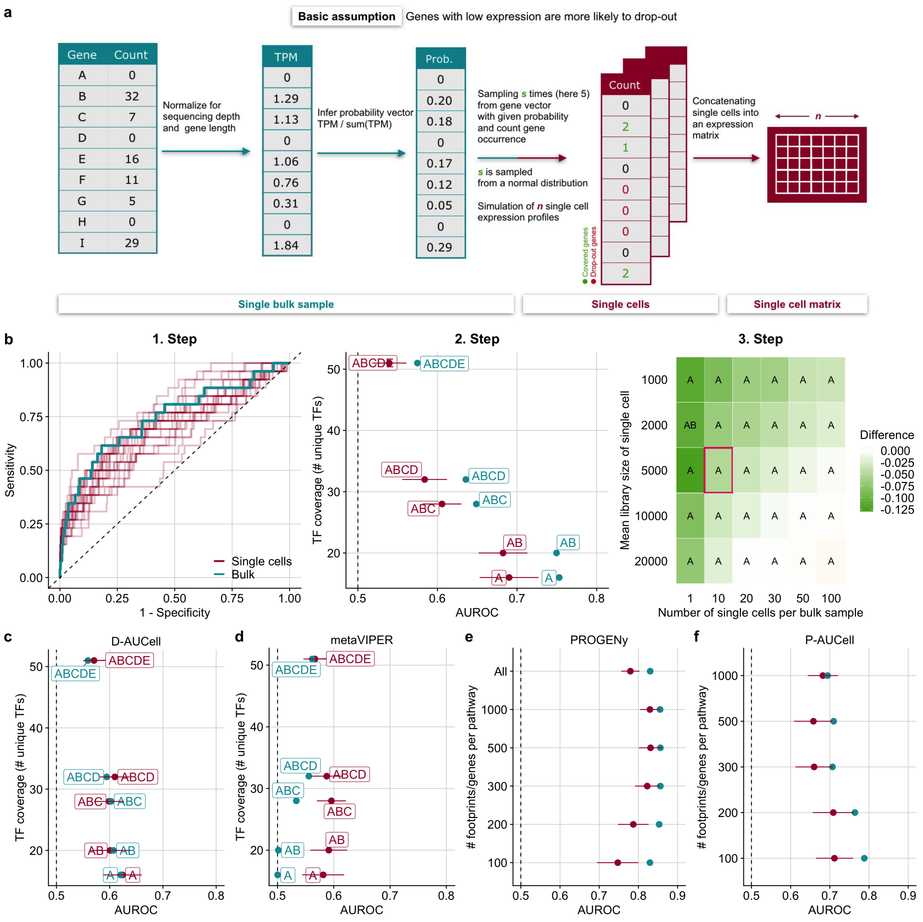


Figure 3.2: Benchmark results of TF and pathway analysis tools on simulated scRNA-seq data. a Simulation strategy of single cells from an RNA-seq bulk sample. b Example workflow of DoRothEA's performance evaluation on simulated single cells for a specific parameter combination (number of cells=10, mean library size=5000). 1. Step: ROC-curves of DoRothEA's performance on single cells (25 replicates) and on bulk data including only TFs with confidence level A. 2. Step: DoRothEA performance on single cells and bulk data summarized as AUROC vs TF coverage. TF coverage denotes the number of distinct perturbed TFs in the benchmark dataset that are also covered by the gene set resource. Results are provided for different combinations of DoRothEA's confidence levels (A, B, C, D, E). Error bars of AUROC values depict the standard deviation and correspond to different simulation replicates. Step 3: Averaged difference across all confidence level combinations between AUROC of single cells and bulk data for all possible parameter combinations. The letters within the tiles indicates which confidence level combination performs the best on single cells. The tile marked in red corresponds to the parameter setting used for previous plots (Steps 1 and 2). c D-AUCell and d metaVIPER performance on simulated single cells summarized as AUROC for a specific parameter combination (number of cells=10, mean library size=5000) and corresponding bulk data vs TF coverage. e, f Performance results of e PROGENy and f P-AUCell on simulated single cells for a specific parameter combination (number of cells=10, mean library size=5000) and corresponding bulk data in ROC space vs number of footprint genes per pathway. b–f The dashed line indicates the performance of a random model.

and b; see the “Methods” section). I repeated the simulation of single cells from each bulk sample template to account for the stochasticity of the simulation procedure. I tested my simulation strategy by comparing the characteristics of the simulated cells to real single cells. In this respect, I compared the count distribution (Supplementary Figure B.4a), the relationship of mean and variance of gene expression (Supplementary Figure B.4b), and the relationship of library size to the number of detected genes (Supplementary Figure B.4c). These comparisons suggested that my simulated single cells closely resemble real single cells and are thus suitable for benchmarking.

Unlike in my first benchmark, I applied the TF and pathway analysis tools directly on single samples/cells and built the contrasts between perturbed and control samples at the level of pathway and TF activities (see the “Methods” section). I compared the performance of all tools to recover the perturbed TFs/pathways. I also considered the performance on the template bulk data, especially for the bulk-based tools DoRothEA and PROGENy, as a baseline for comparison to their respective performance on the single-cell data.

I show, as an example, the workflow of the performance evaluation for DoRothEA (Figure 3.2b, 1. Step). As a first step, I applied DoRothEA to single cells generated for one specific parameter combination and bulk samples, performed differential activity analysis (see the “Methods” section), and evaluated the performance with ROC and PR curves including only TFs with confidence level A. In this example, I set the number of cells to 10 as this reflects an observable loss of information of the original bulk sample and the mean library size to 5000 as this corresponds to a very low but still realistic sequencing depths of scRNA-seq experiments. Each repetition of the simulation is depicted by an individual ROC curve, which shows the variance in the performance of DoRothEA on simulated single-cell data (Figure 3.2b, 1. Step). The variance decreases as the library size and the number of cells increase (which holds true for all tested tools, Supplementary Figure B.5a–e). The shown ROC curves are summarized into a single AUROC value for bulk and mean AUROC value for single cells. I performed this procedure also for different TF confidence level combinations and show the performance change in these values in relation to the number of distinct perturbed TFs in the benchmark that are also covered by the gene set resources that I refer to as TF coverage (Figure 3.2b, 2. Step). For both bulk and single cells, I observe a tradeoff between TF coverage and performance caused by including different TF confidence level combinations in the benchmark. This result is supported by both AUROC and AUPRC (Supplementary Figure B.6a) and corresponds to my previous findings (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019). The performance of DoRothEA on single cells does not reach the performance on bulk, though it can still recover TF perturbations on the simulated single cells reasonably well. This is especially evident for the most confident TFs (AUROC of 0.690 for confidence level A and 0.682 for the confidence level combination AB). Finally, I explore the effect of the simulation parameters library size and the number of cells on the performance by performing the previously described analysis for all combinations of library sizes and cell numbers. I computed the mean difference between AUROC scores of single-cell and bulk data across all confidence level combinations. A negative difference indicates that the tool of interest performs overall better on bulk data than on scRNA-seq data, and a positive difference that it performs better on scRNA-seq. I observed a gradually decreasing negative difference approaching 0 when the size of the library and the number of cells increase (Figure 3.2b, 3. Step, and Supplementary Figure B.7a). Note, however, that the number of cells and thus the amount of lost information of the original bulk sample has a stronger impact on the performance than the mean library size. In addition, I identified the best performing combination of DoRothEA's TF confidence levels for different library sizes and the number of single cells. Thus, the results can be used as recommendations for choosing the confidence levels on data from an experiment with comparable characteristics in terms of sequencing depths.

Similarly to DoRothEA, I also observed for D-AUCell a tradeoff between TF coverage and performance on both single cells and bulk samples when using the same parameter combination as before (Figure 3.2c, similar trend with AUPRC Supplementary Figure B.6b). The summarized performance across all confidence level combinations of D-AUCell on single cells slightly outperformed its performance on bulk samples (AUROC of 0.601 on single cells and 0.597 on bulk). This trend becomes more evident with increasing library size and the number of cells (Supplementary Figure B.7b).

For the benchmark of metaVIPER, I assigned confidence levels to the tissue-specific GTEx regulons based on DoRothEA's gene set classification. This was done for consistency with DoRothEA and D-AUCell, even if there is no difference in confidence among them. Hence, for metaVIPER, I do not observe a tradeoff between TF coverage and performance (Figure 3.2d, similar trend with AUPRC Supplementary Figure B.6c). As opposed to D-AUCell, metaVIPER performed clearly better on single cells than on bulk samples across all confidence level combinations (AUROC of 0.584 on single cells and 0.531 on bulk). This trend increased with increasing library size and number of cells (Supplementary Figure B.7c). However, the overall performance of metaVIPER is worse than the performance of DoRothEA and D-AUCell. In summary, the bulk-based tool DoRothEA performed the best on the simulated single cells followed by D-AUCell. metaVIPER performed slightly better than a random model.

For the benchmark of pathway analysis tools, I observed that PROGENy performed well across different number of footprint genes per pathway, with a peak at 500 footprint genes for both single cells and bulk (AUROC of 0.856 for bulk and 0.831 for single cells, Figure 3.2e, similar trend with AUPRC Supplementary Figure B.6d). A better performance for single-cell analysis with more than 100 footprint genes per pathway is in agreement with the previous general robustness study that suggested that a higher number of footprint genes can counteract low gene coverage. Similarly to the benchmark of TF analysis tools, I studied the effect of the simulation parameters on the performance of pathway analysis tools. I averaged for each parameter combination the performance difference between single cells and bulk across the different versions of PROGENy. For the parameter combination associated with Figure 3.2e (number of cells=10, mean library size=5000), the average distance is negative showing that the performance of PROGENy on bulk was, in general, better than on single-cell data. Increasing the library size and the number of cells improved the performance of PROGENy on single cells reaching almost the same performance as on bulk samples (Supplementary Figure B.7d). For most parameter combinations, PROGENy with 500 or 1000 footprint genes per pathway yields the best performance.

For P-AUCell, I observed a different pattern than for PROGENy as it worked best with 100 footprint genes per pathway for both single cells and bulk (AUROC of 0.788 for bulk and 0.712 for single cells, Figure 3.2f, similar trends with AUPRC Supplementary Figure B.6e). Similar to PROGENy, increasing the library size and the number of cells improved the performance, but not to the extent of its performance on bulk (Supplementary Figure B.7e). For most parameter combinations, P-AUCell with 100 or 200 footprint genes per pathway yielded the best performance.

In summary, both PROGENy and P-AUCell performed well on the simulated single cells, and PROGENy performed slightly better. For pathway analysis, P-AUCell did not perform better on scRNA-seq than on bulk data. I then went on to perform a benchmark analysis on real scRNA-seq datasets.

Benchmark on real single-cell RNA-seq data

After showing that the footprint-based gene sets from DoRothEA and PROGENy can handle low gene coverage and work reasonably well on simulated scRNA-seq data with different statistical methods, I performed a benchmark on real scRNA-seq data. However, single-cell transcriptome profiles of TF and pathway perturbations are very rare. To my knowledge, there are no datasets of pathway perturbations on single-cell level comprehensive enough for a robust benchmark of pathway analysis tools. For tools inferring TF activities, the situation is better: recent studies combined CRISPR knock-outs/knock-down of TFs with scRNA-seq technologies (Dixit et al., 2016; Genga et al., 2019) that can serve as potential benchmark data.

The first dataset is based on the Perturb-seq technology, which contains 26 knock-out perturbations targeting 10 distinct TFs after 7 and 13 days of perturbations (Supplementary Figure B.8a) (Dixit et al., 2016). To explore the effect of perturbation time, I divided the dataset into two sub-datasets based on perturbation duration (Perturb-seq (7d) and Perturb-seq (13d)). The second dataset is based on CRISPRi protocol and contains 141 perturbation experiments targeting 50 distinct TFs (Genga et al., 2019) (Supplementary Figure B.8a). The datasets showed a variation in terms of drop-out rate, the number of cells, and sequencing depths (Supplementary Figure B.8b).

To exclude bad or unsuccessful perturbations in the case of CRISPRi experiments, I discarded experiments when the logFC of the targeted gene/TF was greater than 0 (12 out of 141, Supplementary Figure B.8c). This quality control is important only in the case of CRISPRi, as it works on the transcriptional level. Perturb-seq (CRISPR knock-out) acts on the genomic level, so I cannot expect a clear relationship between KO efficacy and transcript level of the target. Note that the logFCs of both Perturb-seq sub-datasets are in a narrower range in comparison to the logFCs of the

CRISPRi dataset (Supplementary Figure B.8d). The perturbation experiments that passed this quality check were used in the following analyses.

I also considered the SCENIC framework for TF analysis (Aibar et al., 2017). I inferred GRNs for each sub-dataset using this framework (see the "Methods" section). I set out to evaluate the performance of DoRothEA, D-AUCell, metaVIPER, and SCENIC on each benchmark dataset individually.

To perform a fair comparison among the tools, I pruned their gene set resources to the same set of TFs. However, the number of TFs in the dataset-specific SCENIC networks was very low (109 for Perturb-Seq (7d), 126 for Perturb-Seq (13d), and 182 TFs for CRISPRi), yielding a low overlap with the other gene set resources. Therefore, only a small fraction of the benchmark dataset was usable yielding low TF coverage. Nevertheless, I found that DoRothEA performed the best on the Perturb-seq (7d) dataset (AUROC of 0.752, Figure 3.3a) followed by D-AUCell and SCENIC with almost identical performance (AUROC of 0.629 and 0.631, respectively). metaVIPER performed just slightly better than a random model (AUROC of 0.533). Interestingly, all tools performed poorly on the Perturb-seq (13d) dataset. In the CRISPRi dataset, DoRothEA and D-AUCell performed the best with D-AUCell showing slightly better performance than DoRothEA (AUROC of 0.626 for D-AUCell and 0.608 for DoRothEA). SCENIC and metaVIPER performed slightly better than a random model. Given that I included in this analysis only shared TFs across all gene set resources, I covered only 5 and 17 distinct TFs of the Perturb-seq and CRISPRi benchmark dataset. To make better use of the benchmark dataset, I repeated the analysis without SCENIC, which resulted in a higher number of shared TFs among the gene set resources and a higher TF coverage. The higher TF coverage allowed me to investigate the performance of the tools in terms of DoRothEA's confidence level. For both Perturb-seq datasets, I found consistent results with the previous study when the TF coverage increased from 5 to 10 (Figure 3.3b). However, for the CRISPRi dataset, the performance of DoRothEA and metaVIPER remained comparable to the previous study while the performance of D-AUCell dropped remarkably. These trends can also be observed in PR-space (Supplementary Figure B.8e).

In summary, these analyses suggested that the tools DoRothEA and D-AUCell, both interrogating the manually curated, high-quality regulons from DoRothEA, are the best-performing tools to recover TF perturbation at the single-cell level of real data.

Application of TF and pathway analysis tools on samples of heterogeneous cell type populations (PBMC+HEK293T)

In my last analysis, I wanted to test the performance of all tested tools in a more heterogeneous system that would illustrate a typical scRNA-seq data analysis scenario where multiple cell types are present. I used a dataset from the Human Cell Atlas project (Regev et al., 2017) that contains scRNA-seq profiles of human peripheral blood mononuclear cells (PBMCs) and HEK 293T cell line with annotated cell types

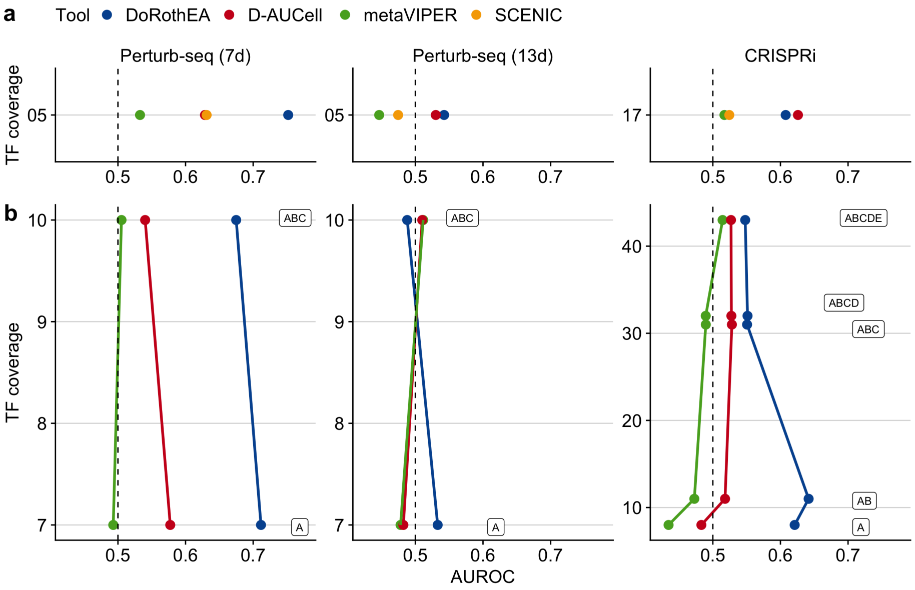


Figure 3.3: Benchmark results of TF analysis tools on real scRNA-seq data. a Performance of DoRothEA, D-AUCell, metaVIPER, and SCENIC on all sub benchmark datasets in ROC space vs TF coverage. b Performance of DoRothEA, D-AUCell, and metaVIPER on all sub benchmark datasets in ROC vs TF coverage split up by combinations of DoRothEA's confidence levels (A-E). a, b In both panels, the results for each tool are based on the same but for the respective panel different set of (shared) TFs. TF coverage reflects the number of distinct perturbed TFs in the benchmark data set that are also covered by the gene sets.

(Mereu et al., 2020). This dataset was analyzed with 13 different scRNA-seq protocols (see the "Methods" section). In this study, no ground truth (in contrast to the previous perturbation experiments) for TF and pathway activities was available. To evaluate the performance of all tools, I assessed the potential of TF and pathway activities to cluster cells from the same cell type together based on a priori annotated cell types. All pathway analysis tools and the TF analysis tools DoRothEA, D-AUCell, and metaVIPER were readily applicable to the dataset, except for SCENIC, where I first had to infer GRNs specific for each dataset (and thus experimental protocol) from the respective data (e.g., Drop-seq regulons inferred from the Drop-seq dataset; see the "Methods" section). The overlap of all protocol-specific SCENIC regulons comprised only 24 TFs (Supplementary Figure B.9a). Including regulons from DoRothEA and GTEx shrank the total overlap down to 20 (Supplementary Figure B.9b). In contrast, high-quality regulons (confidence levels A and B) from DoRothEA and GTEx alone overlapped in 113 TFs. Given the very low regulon overlap between DoRothEA, GTEx, and all protocol-specific SCENIC regulons, I decided to subset DoRothEA and GTEx to their shared TFs while using all available TFs of the protocol-specific SCENIC regulons.

The low overlap of the SCENIC regulons motivated me to investigate the direct functional consequences of their usage. Theoretically, one would expect to retrieve highly similar regulons as they were constructed from the same biological context. I calculated the pairwise (Pearson) correlations of TF activities between the scRNA-seq technologies for each tool. The distribution of correlation coefficients for each tool denotes the consistency of predicted TF activity across the protocols (Supplementary Figure B.10). The tools DoRothEA, D-AUCell, and metaVIPER had all a similar median Pearson correlation coefficient of ~0.63 and SCENIC of 0.34. This suggests that the predicted TF activities via SCENIC networks are less consistent across the protocols than the TF activities predicted via DoRothEA, D-AUCell, and metaVIPER.

To assess the clustering capacity of TF and pathway activities, I performed my analysis for each scRNA-seq technology separately to identify protocol-specific and protocol-independent trends. I assumed that the cell-type-specific information should be preserved also on the reduced dimension space of TF and pathway activities if these meaningfully capture the corresponding functional processes. Hence, I assessed how well the individual clusters correspond to the annotated cell types by a two-step approach. First, I applied UMAP on different input matrices, e.g., TF/pathway activities or gene expression, and then I evaluated how well cells from the same cell type cluster together. I considered silhouette widths as a metric of cluster purity (see the "Methods" section). Intuitively, each cell type should form a distinct cluster. However, some cell types are closely related, such as different T cells (CD4 and CD8) or monocytes (CD14+ and FCGR3A+). Thus, I decided to evaluate the cluster purity at different levels of the cell-type hierarchy from fine-grained to coarse-grained. I started with the hierarchy level 0 where every cell type forms a distinct cluster and ended with the hierarchy level 4 where all PBMC cell types and the HEK cell line form a distinct cluster (Figure 3.4a). My main findings rely on hierarchy level 2. Silhouette widths derived from a set of highly variable genes (HVGs) set the baseline for the

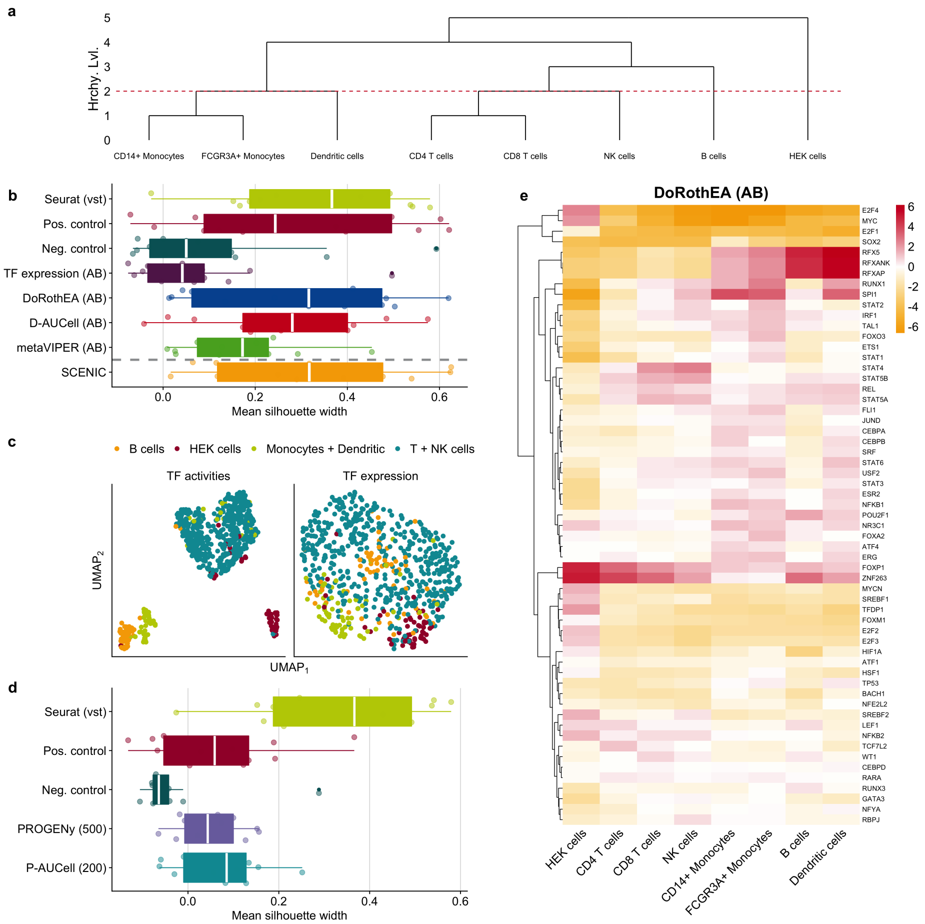


Figure 3.4: Application of TF and pathway analysis tools on a representative scRNA-seq dataset of PBMCs and HEK cells. a Dendrogram showing how cell lines/cell types are clustered together based on different hierarchy levels. The dashed line marks the hierarchy level 2, where CD4 T cells, CD8 T cells, and NK cells are aggregated into a single cluster. Similarly, CD14+ monocytes, FCGR3A+ monocytes, and dendritic cells are also aggregated to a single cluster. The B cells and HEK cells are represented by separate, pure clusters. b, d Comparison of cluster purity (clusters are defined by hierarchy level 2) between the top 2000 highly variable genes and b TF activity and TF expression and d pathway activities. The dashed line in b separates SCENIC as it is not directly comparable to the other TF analysis tools and controls due to a different number of considered TFs. c UMAP plots of TF activities calculated with DoRothEA and corresponding TF expression measured by SMART-Seq2 protocol. e Heatmap of selected TF activities inferred with DoRothEA from gene expression data generated via Quartz-Seq2.

silhouette widths derived from pathway/TF activities. I identified the top 2000 HVGs with Seurat (Butler, Hoffman, Smibert, Papalexi, & Satija, 2018) using the selection method "vst" as it worked the best in my hands at four out of five hierarchy levels (Supplementary Figure B.11). For both TF and pathway activity matrices, the number of features available for dimensionality reduction using UMAP was substantially less (113 TFs for DoRothEA/metaVIPER, up to 400 TFs for SCENIC GRNs and 14 pathways, respectively) than for a gene expression matrix containing the top 2000 HVGs. As the number of available features for dimensionality reduction is different between HVGs, TFs, and pathways, I compare the cluster purity among these input features, to a positive and negative control. The positive control is a gene expression matrix with the top n HVGs and the negative control is a gene expression matrix with randomly chosen n HVGs out of the 2000 HVGs (n equals 14 for pathway analysis and 113 for TF analysis). It should be noted that in terms of TF analysis, the positive and negative control is only applicable to DoRothEA, D-AUCell, and metaVIPER as they share the same number of features. As the protocol-specific SCENIC GRNs differ in size (Supplementary Figure B.9a), each network would require its own positive and negative control.

To evaluate the performance of the TF activity inference methods and the utility of TF activity scores, I determined the cluster purity derived from TF activities predicted by DoRothEA, D-AUCell, metaVIPER, and SCENIC, TF expression, and positive and negative controls. scRNA-seq protocols and input matrices used for dimensionality reduction affected cluster purity significantly (two-way ANOVA p values <2.2e-16 and 4.32e-12, respectively, p values and estimations for corresponding linear model coefficients in Supplementary Figure B.12a; see the "Methods" section). The cluster purity based on TF activities inferred using DoRothEA and D-AUCell did not differ significantly (Figure 3.4b, corresponding plots for all hierarchy levels in Supplementary Figure B.12b). In addition, the cluster purity of both tools was not significantly worse than the purity based on all 2000 HVGs, though I observed a slight trend indicating a better cluster purity based on HVGs. This trend is expected due to the large difference in available features for dimensionality reduction. Instead, a comparison to the positive and negative controls is more appropriate. Both DoRothEA and D-AUCell performed comparably to the positive control but significantly better than the negative control across all scRNA-seq protocols (TukeyHSD post-hoc-test, adj. p value of 1.26e-4 for DoRothEA and 7.09e-4 for D-AUCell). The cluster purity derived from metaVIPER was significantly worse than for DoRothEA (TukeyHSD post-hoc-test, adj. p value of 0.054) and tend to be worse than D-AUCell (TukeyHSD post-hoc-test, adj. p value of 0.163) as well. metaVIPER was not significantly better than the negative control. The cluster purity from SCENIC was significantly better than the negative control (TukeyHSD post-hoc-test, adj. p value of 1.11e-6) and comparable to the positive control and thus to DoRothEA and D-AUCell. However, as mentioned above, SCENIC is only partially comparable to the controls and other tools due to the different number of TFs.

Regardless of the underlying TF activity tool, except for metaVIPER, the cluster purity derived from TF activities outperformed significantly the purity derived from

TF expression (TukeyHSD post-hoc-test, adj. p value of 5.89e-6 for DoRothEA, 3.85-e5 for D-AUCell, and 4.0e-8 for SCENIC). This underlines the advantage and relevance of using TF activities over the expression of the TF itself (Figure 3.4c). With a comparable performance to a similar number of HVG and also to 2000 HVGs, I concluded that TF activities serve—independently of the underlying scRNA-seq protocol—as a complementary approach for cluster analysis that is based on generally more interpretable cell type marker.

To evaluate the performance of pathway inference methods and the utility of pathway activity scores, I determined cluster purity with pathway matrices generated by different PROGENy versions and P-AUCell. I used 200 and 500 footprint genes per pathway for PROGENy and P-AUCell, respectively, since they provided the best performance in the previous analyses. As observed already for the TF analysis tools, scRNA-seq protocols and matrices used for dimensionality reduction affected cluster purity significantly (two-way ANOVA p values of 2.84e-7 and 1.13e-13, respectively, p values and estimations for corresponding linear model coefficients in Supplementary Figure B.13a; see the "Methods" section). The cluster purity derived from pathway activity matrices is not significantly different between PROGENy and P-AUCell, while worse than all HVGs (TukeyHSD post-hoc-test, adj. p value of 4.07e-10 for PROGENy and 4.59e-9 for P-AUCell, Figure 3.4d, corresponding plots for all hierarchy levels in Supplementary Figure B.13b). This is expected due to the large difference in the number of available features for dimensionality reduction (2000 HVGs vs 14 pathways). The cluster purity of both approaches was comparable to the positive control but significantly better than the negative control (TukeyHSD post-hoc-test, adj. p value of 0.077 for PROGENy and 0.013 for P-AUCell vs negative control). In summary, this study indicated that the pathway activities contain relevant and cell-type-specific information, even though they do not capture enough functional differences to be used for effective clustering analysis. Overall, the cluster purity of cells represented by the estimated pathway activities is worse than the cluster purity of cells represented by the estimated TF activities.

In addition, I observed that TF and pathway matrices derived from Quartz-Seq2 protocol yielded for hierarchy level 2 in significantly better cluster purity than all other protocols, which is in agreement with the original study of the PBMC + HEK293T data (Supplementary Figures B.12a and B.13a) (Mereu et al., 2020).

TF and pathway activity scores are more interpretable than the expression of single genes. Hence, I were interested to explore whether I could recover known cell-type-specific TF and pathway activities from the PBMC data. I decided to focus on the dataset measured with Quartz-Seq2 as this protocol showed in my and in the original study superior performance over all other protocols (Mereu et al., 2020). I calculated mean TF and pathway activity scores for each cell type using DoRothEA, D-AUCell, metaVIPER, and SCENIC (using only TFs with confidence levels A and B, Figure 3.4e and Supplementary Figure B.14a–c, respectively), PROGENy with 500 and P-AUCell with 200 footprint genes per pathway (Supplementary Figure B.14d and e). In terms of TF activities, I observed high RFXAP, RFXANK, and RFX5 activity (TFs responsible for MHCII expression) in monocytes, dendritic cells, and B cells (the main antigen-presenting cells of the investigated population (Burd et al., 2004)) (Supplementary Figure B.14a and b). Myeloid lineage-specific SPI1 activity (Zakrzewska et al., 2010) was observed in monocytes and dendritic cells. The high activity of repressor TF (where regulation directionality is important) FOXP1 in T lymphocytes (Feng et al., 2011) was only revealed by DoRothEA. Proliferative TFs like Myc and E2F4 had also high activity in HEK cells.

Regarding pathway activities, I observed across both methods, in agreement with the literature, high activity of NFkB and TNFa in monocytes (T. Liu, Zhang, Joo, & Sun, 2017) and elevated Trail pathway activity in B cells (Supplementary Figure B.14d and e) (Staniek et al., 2019). HEK cells, as expected from dividing cell lines, had higher activity of proliferative pathways (MAPK, EGFR, and PI3K, Supplementary Figure B.14d). These later pathway activity changes were only detected by PROGENy but not with AUCell, highlighting the importance of directionality information.

In summary, the analysis of this mixture sample demonstrated that summarizing gene expression into TF activities can preserve cell type-specific information while drastically reducing the number of features. Hence, TF activities could be considered as an alternative to gene expression for clustering analysis.

I also showed that pathway activity matrices contain cell-type-specific information, too, although I do not recommend using them for clustering analysis as the number of features is too low. In addition, I recovered known pathway/TF cell-type associations showing the importance of directionality and supporting the utility and power of the functional analysis tools DoRothEA and PROGENy.

3.5 Discussion

In this paper, I tested the robustness and applicability of functional analysis tools on scRNA-seq data. I included both bulk- and single-cell-based tools that estimate either TF or pathway activities from gene expression data and for which well-defined benchmark data exist. The bulk-based tools were DoRothEA, PROGENy, and GO gene sets analyzed with GSEA (GO-GSEA). The functional analysis tools specifically designed for the application in single cells were SCENIC, AUCell combined with DoRothEA (D-AUCell) and PROGENy (P-AUCell) gene sets, and metaVIPER.

I first explored the effect of low gene coverage in bulk data on the performance of the bulk-based tools DoRothEA, PROGENy, and GO-GSEA. I found that for all tools the performance dropped with decreasing gene coverage but at a different rate. While PROGENy was robust down to 500 covered genes, DoRothEA's performance dropped markedly at 2000 covered genes. In addition, the results related to PROGENy suggested that increasing the number of footprint genes per pathway counteracted low gene coverage. GO-GSEA showed the strongest drop and did not perform better than a random guess below 2000 covered genes. Comparing the global performance across all pathways of both pathway analysis tools suggests that footprint-based gene sets are superior over gene sets containing pathway members (e.g., GO gene sets) in recovering perturbed pathways. This observation is in agreement with previous studies conducted by me and others (Parikh, Klinger, Xia, Marto, & Blüthgen, 2010; Schubert et al., 2018). However, both PROGENy and GO-GSEA performed poorly for some pathways, e.g., WNT pathway. I reason that this observation might be due to the quality of the corresponding benchmark data (Christian H. Holland, Szalai, & Saez-Rodriguez, 2020). Given this fact and that GO-GSEA cannot handle low gene coverage (in my hands), I concluded that this approach is not suitable for scRNA-seq analysis. Hence, I decided to focus only on PROGENy as bulk-based pathway analysis tool for the following analyses.

Afterward, I benchmarked DoRothEA, PROGENy, D-AUCell, P-AUCell, and metaVIPER on simulated single cells that I sampled from bulk pathway/TF perturbation samples. I showed that my simulated single cells possess characteristics comparable to real single-cell data, supporting the relevance of this strategy. Different combinations of simulation parameters can be related to different scRNA-seq technologies. For each combination, I provide a recommendation of how to use DoRothEA's and PROGENy's gene sets (in terms of confidence level combination or number of footprint genes per pathway) to yield the best performance. It should be noted that my simulation approach, as it is now, allows only the simulation of a homogenous cell population. This would correspond to a single cell experiment where the transcriptome of a cell line is profiled. In future work, this simulation strategy could be adapted to account for a heterogeneous dataset that would resemble more realistic single-cell datasets (Peng, Zhu, Yin, & Tan, 2019; Zappia, Phipson, & Oshlack, 2017).

In terms of TF activity inference, DoRothEA performed best on the simulated single cells followed by D-AUCell and then metaVIPER. Both DoRothEA and D-AUCell shared DoRothEA's gene set collection but applied different statistics. Thus, I concluded that, in my data, VIPER is more suitable to analyze scRNA-seq data than AUCell. The tool metaVIPER performed only slightly better than a random model, and since it uses VIPER like DoRothEA, the weak performance must be caused by the selection of the gene set resource. DoRothEA's gene sets/TF regulons were constructed by integrating different types of evidence spanning from literature curated to predicted TF-target interactions. For metaVIPER, I used 27 tissue-specific GRNs constructed in a data-driven manner with ARACNe (Margolin et al., 2006) thus containing only predicted TF-target interactions. The finding that especially the high-confidence TF regulons from DoRothEA outperform pure ARACNe regulons is in agreement with previous observations (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019; Keenan et al., 2019) and emphasizes the importance of combining literature curated resources with in silico predicted resources. Moreover, I hypothesize based on the pairwise comparison that for functional analysis, the choice of gene sets is of higher relevance than the choice of the underlying statistical method.

As one could expect, the single-cell tools D-AUCell metaVIPER performed better on single cells than on the original bulk samples. This trend becomes more pronounced with increasing library size and number of cells. However, the bulk-based tools performed even better on the simulated single cells than the scRNA specific tools.

Related to pathway analysis, both PROGENy and P-AUCell performed well on the simulated single cells. The original framework of PROGENy uses a linear model that incorporates individual weights of the footprint genes, denoting the importance and also the sign of the contribution (positive/negative) to the pathway activity score. Those weights cannot be considered when applying AUCell with PROGENy gene sets. The slightly higher performance of PROGENy suggests that individual weights assigned to gene set members can improve the activity estimation of biological processes.

Subsequently, I aimed to validate the functional analysis tools on real single-cell data. While I could not find suitable benchmark data of pathway perturbations, I exploited two independent datasets of TF perturbations to benchmark the TF analysis tools which I extended with SCENIC. These datasets combined CRISPR-mediated TF knock-out/knock-down (Perturb-Seq and CRISPRi) with scRNA-seq. It should be noted that pooled screenings of gene knock-outs with Perturb-seq suffer from an often faulty assignment of guide-RNA and single-cell (Hegde, Strand, Hanna, & Doench, 2018). Those mislabeled data confound the benchmark as the ground-truth is not reliable. In addition, my definition of true-positives and true-negatives is commonly used for such analyses (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019; Keenan et al., 2019; Nguyen, Shafi, Nguyen, & Draghici, 2019), but it might be incorrect due to indirect and compensatory mechanisms (Smits et al., 2019). These phenomena can confound the results of this type of benchmarks.

Nevertheless, I showed that DoRothEA's gene sets were globally effective in inferring TF activity from single-cell data with varying performance dependent on the used statistical method. As already shown in the in silico benchmark, D-AUCell showed a weaker performance than DoRothEA, supporting that VIPER performs better than AUCell. Interestingly, metaVIPER's performance was no better than random across all datasets. metaVIPER used the same statistical method as DoRothEA but different gene set resources. This further supports my hypothesis that the selection of gene sets is more important than the statistical method for functional analysis. This trend is also apparent when comparing the performance of SCENIC and D-AUCell as both rely on the statistical method AUCell but differ in their gene set resource. SCENICs' performance was consistently weaker than D-AUCell. In addition, I found that the gene regulatory networks inferred with the SCENIC workflow covered only a limited number of TFs in comparison to the relatively comprehensive regulons from DoRothEA or GTEx.

Furthermore, the perturbation time had a profound effect on the performance of the tools: while DoRothEA and D-AUCell worked well for a perturbation duration of 6 (CRISPRi) and 7 days (Perturb-Seq (7d)), the performance dropped markedly for 13 days. I reasoned that, within 13 days of perturbation, compensation effects are taking place at the molecular level that confound the prediction of TF activities. In addition, it is possible that cells without a gene edit outgrow cells with a successful knock-out after 13 days as the knock-out typically yield in a lower fitness and thus proliferation rate.

In summary, DoRothEA subsetted to confidence levels A and B performed the best on real scRNA-seq data but at the cost of the TF coverage. The results of the in silico and in vitro benchmark are in agreement. Accordingly, I believe that it is reasonable to assume that also PROGENy works on real data given the positive benchmark results on simulated data.

Finally, I applied the tools of interest to a mixture sample of PBMCs and HEK cells profiled with 13 different scRNA-seq protocols. I investigated to which extent pathway and TF matrices retain cell-type-specific information, by evaluating how well cells belonging to the same cell type or cell type family cluster together in reduced dimensionality space. Given the lower numbers of features available for dimensionality reduction using TF and pathway activities, cell types could be recovered equally well as when using the same number of the top highly variable genes. In addition, I showed that cell types could be recovered more precisely using TF activities than TF expression, which is in agreement with previous studies [ding_2018]. This suggests that summarizing gene expression as TF and pathway activities can lead to noise filtering, particularly relevant for scRNA-seq data, though TF activities performed better than pathway activities which is again attributed to the even lower number of pathways. Specifically, TF activities computed with DoRothEA, D-AUCell, and SCENIC yielded a reasonable cluster purity. It should be noted that, while DoRothEA and D-AUCell rely on independent regulons, the SCENIC networks are constructed from the same dataset they are applied to. This poses the risk of overfitting.

My analysis suggested at different points that the performance of TF and pathway analysis tools is more sensitive to the selection of gene sets than the statistical methods. In particular, manually curated footprint gene sets seem to perform generally better. This hypothesis could be tested in the future by decoupling functional analysis tools into gene sets and statistics. Benchmarking all possible combinations of gene sets and statistics (i.e., DoRothEA gene sets with a linear model or PROGENy gene sets with VIPER) would shed light on this question which I believe is of high relevance for the community.

3.6 Conclusions

my systematic and comprehensive benchmark study suggests that functional analysis tools that rely on manually curated footprint gene sets are effective in inferring TF and pathway activity from scRNA-seq data, partially outperforming tools specifically designed for scRNA-seq analysis. In particular, the performance of DoRothEA and PROGENy was consistently better than all other tools. I showed the limits of both tools with respect to low gene coverage. I also provided recommendations on how to use DoRothEA's and PROGENy's gene sets in the best way dependent on the number of cells, reflecting the amount of available information, and sequencing depths. Furthermore, I showed that TF and pathway activities are rich in cell-type-specific information with a reduced amount of noise and provide an intuitive way of interpretation and hypothesis generation. I provide my benchmark data and code to the community for further assessment of methods for functional analysis.

3.7 Availability of data and materials

The code to perform all presented studies is written in R (Gentleman et al., 2004; R Core Team, 2020; Wickham, 2016) and is freely available on GitHub. The datasets supporting the conclusions of this article are available at Zenodo.

Chapter 4

Transcriptomic cross-species analysis of chronic liver disease reveals consistent regulation between humans and mice

4.1 Preface

The text of the following chapter is based on an early draft of the later published manuscript “Transcriptomic cross-species analysis of chronic liver disease reveals consistent regulation between humans and mice” (Christian H. Holland et al., 2021) that was originally written by myself. Unless otherwise stated I performed all bioinformatics-related analyses myself.

4.2 Background

In recent decades the number of patients suffering from chronic liver disease (CLD) has increased (Younossi et al., 2020). To study the development and progression of CLD, with the ultimate aim to identify therapeutic targets and test drug candidates mouse models are often used. However, their use for translational research has been discussed controversially due to known large interspecific differences (Leist & Hartung, 2013). An earlier comparison of gene expression in mouse models of non-alcoholic fatty liver disease (NAFLD) and human liver tissue reported large interspecies differences, with very little overlap in expression changes in mice and humans (Teufel et al., 2016). This finding questions whether experiments in mice allow conclusions about the pathophysiology of CLD in humans. To shed light on this question I revisited the changes in gene expression in liver disease mouse models and compared them to patient cohorts of CLD.

Previous comparisons between mice and humans were limited by exposure of mice to harmful substances and typically relied on exposure periods of only a few months or even weeks (Campos et al., 2020; Teufel et al., 2016). This may seem appropriate because a 1-year-old mouse is comparable to a 40-year-old, extrapolated from the relative lifespans of humans and mice. However, it remains to be seen whether such assumptions is justified, i.e., whether the disease progresses faster in mice, or whether comparable time frames of damage are necessary to obtain similar phenotypes. To answer this question, I analyzed gene expression data from a mouse model that was treated with CCl_4 for up to a year and compared it to a collection of human CLD expression data. Furthermore, I characterized changes in gene expression in mice with either chronic or acute liver injury. This was justified by a recent study that suggested that the expression changes observed in mouse models of chronic and acute damage are similar to those observed in human CLD (Campos et al., 2020). The liver of mice responded to acute injuries by simultaneously upregulating inflammation- and downregulating metabolism-associated gene regulatory networks, both controlled by a common upstream master regulator. Chronic mouse models and even patients with CLD responded similarly (Campos et al., 2020). To define the therapeutic window, it may be important to distinguish between genes altered exclusively in chronic insult and genes altered during both chronic and acute injury. If a gene that is altered exclusively in chronic damage and is relevant to the progression of the disease, it is recommended to focus therapeutically on advanced stages when its expression increases; in contrast, genes altered under both chronic and acute conditions may be inhaled earlier in the course of the disease.

In the present study, bioinformatics analysis confirmed the previously reported large interspecies differences, but also revealed substantial sets of genes that respond similarly in both species. Interestingly, the categories of genes that were exclusively altered after acute or chronic injury compared to those conserved in both damage scenarios differed in their similarity to human CLD. Genes conserved in both the acute and chronic settings showed higher similarity to the human situation, and importantly, the similarity to human CLD increased when mice were exposed for longer periods of up to one year. These findings were compiled into a data resource linking expression profile alterations of individual genes in CLD, their differential regulation in mice and humans, and categorization into acute, chronic, or conserved response sets. This resource is accessible via an online application to facilitate the intuitive exploration of the role of genes in human and mouse CLD.

4.3 Materials and Methods

A detailed description is available in Supplementary Material and Methods (Section C.1).

Mouse models and human datasets

The present analysis included genome-wide transcriptome data from seven mouse models (one with chronic and six with acute liver damage) comprising 227 mice (Table

4.1; Supplementary Figure C.1A) and five studies of human CLD with a total of 372 patients (Table 4.1; Supplementary Figure C.1B). The analyzed datasets were either generated by the group of Jan Hengstler (all mouse models except tunicamycin) or downloaded from public sources (acute tunicamycin model and all human datasets). Additionally, I analyzed nine published sets of differentially expressed genes of CLD mouse models, for which the corresponding raw data was not available (Teufel et al., 2016).

For the mouse models generated by the group of Jan Hengstler, a detailed description of treatment protocols, collection and processing of liver tissue, histopathology, RNA isolation, RNA-sequencing, or Affymetrix gene array analysis, and immunostaining were provided by them and is given in the supplement. In brief, for induction of chronic liver injury, 8-10 weeks-old male C57BL/6N mice were injected with CCl_4 (1 g/kg b.w.; i.p. in olive oil) twice a week for 2, 6, and 12 months (Ghallab, Myllys, et al., 2019). For acute APAP intoxication, the mice received a single dose of 300 mg/kg i.p. in warm PBS (Ghallab, Myllys, et al., 2019). For acute intoxication with CCl_4, a single dose (1.6 g/kg, i.p.) was administered (Ghallab, Myllys, et al., 2019). For acute intoxication with LPS, a single dose of 750 µg/kg was intraperitoneally injected (Godoy et al., 2016). Partial hepatectomy (PH) and bile duct ligation (BDL) were performed as previously published (Ghallab, Hofmann, et al., 2019; Mitchell & Willenbring, 2008).

Biopsies from patients with primary sclerosing cholangitis (PSC) and alcoholic liver disease were used for validation by immunostaining performed by Ahmed Ghallab.

Table 4.1: Mouse models and patient cohorts with genome-wide expression data of liver tissue.

Organism	Damage	Treatment	N	Accession ID	Reference
Mouse	Chronic	CCl4 (Up to 12 months)	36	GSE167216	Ghallab, Myllys, et al. (2019) and present study
Mouse	Acute	APAP (Up to 16 days)	49	GSE167032	Present study
Mouse	Acute	CCl4 (Up to 16 days)	46	GSE167033	Campos et al. (2020) and present study
Mouse	Acute	PH (Up to 3 months)	52	GSE167034	Present study
Mouse	Acute	BDL (Up to 3 weeks)	29	GSE166867	Ghallab, Hofmann, et al. (2019) and present study
Mouse	Acute	LPS (24 hours)	8	GSE166488	Godoy et al. (2016)
Mouse	Acute	Tunicamycin (6 hours)	7	GSE29929	Teske et al. (2011)
Human	Chronic	Mild vs advanced NAFLD	72	GSE49541	Moylan et al. (2014)

Organism	Damage	Treatment	N	Accession ID	Reference
Human	Chronic	Full-spectrum of NAFLD	78	GSE130970	Hoang et al. (2019)
Human	Chronic	NAFLD and NASH	46	GSE48452	Ahrens et al. (2013)
Human	Chronic	NASH, NAFLD, PBC and PSC	109	GSE61260	Horvath et al. (2014)
Human	Chronic	HCV and NAFLD	67	E-MTAB-6863	Ramnath et al. (2018)

Processing and analysis of transcriptomic data

Raw data of publicly available transcriptome studies were downloaded from Gene Expression Omnibus or ArrayExpress. Microarray and RNA-sequencing data were processed and normalized with the R/Bioconductor packages oligo, limma, and edgeR. FASTQ files from the chronic CCl_4 study were aligned using the web application Biojupies (Torre, Lachmann, & Ma'ayan, 2018). Differential gene expression analysis (DGEA) between two conditions was performed using the R/Bioconductor package limma. A gene was considered differentially expressed with a false discovery rate (FDR) ≤ 0.05 and an absolute log-fold change (logFC) ≥ 1. The result of a DGEA is referred to as a (gene) signature.

Time-series clustering

Time-series gene expression data were clustered with the software program STEM (Short Time-series Expression Miner, version 1.3.1216; Ernst, Nau, & Bar-Joseph (2005); Ernst & Bar-Joseph (2006)) using logFC from the preceding DGEA.

Comparison of gene set similarity

The similarity or overlap between two gene sets was summarized either as Jaccard Index or Overlap Coefficient. Unless otherwise stated, the gene sets were composed of the top 500 differentially expressed genes based on limma's moderated t-value.

Testing direction of regulation with Gene Set Enrichment Analysis

To test whether the differentially expressed genes of a specific study are consistently regulated in an independent study, Gene Set Enrichment Analysis (GSEA) was performed. The gene sets comprised the top 500 up- and downregulated genes extracted based on the moderated t-statistic. GSEA was performed with the R/Bioconductor package fgsea (version 1.14.0; Sergushichev (2016)).

Construction and ranking of exclusive chronic, exclusive acute, and common gene sets

First, the union of all differentially expressed genes (FDR $\leq$ 1e-4 and $|\text{logFC}| \geq 1$) was built across all the signatures derived from the chronic or acute mouse experiments. Subsequently, genes were classified as exclusively chronic, exclusively acute, or commonly regulated in chronic and acute. For each of these three different gene set classes, a custom metric was computed to rank genes within these sets, e.g. the metric for the exclusive chronic gene set prioritized genes that were deregulated in chronic but not in acute signatures. Similarly, the metric for exclusive acute genes highly ranked genes that were deregulated in acute but not in chronic signatures. Finally, the metric for the common genes ranked genes as 'high' that were altered in both chronic and acute signatures.

Functional characterization of transcriptomic data with various gene set resources

Transcriptomic data were functionally characterized using biological processes from Gene Ontology (GO), DoRothEA's regulons (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019; Christian H. Holland, Szalai, & Saez-Rodriguez, 2020), and PROGENy's pathway responsive-genes (Christian H. Holland, Szalai, & Saez-Rodriguez, 2020; Schubert et al., 2018), applying overrepresentation analysis. The number of background genes was set to 20,000 and the minimal gene set size to 10.

Identification of consistently deregulated genes in patients and chronic CCl_4 mouse model

To identify consistently deregulated genes in patients of CLD and the chronic CCl_4 mouse model, the top 500 up- and downregulated genes, respectively, were extracted from each human signature. Subsequently, those gene sets were enriched in the signatures from the chronic CCl_4 mouse model by using the R/Bioconductor package fgsea (version 1.14.0; Sergushichev (2016)) with 1000 permutations. Afterward, the leading-edge genes were extracted from each enrichment if they met two criteria: i) FDR $\leq$ 0.05, and ii) consistent regulation between human and mouse data. Leading-edge genes that appeared in at least three studies per chronic time point were considered consistently deregulated in humans and mice.

4.4 Results

Chronic liver damage in mice

Ahmed Ghallab performed a study of chronic liver damage in mice with twice-weekly administrations of the hepatotoxic compound CCl_4 for up to twelve months. Six animals per group were analyzed by RNA-sequencing after two, six, and twelve months (Figure 4.1A). Histological analyses showed progressive fibrogenesis particularly between months 6 and 12 accompanied by increased transaminase enzyme activities in blood (Figure 4.1B and C). Principal component analysis (PCA) showed that mice from the individual treatment groups clustered together; chronic CCl_4 intoxication led to a shift inversely along with principal component 1 (PC1), while the solvent oil caused a shift along PC2 (Supplementary Figure C.2A). The number of differentially expressed genes increased particularly between months six and twelve (Supplementary Figure C.2B and C). Overlap analysis of the differential genes showed that most genes deregulated at an earlier time were also altered later and additional differential genes occurred (Figure 4.1D). Among the genes in the overlap of all three exposure periods with the highest fold-changes are the extracellular matrix protein Col28a1, two sulfotransferase isoforms (Sult2a1 and 2a2), the basement membrane glycoprotein Tinag, the positive regulator of PP1 phosphatase Ppp1r42 that plays a role in centrosome separation (up); three members of the lipocalin family (Mup12, 15 and 19); the DBH-like monooxygenase protein 1, Moxd1 (down) (Figure 4.1E). Time-dependent clustering resulted in 7 clusters, four clusters with up-, two with down, and one with initially up and later downregulated genes (Figure 4.1F). Upregulated clusters were enriched in inflammation and proliferation-associated genes, with Lyl1 and Maf as the most overrepresented transcription factors and TNFa as well as NFkB as the most overrepresented pathways. The dominant GO terms of the downregulated clusters were all metabolism-associated with HNF1a and HNF4a as most significantly overrepresented TFs. The initially up- and later downregulated cluster contained mostly extracellular matrix-associated genes.

In summary, transcriptomics in agreement with histological analyses consistently supports progressive inflammation and fibrosis with a relatively mild phenotype until month 6 in contrast to massive progression between months 6 and 12.

Acute liver damage in mice

Next, Ahmed Ghallab and I studied time-resolved acute liver damage after single doses of CCl_4 and APAP, as well as partial hepatectomy (PH), and common bile duct ligation (BDL), and single time points after lipopolysaccharide (LPS) and tunicamycin (Table 4.1). While I was responsible for the bioinformatics analysis Ahmed Ghallab performed imaging and chemistry analyses. As an example, the APAP model is presented in Figure 4.2, corresponding summaries of the other mouse are shown in Supplementary Figures C.4-C.9.

To induce acute liver injury the mice were treated once with a hepatotoxic but not lethal dose of 300 mg/kg body weight APAP and the transcriptome was profiled at 9 time points after injection spanning from 1 hour to 16 days (Figure 4.2A). Histological analyses showed pericentral necrotic tissue on day 1 with almost complete regeneration until day 8 without the formation of fibrosis (Figure 4.2B). Infiltration of CD45 positive immune cells was observed between days 1 and 4. Clinical chemistry showed a transient increase in liver enzymes (Figure 4.2C). Thus, the histological alterations correspond

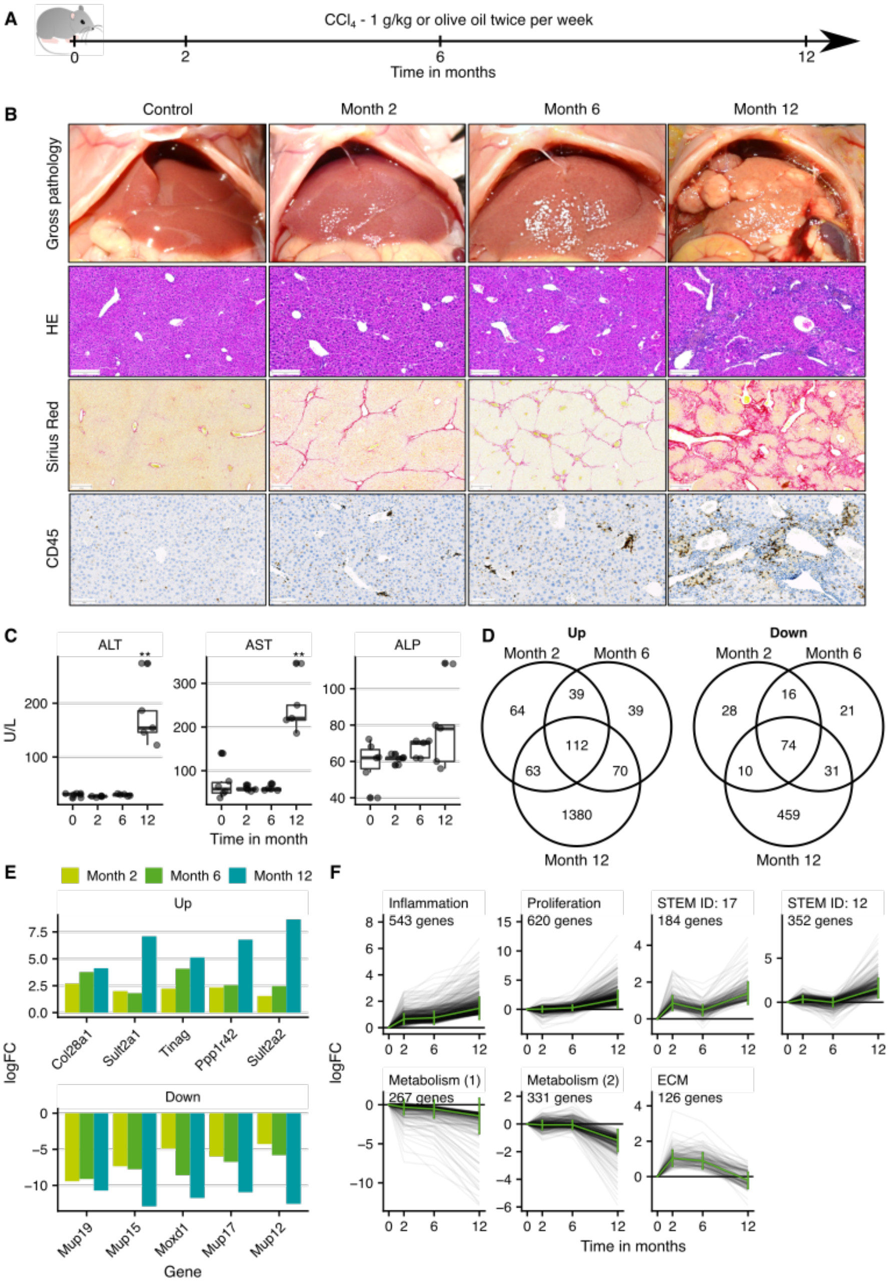


Figure 4.1: Gene expression changes in the CCl 4 mouse model of CLD. A. Experimental design. Six mice were analyzed in each treatment group. B. Histological analyses with hematoxylin and eosin (HE) staining, visualization of fibrosis by Sirius red, and infiltration of immune cells by CD45; scale bars: 200 µm (HE; Sirius red) and 100 µm (CD45). C. Clinical chemistry of alanine transaminase (ALT), aspartate transaminase (AST) and alkaline phosphatase (ALP) activities in plasma. D. Overlaps of up- and downregulated genes. E. Genes in the overlap of the three exposure periods with the highest fold changes. F. Time-resolved clustering of deregulated genes with the dominant GO terms or the default profile names (STEM ID), if no significantly overrepresented GO term was obtained. The panels B and C were provided by Ahmed Ghallab.

well to previous studies of APAP intoxication in mice (Sezgin et al., 2018). In the PCA space, differences to the controls were largest between 12 hours and day 2 and subsequently returned towards the control levels that is also reflected in the number of differentially expressed genes (Figure 4.2D and E; Supplementary Figure C.3A and B). The strongest upregulated genes were the chaperone Hspa1a (12 hours), Chil3, a protein secreted by macrophages that is involved in inflammatory processes (day 1), and calcium-binding protein S100a6 that is involved in the response to different types of cell stress (day 2). Among the most downregulated genes were numerous cytochrome P450 enzymes, including Cyp7a1, a key enzyme in bile acid synthesis, numerous further metabolic enzymes such as Acot1 that plays a role in fatty acid metabolism, and also Inhbe, a member of the TGF-beta superfamily (Figure 4.2F). Clustering of gene expression trajectories resulted in seven clusters, three with up-and four with downregulated genes (Figure 4.2G). Upregulated clusters were enriched in stress response, migration, and proliferation-associated genes with Atf3, Sp1, and E2F4 as enriched transcription factors and TGF-beta as the most enriched pathway. Genes of the downregulated clusters were predominantly metabolism-associated with HNF4a and Cebpa as significantly overrepresented TFs. A common feature of the studies with acute time series is that the maximal number of deregulated genes was reached at 24 or 48 hours after the intervention and returned completely or almost completely to control levels within 16 days. An exception was BDL, where expression changes persisted due to the irreversible obstruction of the bile duct. To investigate the consistency of the gene signature across the six acute mouse models I first set out to compare the similarity of the top 500 differentially expressed genes. For mouse models with several time points and a reversible phenotype, the time point with the strongest deregulated expression profile was used (Supplementary Figure C.10), while the 24 hour time point was selected for the BDL model.

Globally, I found a low gene overlap across the six models (mean Jaccard Index of 0.058; Figure 4.3A), with the highest similarity between APAP and CCl_4 (Jaccard Index of 0.157) and the lowest between LPS and tunicamycin (Jaccard Index of 0.012). This pairwise comparison revealed that each treatment yields a distinctive set of top differentially expressed genes which at first glance could be interpreted as inconsistency across them. Next, I tested whether the direction of regulation of the top differentially expressed genes is conserved within the six acute mouse models. For this purpose, I performed a mutual enrichment of the top 500 up-and downregulated genes in the acute contrasts. This analysis revealed that the different sets of upregulated genes were coordinately upregulated in all other acute contrasts (Figure 4.3B). The same applied to the sets of downregulated genes. As the only outlier, the top 500 upregulated genes after LPS treatment tended to be enriched among the downregulated genes of the tunicamycin model (FDR=0.052). These systematic comparisons showed that although the top differentially expressed genes were distinct the direction of regulation was consistent.

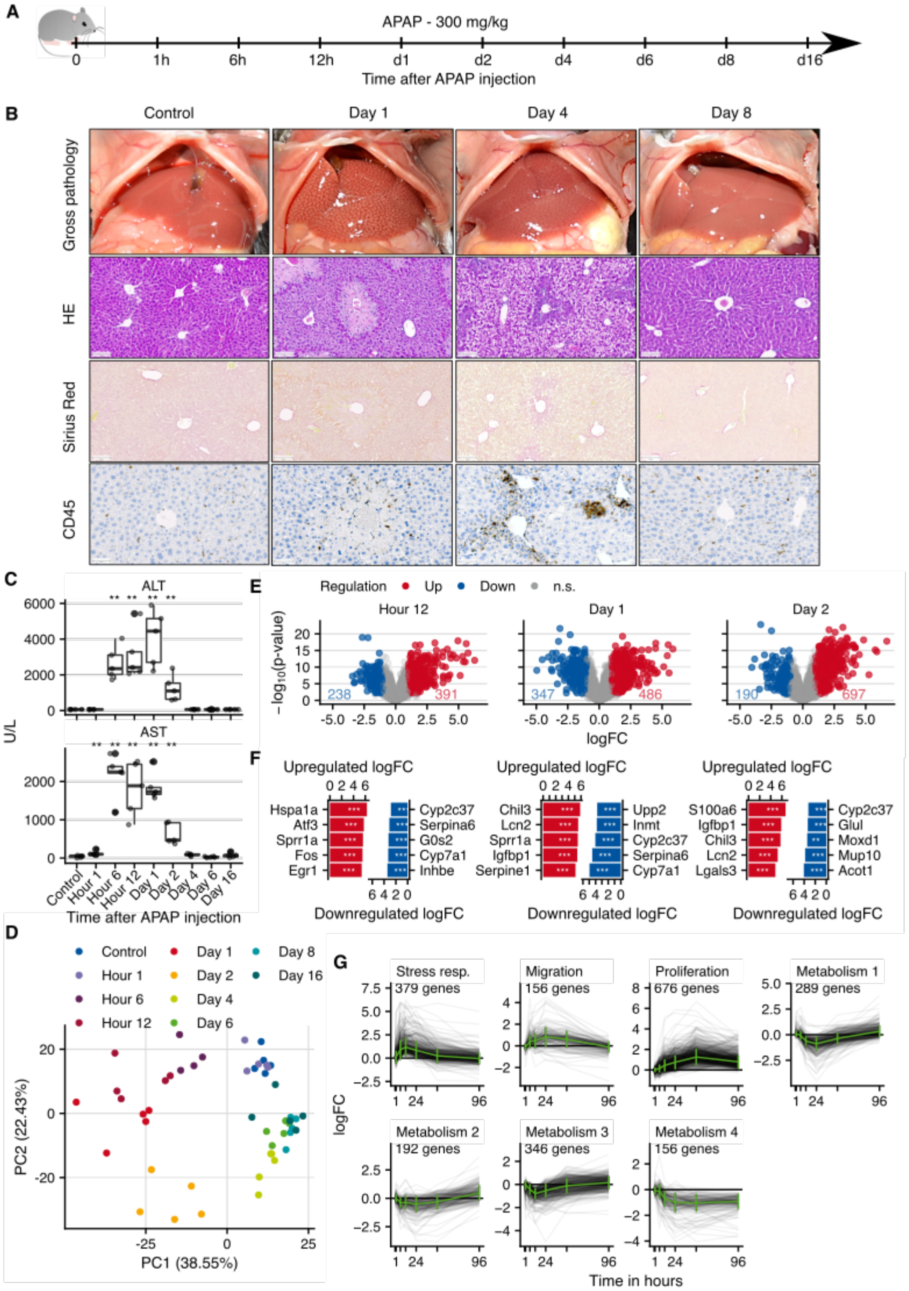


Figure 4.2: Gene expression changes in a mouse model of acute liver damage induced by administration of 300 mg/kg b.w. acetaminophen (APAP). A. Experimental design. Five mice were analyzed in each treatment group. B. Histological analyses with hematoxylin and eosin (HE) staining, lack of fibrosis visualized by Sirius red, and infiltration of immune cells by CD45 immunostaining; scale bars: 100 µm (HE; Sirius red) and 50 µm (CD45). C. Clinical chemistry of alanine transaminase (ALT) and aspartate transaminase (AST) activity in plasma. D. PCA analysis of global expression changes. E. Volcano plots of gene expression changes at 12 hours, days 1 and 2 after APAP administration. F. Genes with the highest logFCs. G. Time-resolved clustering of deregulated genes. The panels B and C were provided by Ahmed Ghallab.

Exclusively and commonly regulated genes of chronic and acute liver damage in mice

To identify a set of exclusive chronic, exclusive acute, and common acute and chronic genes, I integrated and analyzed all available acute and chronic time points. I constructed a pool of chronic and acute genes by taking the union of all differentially expressed genes from the chronic and acute mouse models. The unified genes of the acute and also chronic studies showed consistent up- or downregulation across the individual contrasts, respectively (Supplementary Figure C.11). Set comparisons between the pools of unified acute and chronic genes revealed 834 exclusive chronic, 2777 exclusive acute, and 586 common genes (Figure 4.3C). To identify the top 100 genes in each category, I developed a custom metric that ranks the genes based on their expression in the chronic and selected acute contrasts.

In the top 100 exclusive chronic genes, 97 genes were up-and only 3 genes were downregulated (Figure 4.3D). Subsequently, I functionally characterized all up and downregulated exclusive chronic genes, by overrepresentation analysis. As gene set resources I used DoRothEA's transcription factor regulons (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019; Christian H. Holland, Szalai, & Saez-Rodriguez, 2020), PROGENy's pathway footprints (Christian H. Holland, Szalai, & Saez-Rodriguez, 2020; Schubert et al., 2018), and GO-terms of biological processes. I found the target genes from the transcription factors Hif1a and Klf as well as the footprint genes from the TGF-beta pathway overrepresented in the set of upregulated exclusive chronic genes (Figure 4.3E and F). Text analysis of overrepresented GO terms revealed that most processes were associated with development and morphogenesis (Figure 4.3G). By manual classification I found 38 GO terms related to "Development and Morphogenesis" and 24 GO terms related to "Migration," while the latter was more pronounced among the most significantly overrepresented GO terms. Functional characterization of the downregulated exclusive chronic genes highlighted the TFs Stat3, Sox2, and Hoxb13. Pathways as well as GO-terms, however, did not result in any significant associations.

For the exclusive acute gene set, I also extracted the top 100 exclusive acute genes (Supplementary Figure C.12A). The upregulated genes were associated with the TFs Myc, Trp53, and the pathways MAPK, EGFR, and TNFa (Supplementary Figure C.12B and C). GO terms were dominated by metabolic processes, however, I also identified a cluster of endoplasmic reticulum stress-related processes among the most significant GO-terms (Supplementary Figure C.12D). Downregulated genes were associated with the TFs Hnf4a, Ubtf, and Zpf263 and the pathway Androgen, Estrogen, EGFR, and MAPK (Supplementary Figure C.12E and F). GO terms were also almost exclusively related to metabolic processes (Supplementary Figure C.12G).

Out of the top 100 common genes, 53 were consistently up- and 47 downregulated (Figure 4.3H). Within this set, genes with inverse regulation in acute and chronic scenarios were extremely rare. Functional analysis of the upregulated genes identified the TF Klf5 and the pathways NFKb and TNFa as relevant (Figure 4.3I and J).

Among the overrepresented GO terms, the term cell-cycle had the highest frequency (Figure 4.3K). In agreement, I identified a cluster of 37 GO terms that corresponds to the biological process of proliferation and represents the most significant GO terms. Downregulated genes were associated with the TFs Hnf4a and Nr4a1 (Supplementary Figure C.13A) and biological processes were dominated by metabolic and catabolic related GO-terms (Supplementary Figure C.13B). Thus, the integration of chronic and acute mouse data unveiled genes that are deregulated in both damage models and can be differentiated from gene sets exclusively deregulated in acute or chronic liver damage.

Similarities between humans and mice

I performed a cross-species analysis to evaluate how well the altered gene expression in the chronic CCl_4 mouse model reflects the transcriptomic changes in humans that suffer from CLD. For this purpose, I collected genome-wide gene expression data from 5 publicly available patient cohorts with a total of 372 patients and five etiologies (Table 4.1, Supplementary Figure C.1B). These studies allowed us to calculate a total of 15 contrasts due to different disease stages and phenotypes.

Similar to the acute mouse models I first analyzed inter-study consistency comparing the similarity of the top 500 differentially expressed genes from each signature. Differential genes obtained from studies of the same groups of authors showed a higher degree of similarity (Supplementary Figure C.14). The highest similarity of two independent contrasts was observed between NAFLD 7 and HCV 6 (Jaccard Index of 0.154). In summary, the similarity of the top differentially expressed genes in humans appeared to be low. However, the mutual enrichment of the top 500 up-and downregulated genes demonstrated a very high consistency of the direction of regulation within contrasts of the same group of authors but also observed still relatively high accordance across the cohorts reported by different authors (Figure 4.4A). Partially, the direction of regulation of genes from the cohorts of patients with PSC, PBC, and NAFLD did not match well with the other contrasts. However, all other pairwise comparisons yielded convincing consistent results. Similar to the analysis of the mouse studies, this systematic comparison shows that similarities between different studies can better be identified by an enrichment analysis that considers the orientation (up, down) of expression changes than just focussing on the top differentially expressed genes.

Considering that previous studies reported only very low overlap between differentially expressed genes of humans and mice in CLD (Teufel et al., 2016) and the above-described limitations of this type of comparison I performed a cross-species enrichment analysis between the chronic CCl_4 model and the set of human data. For this purpose, I enriched the top 500 up-and downregulated genes from each human contrast in the three signatures from the individual time points of the chronic CCl_4 experiment in mice. I found a high degree of accordance where all human gene sets were significantly and consistently enriched at any time point of the chronic CCl_4 mouse signatures except for the up- and-downregulated genes of the PBC contrast, the downregulated genes of NAFLD contrast, and the upregulated genes of the NAFLD contrast; instead,

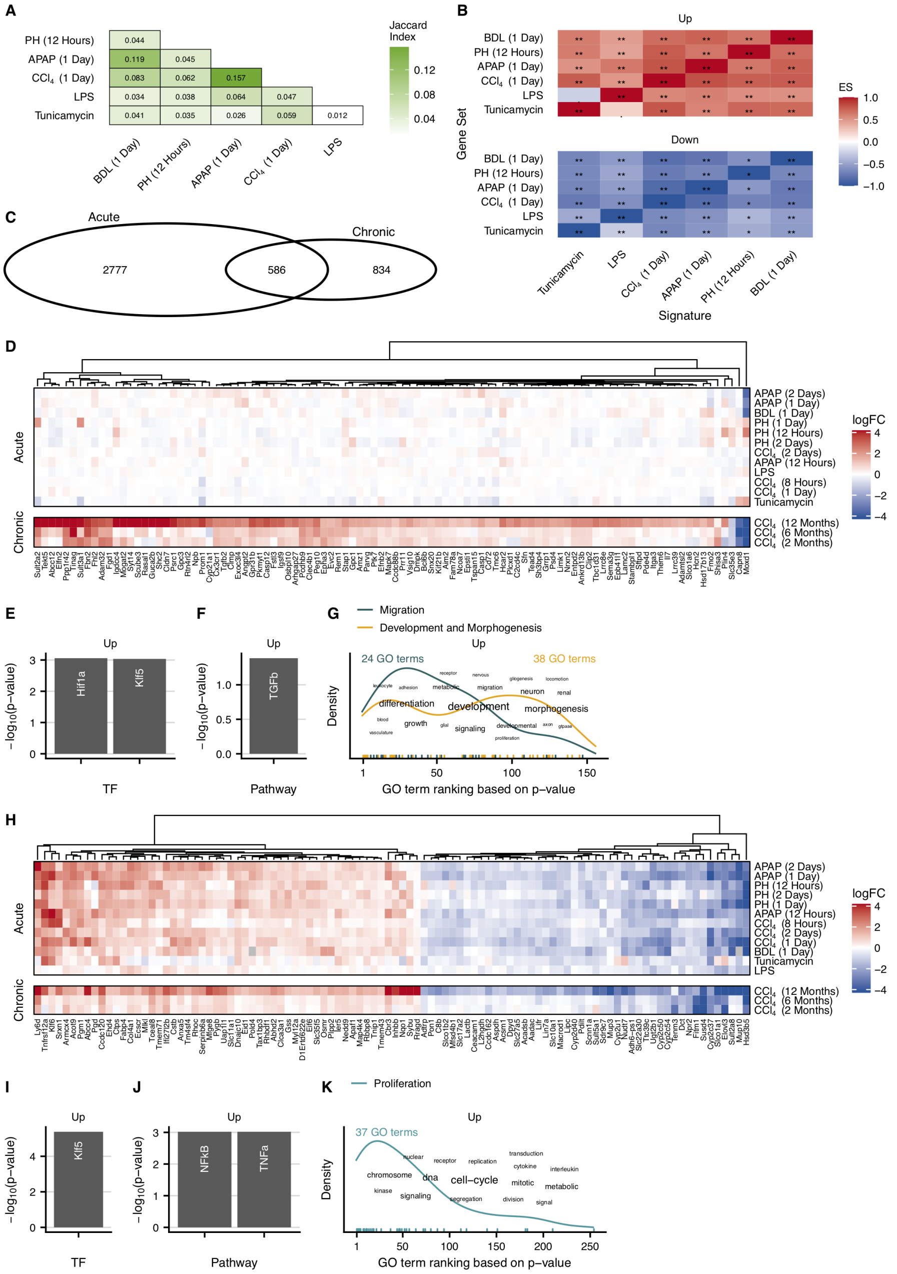


Figure 4.3: Comparison of gene expression changes in acute and chronic mouse models. A. Analysis of the similarity of the acute data sets. As a measure of similarity, the Jaccard index was calculated at the indicated time periods after the acute challenge. B. Pairwise enrichment analysis of the top 500 up- and downregulated genes (ES: enrichment score). C. Overlap of the unified acute and chronic (2, 6, and 12 months CCl 4) genes. D. Heatmap of genes exclusively deregulated in the chronic mouse model. E-G. Overrepresented transcription factors, identified by DoRothEA (E), pathways obtained by PROGENy (F), and GO terms (G) in the upregulated exclusive chronic genes. H. Heatmap of genes commonly deregulated in the chronic and acute mouse models. I-K. DoRothEA (I), PROGENy (J), and GO (K) overrepresentation of the genes upregulated in the acute and chronic mouse models.

I found that even the top 500 upregulated genes of the PBC contrast were significantly enriched among the downregulated genes of the 6-month time point of the chronic CCl_4 signature (Figure 4.4B).

Having shown that the expression profiles of the here established CCl_4 mouse model and the patients contain similar features, I set out to identify the genes which are consistently deregulated in humans and mice. My strategy is based on the leading-edge genes of the above-conducted enrichment analyses. By default, the leading-edge is defined as the subset of the gene set that mainly accounts for the enrichment score. I extracted the leading-edge genes only from those enrichment analyses that led to significant and consistent results in terms of the direction of regulation. Subsequently, I kept only those leading-edge genes that were identified in at least three human studies per time point in the chronic mouse model. The remaining leading-edge genes were defined as consistently deregulated genes in humans and mice in at least a single time point of the chronic mouse model. Across all time points, I identified 126 consistently up- and 102 consistently downregulated genes, whereby 45 (up) and 23 (down) genes were shared among all three chronic time points (Figure 4.4C).

To study whether it is possible to map those genes to individual cell types of the liver I integrated single-cell RNA-sequencing (scRNA-seq) data with the existing bulk data. For this purpose, I re-analyzed a publicly available scRNA-seq dataset of cirrhotic patients and healthy controls annotated with different cell types. Differentially expressed genes between cirrhotic and healthy patients were identified for each cell type individually ($FDR \leq 0.05$ & $|logFC| \geq 0.25$). The resulting cell-type associated sets of differential genes overlapped with 50 of the above described 228 genes that are consistently regulated in humans and mice; 41 of the 50 genes were exclusive for a single cell type. From the total 228 consistently deregulated genes I identified the top 100 genes with respect to the highest average logFC across all human and mouse bulk contrasts (Figure 4.4D). Of those 100 genes, 79 were up-and 21 were downregulated and 31 were mapped to a specific cell type. Overall, the direction of regulation was consistent between bulk and scRNA-seq data. Finally, a functional characterization was performed separately for all up-and downregulated genes. The upregulated genes were significantly associated with the pathways TGFb and TFNa and the TFs SP1, RELA, and NFkB1 (Figure 4.4E and F). Biological processes related to migration and development-and-morphogenesis functionally characterize the upregulated genes. The downregulated genes were dominated by metabolic processes including the pathway Androgen (Supplementary Figure C.15A and B). Remarkably, I found that the upregulated genes mapped frequently to cholangiocytes, endothelial as well as mesenchymal cells and macrophages, and downregulated genes to hepatocytes. Mapping the consistently deregulated genes back to the 12-month signature which contains all altered genes of the chronic CCl_4 mouse model revealed that the log-fold changes for upregulated genes are generally higher compared to the downregulated genes (Figure 4.4H). Genes consistently regulated between humans and mice did not show particularly low p-values or high fold-changes compared to all deregulated genes upon CCl_4 treatment. In summary, my cross-species enrichment analysis identified a set of 228 genes that are consistently deregulated in humans and mice during CLD.

The consistent regulation in humans and mice, as well as the cell type specificity of gene expression, was confirmed at the protein level, using commercially available antibodies against three of the consistently upregulated genes, LTBP2, ANXA5, and AKR1B10. A strong increase in the extracellular matrix protein LTBP2 occurred after 12 months of CCl_4 treatment in mice, and was expressed in the fibrotic tissue but not in the hepatocytes (Figure 4.5A). For analysis of the human situation, my collaborators Ahmed Ghallab and Jan Hengstler used independent biopsies of PSC patients. Since CCl_4 induces pericentral necrosis in mice and a similar zonation is known for alcohol-related liver disease (ARLD), they additionally tested biopsies of ARLD patients. Similar to mice, LTBP2 was expressed in the fibrotic streets of PSC (Figure 4.5B) and ALRD (Figure 4.5C) patients and the staining intensity increased with the fibrosis stage. Expression in fibrotic tissue also confirmed the results of the scRNA-seq analysis that identified a mesenchymal cell type preference of LTBP2. Immunostaining of the aldo-keto reductase AKR1B10 and of annexin V (ANXA5) also showed similar upregulation in mouse and human CLD (Supplementary Figure C.16).

Next, I placed the deregulated genes identified in the present chronic CCl_4 study into the context of previously published chronic mouse models and compared them to human CLD. The 12-month chronic CCl_4 model resulted in a much higher number of differentially expressed genes than all other previously published mouse models (2721 up- and 1437 downregulated genes; FDR ≤ 0.05 & $|\text{logFC}| \geq log_2(1.5)$; Supplementary Figure C.17A). In contrast, a model of the 18-week high-fat diet feeding yielded only 16 up and 19 downregulated genes, when applying the same cutoffs (Teufel et al., 2016). Similarity analysis of the differentially expressed genes between the chronic mouse models and the patient cohorts showed that the 12-month time point of the chronic CCl_4 model was more similar to human data than all other models (mean overlap coefficient of 0.37; Supplementary Figure C.17B). However, this analysis is biased towards the total number of differentially expressed genes. To study how well mouse models reflect expression changes in human CLD independently from the total number of differentially expressed genes, I first pooled the differentially expressed genes of the same human etiology (NAFLD, NASH, HCV, PSC, PBC) to a unified set of altered genes. NAFLD and HCV showed higher numbers of differential genes than NASH, PSC, and PBC. The majority of differentially expressed genes occurred in a single disease (84.9%), and 12.2, 2.4, 0.5% were altered in 2, 3, or 4 of the 5 investigated diseases, respectively (Figure 4.6A). No single gene was differentially expressed in all etiologies. To quantify the similarity between the individual chronic mouse models and the different human disease-specific gene sets, I computed precision and recall metrics. Recall denotes the ratio of altered human genes that are also altered in mice with respect to all altered human genes. Precision denotes the ratio of genes altered in mice that are also altered in humans with respect to all altered mouse genes. In general, precision and recall of the chronic mouse models for the different human disease etiologies were highly variable (Figure 4.6B). Precision and recall pairs were highest for NAFLD and lowest for PBC. The 12-month chronic CCl_4 model showed the highest recall among all etiologies. Moreover, recall of 12-months CCl_4 was

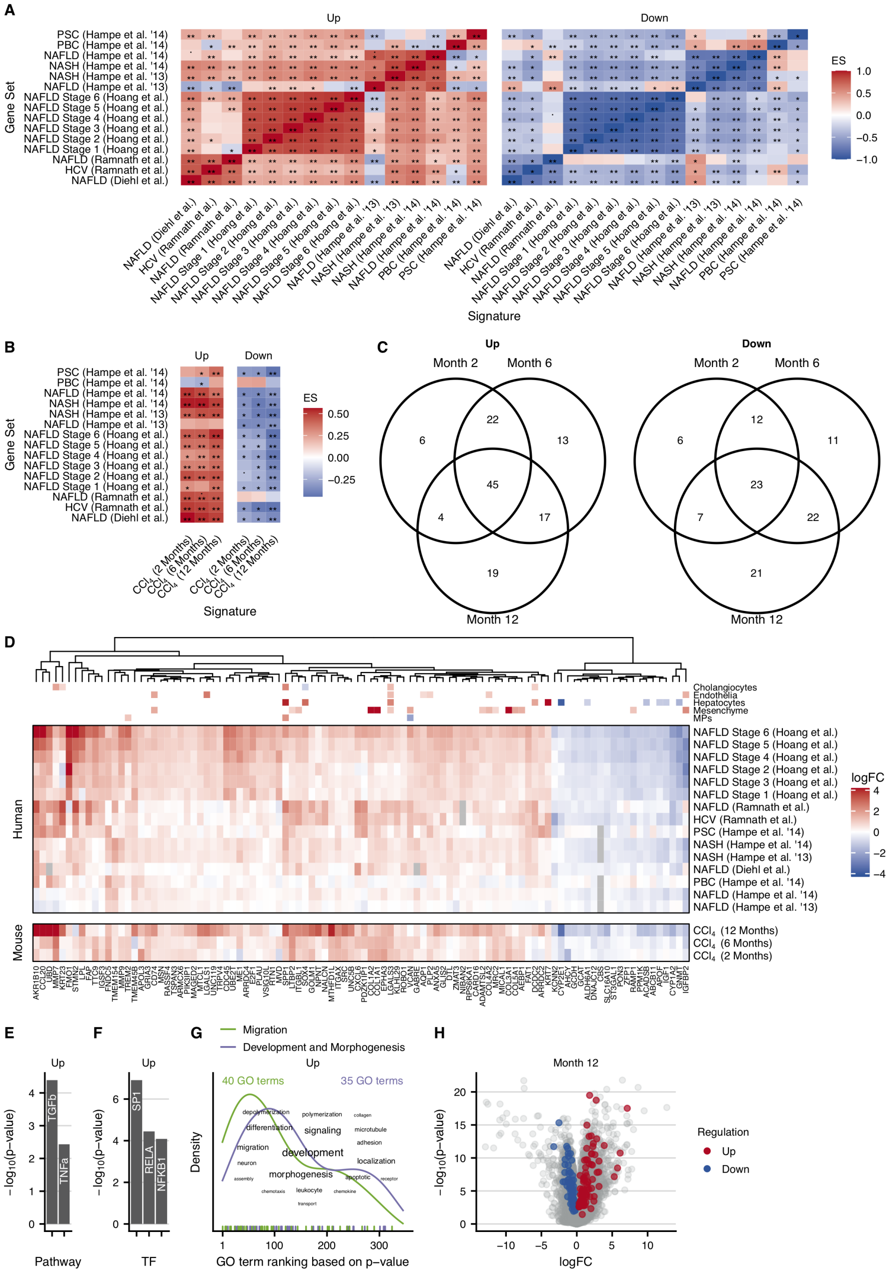


Figure 4.4: Human studies of liver disease and their similarities to the chronic CCl 4 mouse model. A. Pairwise enrichment analysis of the top 500 up- or downregulated genes of the human studies (ES: enrichment score). B. Similarity between the human studies and the chronic CCl 4 mouse model by pairwise enrichment analysis of the 500 top up- and downregulated genes. C. Overlaps of up- and downregulated genes in the chronic mouse model after 2, 6, and 12 months of CCl 4 administration that are consistently regulated in the human studies. D. Heatmap of the top 100 genes consistently regulated in the human studies and in the chronic CCl 4 mouse model. E-G. Characterization of the consistently deregulated genes in humans and mice by analysis of enriched pathways (E), transcription factors (F) and GO terms (G). H. Volcano plot of genes consistently deregulated in mouse and man (red and blue symbols) projected onto all genes deregulated in the chronic mouse model with CCl 4 (grey symbols).

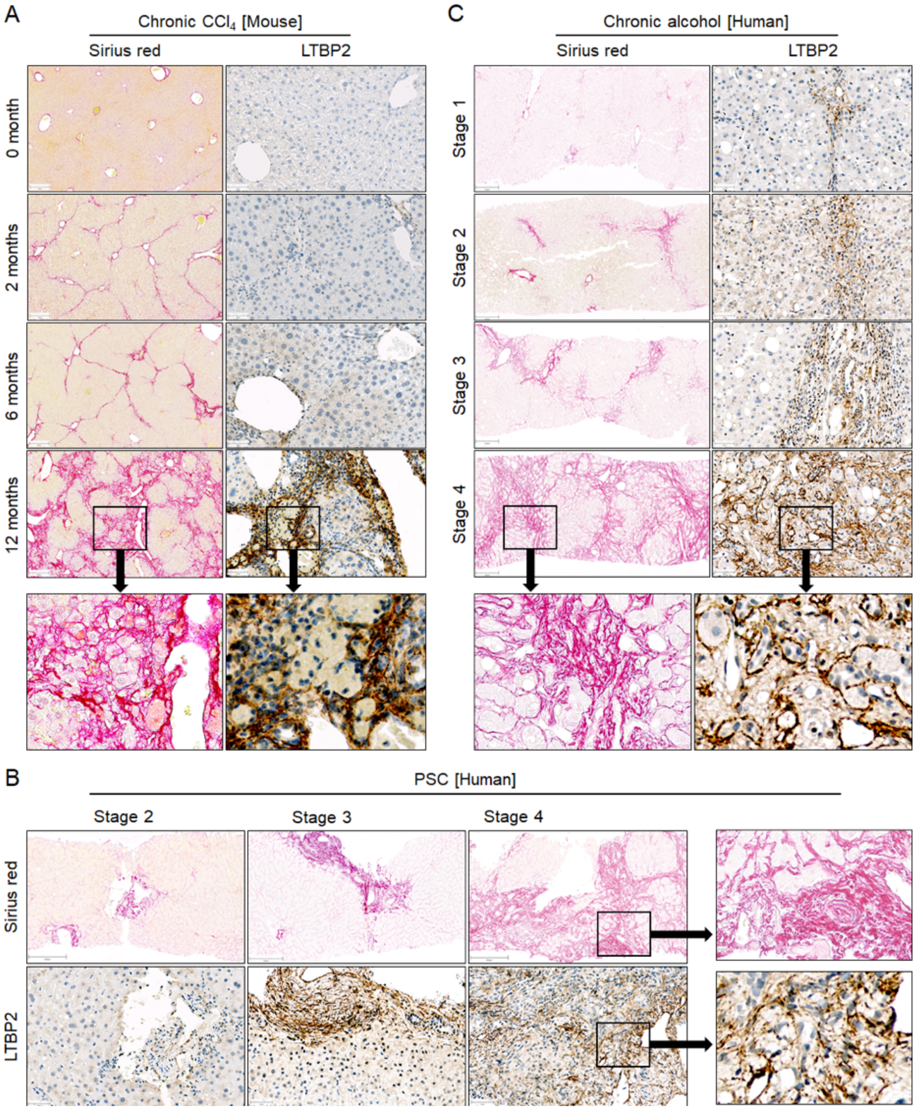


Figure 4.5: The extracellular matrix protein LTBP2 increases in CLD of mice and humans. A. Liver tissue of mice at different time periods after CCl 4 treatment. B. Liver tissue of patients with different stages of PSC. C. Liver tissue of patients with ARLD of different stages. Stainings were performed with Sirius red (scale bars 200 µm) to visualize fibrosis and with antibodies against LTBP2 (scale bars 200 µm) in liver tissue of the same patients. The entire figure was provided by Ahmed Ghallab.

always higher than that of the 6 or 2 months damage periods. The western-type diet model had the highest precision related to the upregulated genes of NAFLD.

Finally, I revisited the exclusively and commonly regulated genes of chronic and acute mouse models (Figure 4.3) to study their similarity to human CLD (Figure 4.6C). As expected, exclusive acute mouse genes showed the lowest recall and precision with respect to CLD. Remarkably, common genes (deregulated in acute and chronic mouse models) resulted in higher metrics than the exclusive chronic genes for several comparisons, particularly with respect to NAFLD.

Gene browser for comparison of human and mouse liver disease

To facilitate the assessment of the translational relevance of mouse models for specific human liver diseases, I established an open-access gene browser. This application provides for any gene of interest mean expression changes in the individual human and mouse studies, categorization into acute, chronic, or common response sets, the associated cell type, and if the gene is consistently altered in mice and humans.

4.5 Discussion

In the field of CLD, mouse models were successfully used for several preclinical developments (Jansen et al., 2017), though, it is also known that there exist large interspecies differences in the pathophysiology of the human and mouse liver. A former study analyzing the overlap of differentially expressed genes of liver tissue between CLD patients and mouse models with chronic liver injury reported only very little overlap, however, I have revisited this comparison for three reasons. First, the number of CLD patient cohorts with profiled transcriptome has increased in recent years. While the previous mouse-human comparison from 2016 included data from 25 patients with NAFLD and 27 with NASH (Teufel et al., 2016) I included a total 372 expression profiles of patients with NAFLD (n = 147), NASH (n = 42), HCV (n = 23), PBC (n = 11) and PSC (n = 14). Second, it is currently unclear, if longer exposure periods in chronic mouse models will improve the similarity between mice and human CLD. Third, a recent study suggested that a stress response with upregulated inflammatory and downregulated metabolic genes occurs similarly in acute and chronic mouse models and in human CLD. Thus, a comprehensive mouse-human comparison should differentiate between acute, chronic, and common expression responses.

Of all the chronic studies analyzed, the mouse model with 12 months of CCl_4 administration resulted in the highest recall of significantly altered genes in human liver disease. In detail, 38, 40, 25, 34, and 33% of all significantly upregulated genes in NAFLD, NASH, HCV, PBC, and PSC, respectively, were upregulated in the 12-month CCl_4 mouse model as well (Figure 4.6B). Considering that a previous study reported a low overlap in differentially expressed genes between both species (Teufel et al., 2016) my results show a remarkable higher similarity. I submit that the long exposure

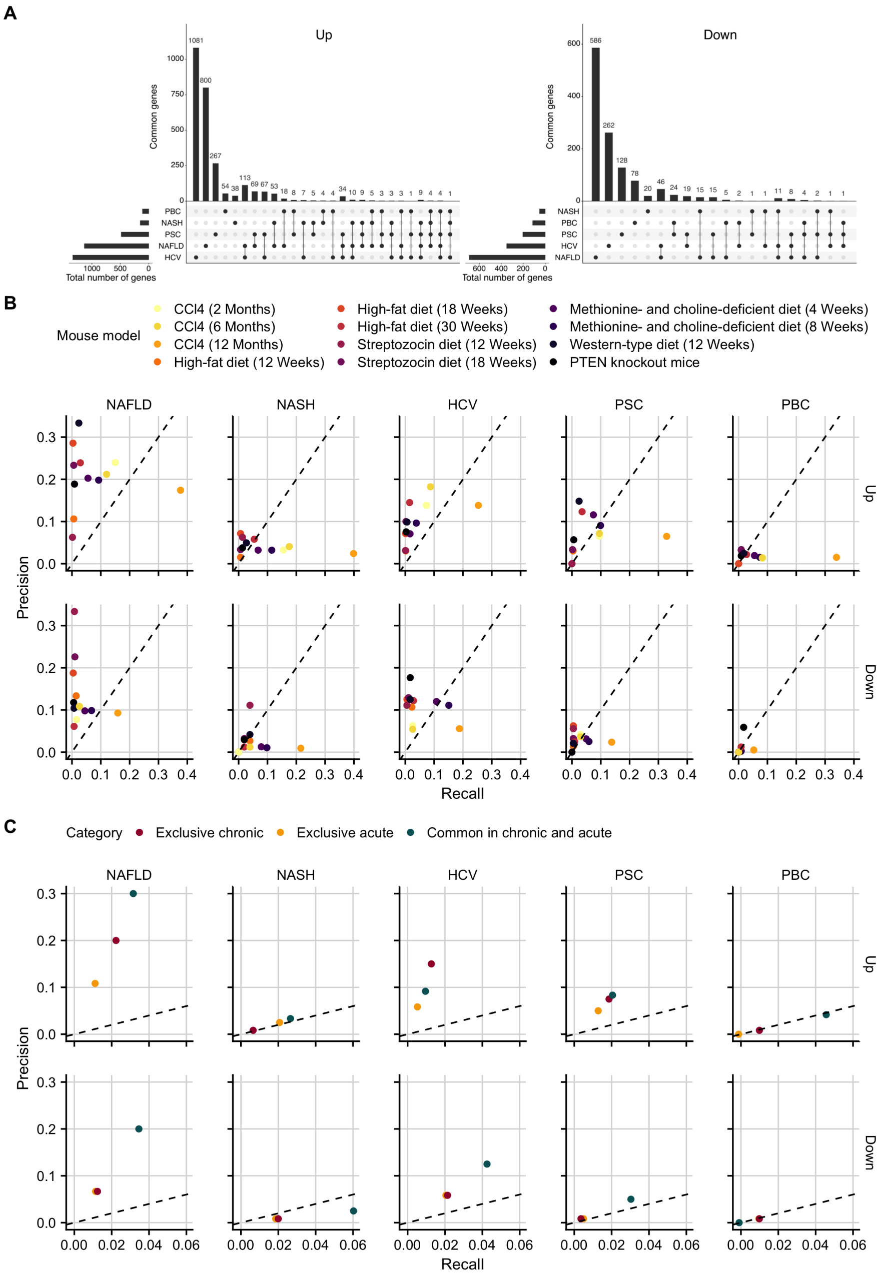


Figure 4.6: Recall and precision of 12 individual mouse models with respect to the human liver diseases NAFLD, NASH, HCV, PBC, and PSC. A. Gene sets that are uniquely or commonly deregulated in individual human diseases. B. Recall and precision of the individual mouse models with respect to the five human liver diseases. All genes with FDR <= 0.05 and |logFC| >= log2(1.5) were included. C. Comparison of exclusive chronic, exclusive acute and common genes in acute and chronic mouse models to human data. To allow a direct comparison, the top 120 genes of each category were included.

time of mice is an important factor for the larger human-mouse overlap The recall of upregulated genes for human NASH was only 0.16 and 0.18 for the models where mice were exposed for 2 or 6 months but increased to ~0.4 after 12 months exposure. This trend was also observed in the other liver disease etiologies. Consequently, the relatively short exposure times used in previous mouse studies may explain the small overlap in differentially expressed genes between humans and mice. Although the recall of downregulated genes was generally lower than that of upregulated genes, the same basic observations were made with respect to time. In contrast to the high recall, precision was lower for the 12-month CCl_4 mouse model, suggesting that in addition to a number of human-relevant genes, many other genes are significantly deregulated in mice which is not the case for humans. Of the 12 different mouse models of CLD analyzed, the western-type diet (WTD) had the highest precision (0.33) for human NAFLD. This was not unexpected given the similarity in disease etiology. However, the WTD mouse model had a much lower recall (0.02), which may be due to the relatively short feeding period of only three months. Assuming that WTD has a similar time dependence as the CCl_4 model additional studies with longer feeding periods could improve the human-mouse similarity.

From a bioinformatics perspective, overlap analysis of differentially expressed genes may not be the optimal approach to identify a consensus set of genes altered in a specific human CLD and in a mouse model. Instead, I propose enrichment analysis to be superior, as it focuses on the direction of regulation and not the effect size of individual genes. Following this strategy, a more sophisticated comparison of the chronic CCl_4 mouse model and human studies identified a set of 228 genes with similar regulation in mouse and human CLD. These genes are enriched in the GO terms migration, development, and morphogenesis, and several are associated with immune cells and ductular reactions. Including scRNA-seq, I found that the upregulated genes were mostly expressed in cholangiocytes, macrophages, endothelial and mesenchymal cells; whereas, the downregulated genes were expressed in hepatocytes. This finding is consistent with the known characteristics of CLD, in which the number of cholangiocytes increases due to ductal reactions, while macrophages, mesenchymal cells, and endothelial cells are involved in inflammatory processes related to GO terms identified among the upregulated genes, such as migration and adhesion. In contrast, genes related to hepatocyte metabolism are mainly downregulated. For selected upregulated genes in mouse and human CLD, their protein expression was analyzed as well. This analysis highlighted the extracellular matrix protein LTBP2, which is involved in anchoring the latent form of TGF-beta to the ECM and plays a role in cell adhesion and tumor promotion (Chen et al., 2019; Michel et al., 1998). Accordingly, LTBP2 staining is negative in the liver of normal mice, weakly positive after 2 and 6 months of CCl_4 treatment, and strongly signaled at 12 months. Similarly, positive LTBP2 staining was observed in human PSC and alcoholic liver fibrosis, which increased with the fibrosis stage. I propose for further studies to analyze whether upregulation of LTBP2 is associated with TGF-beta signaling activity, as the TGF-beta pathway is significantly activated. Two other genes that were both upregulated in human and mouse CLD, namely aldoketoreductase AKR1B10 and annexin V, showed a similar pattern in gene and protein expression. Although a more systematic and comprehensive validation is still needed, these preliminary immunostainings suggest that the mouse-human consensus set identified here contains genes that can be validated at the protein level.

To compare the exclusively acute, exclusively chronic, and commonly altered genes in mice with the genes deregulated in human CLD, the group of Jan Hengstler generated additional data on acute challenges in mice. In the different acute challenges induced by chemical and surgical insults, I classified expression changes in exclusively acute, exclusively chronic, and common gene sets. The separation of the three groups seemed relevant to human CLD as exclusively acute genes showed only little overlap with human genes, which is not surprising given the chronic nature of CLD. However, the common genes showed the highest similarity to patient CLD, especially for NAFLD and NASH. This was surprising given the difference in CCl_4 damage compared to the hypercaloric, high-fat etiology of human NAFLD/NASH. One explanation may be the recently published hypothesis that different types of injury cause similar expression changes in mouse and human liver, with inflammatory genes upregulated and metabolic genes downregulated (Campos et al., 2020). Exclusively chronic genes with enriched GO processes of development and morphogenesis also showed relatively high similarity, but lower than for the common genes.

In conclusion, my analyses led to the identification of genes that are similarly regulated in human and mouse liver disease. Although major species differences exist, the currently best available mouse models reach a recall of 0.4 and precision of 0.33 with respect to the genes significantly altered in human liver diseases.

4.6 Availability of data and materials

The code to perform all presented analyses is written in R (Gentleman et al., 2004; R Core Team, 2020; Wickham, 2016) and is freely available on Github. Reproducibility of all analyses is ensured by the R package workflowr (version 1.6.2; Blischak, Carbonetto, & Stephens (2019)) by deploying all my analysis scripts at a dedicated webpage. All datasets required to execute the code are available at Zenodo. The transcriptomics raw data of the here analyzed chronic and acute mouse models are available as superseries at GEO under accession number GSE166868.

Chapter 5

General conclusion and outlook

Chapter 2 and 3 of this thesis focussed on broadening the scope of the functional analysis tools PROGENy and DoRothEA by thorough benchmarking studies.

In the past, both tools had been shown to provide valuable mechanistic insight by inferring pathway and transcription factor activities from human bulk transcriptome data (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019; Schubert et al., 2018). Motivated by the fact that in my project within the "Liver Systems Medicine" (LiSyM) network many different gene expression data sets from mouse models were generated I strived for analyzing and characterizing also data from this model organism with PROGENy and DoRothEA. However, it was not clear whether both tools could provide biologically meaningful insight from mouse transcriptome data. For this purpose, I developed a systematic benchmarking pipeline where I showed that it is possible to transfer the regulatory knowledge of PROGENy and DoRothEA from human to mouse to functionally characterize also mice data.

With the emergence of scRNA-seq data, there was a growing need for functional analysis tools to analyze this novel data type. In the early days of this technology, many tools developed for bulk transcriptome analysis were readily applied to scRNA-seq data without any reasonable justification. My benchmarking study about the robustness and applicability of transcription factor and pathway analysis tools on scRNA-seq data was one of the first attempts to systematically evaluate the performance of bulk and scRNA-seq based tools. In summary, I was able to show that PROGENy and DoRothEA i) are robust against low gene coverage, i.e. drop-outs, ii) detect experimentally perturbed TFs/pathways with moderate accuracy iii) preserve cell-type-specific information while reducing noise in parallel, and iv) provide biologically meaningful activity scores.

Both benchmark studies were highly dependent on collecting and curating appropriate pathways and TF perturbation experiments as ground truth for the benchmark. Hence I mined the largest publicly available repositories of gene expression data such as Gene Expression Omnibus and Array Express to identify suitable experiments. For the cross-species benchmark, this endeavor was significantly facilitated by mining the then recently published CREEDS database containing the metadata of thousands of manually curated microarray data of drug and gene perturbation experiments for humans and mice (Zichen Wang et al., 2016). Although the scRNA-seq benchmark study was by far more complex in terms of included data than the cross-species benchmark, I needed also for the latter project a large collection of pathway and TF perturbation experiments. I was able to expand my previous collection of perturbation experiments with mostly further TF perturbation experiments that were previously collected and curated by Keenan et al. (2019) for the benchmark of the TF analysis tools ChEA3. In summary, both benchmark studies' feasibility and ultimate success were primarily made possible by the scientific community, who made their datasets or databases freely and publicly available.

Chapter 4 of this thesis demonstrated how functional analysis tools can provide meaningful insight from transcriptome data. In particular, I studied the similarities and differences in gene expression changes of acute and chronic liver disease in humans and mice. By a systematic analysis, I was able to identify gene sets containing i) genes similar altered between mouse models with chronic damage and liver disease patients or ii) genes exclusively and commonly regulated in chronic and acute liver damage in mice. Each gene set was systematically characterized by applying the tools PROGENy and DoRothEA which was made possible for the mouse-based gene sets by my previous cross-species benchmark. By integrating scRNA-seq I matched commonly deregulated genes in humans and mice to liver-specific cell types. In the future, I envision that the research of the liver and its diseases will benefit greatly from scRNA-seq data, which makes it possible to study the interplay of the individual liver and immune cell types on an unprecedented scale. The first corresponding large-scale data sets have recently been published (Cao et al., 2020; Dobie et al., 2019; Kim, Wu, Allende, & Nagy, 2021; Krenkel, Hundertmark, Ritz, Weiskirchen, & Tacke, 2019; Ramachandran et al., 2019; Segal et al., 2019).

As a side product of the scRNA-seq benchmark study, the results suggested that the performance of TF and pathway analysis tools is more sensitive to the quality of the used prior knowledge in the form of gene sets than the selected statistic to analyze them. This hypothesis laid the foundation for a crowdsourced follow-up project named decoupleR to systematically explore the impact of gene sets and statistics on the performance of functional analysis tools. Initial analyses confirm the hypothesis that well-curated gene sets are the most critical component for this type of analysis. Accordingly, and to make a significant step forward in the development of pathway and TF activity analysis tools, it is crucial to improve the quality of the used prior knowledge. The ever-increasing amount of generated transcriptome data promises a valuable data mine for this purpose. Regarding PROGENy, new pathway footprint signatures could be created or existing ones could be improved by exploiting the vast number of perturbation experiments from the Connectivity Map that systematically generated more than 1,500,000 perturbation signatures (Lamb et al., 2006; Subramanian et al., 2017). In addition, my cross-species benchmark suggests that mouse data could be integrated, but, to avoid additional confounding factors, I recommend relying on human data if possible. DoRothEA's regulons could be improved by integrating further data modalities such as information about chromatin accessibility generated via ATAC-seq. A recently published cell atlas of chromatin accessibility across 25 human tissues could be a precious data resource to tackle this challenge (Zhang et al., 2021)

Next to the general improvement of the consensus gene sets, there is a pressing need to derive and construct also cell-type-specific gene sets. This is particularly important for gene regulatory networks as different cell types can have fundamentally different gene regulatory programs. Currently, most attempts rely on reverse engineering of such networks from gene expression data of specific cell types or tissues. However, these approaches are mainly based on co-expression or mutual information so that there are many indirect and thus false-positive TF-target interactions (Barbosa, Niebel, Wolf, Mauch, & Takors, 2018). In Garcia-Alonso et al. (2019), it was shown that a consensus gene regulatory network constructed from various tissues and cell types still outperforms purely data-driven cell-type/tissue-specific networks. However, as soon as the generation of cell-type-specific improves cell type-specific information will be the preferred resource.

In recent years the first platforms to profile the transcriptome spatially resolved became available. This technology is referred to as spatial transcriptomics and promises to study the organization of cells in tissue in unprecedented detail. Hopes and expectations related to spatial transcriptomics were reflected by being awarded the method of the year 2020 by Nature Methods (“Method of the year 2020,” n.d.). In general, spatial transcriptomics resides in terms of covered genes and the number of cells per sample between scRNA-seq and bulk transcriptomics. Considering that I have shown that PROGENy and DoRothEA can be applied to scRNA-seq data and as originally intended to bulk transcriptomics it is reasonable to assume that they should also deliver biologically meaningful results for spatial transcriptome data, though a thorough benchmark study is still outstanding. Nevertheless, both tools have been recently successfully applied to one of the first spatial transcriptome data set of human myocardial infarction providing mechanistic insight into the differentiation of cardiac myofibroblast (Kuppe et al., 2020).

Even though pathway and TF activities alone are meaningful readouts of a cell’s/system’s state they must not be the endpoint of an analysis pipeline. Instead, these can be interpreted as features for further and more sophisticated downstream analyses. For example, Liu et al. (2019) utilize pathway and TF activities to identify and contextualize a causal signaling network from gene expression data using the tool CARNIVAL. Moreover, Tanevski et al. (2020) exploit these activities either as a predictor or response variable for a machine learning model named MISTy that aims to explain inter-cellular signaling from spatial transcriptome data.

In summary, I am convinced that the feature space of pathway and TF activities can contribute significantly to decipher the key mechanisms of diseases. For example, as the company DarwinHealth demonstrates, identifying master regulators in the field of personalized healthcare successfully helps identify the right drug at the right time for the right patient (Alvarez et al., 2018). Still, I am looking forward to seeing the impact of the next generation of these types of tools relying on substantially improved and extended prior knowledge.

Appendix A

Transfer of regulatory knowledge from human to mouse for functional genomics analysis

A.1 Supplementary Figures

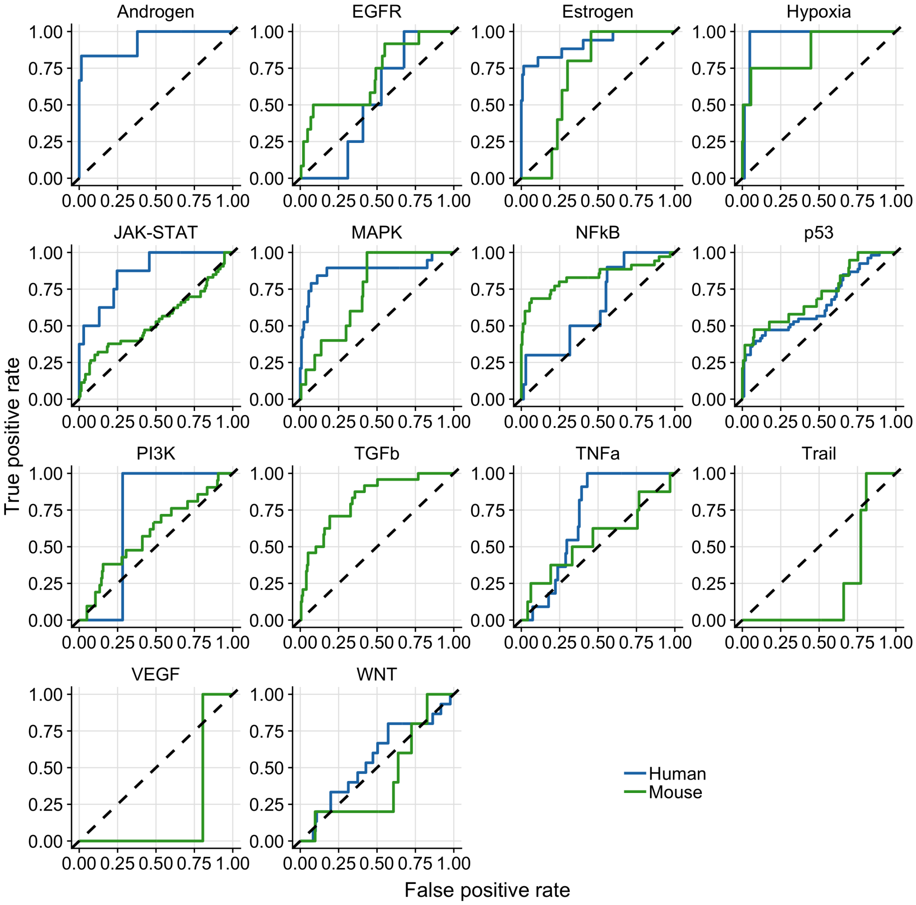


Figure A.1: Results of pathway-wise ROC-curves analysis. The dashed line indicate the performance of a random model. Missing mouse or human ROC-curves are due to missing benchmark data for the corresponding pathway.

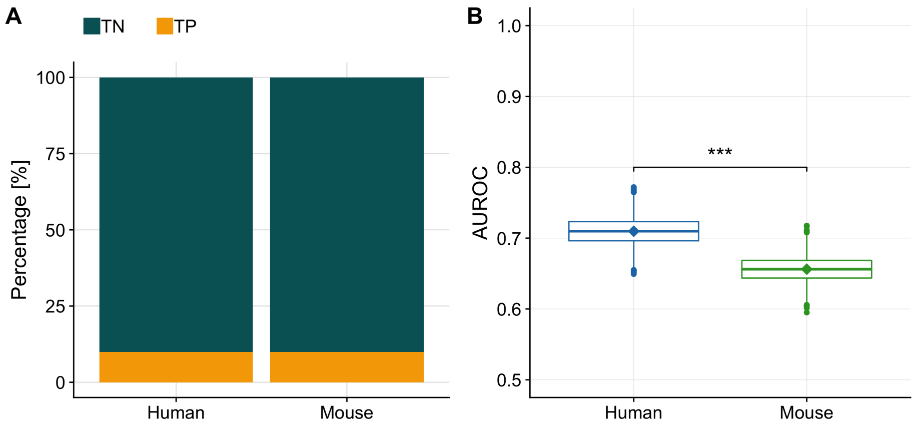


Figure A.2: (A) Barplot showing the imbalance of true negatives (TN) and true positives (TP) in our benchmark dataset for human and mouse. (B) Distribution of AUROC's computed for human and mouse separately from a balanced dataset (generated by downsampling the TN to equal the number of TP). The diamonds indicate the AUROC of the unbalanced dataset.

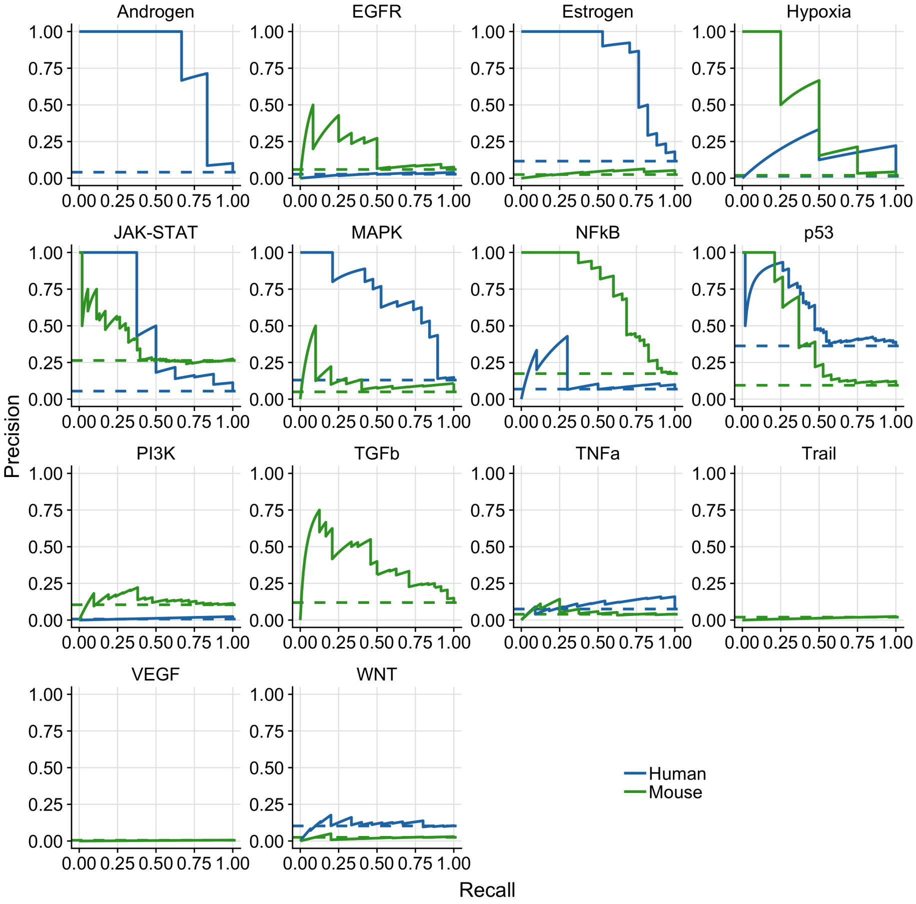


Figure A.3: Results of pathway-wise PR-curves analysis. The dashed line indicates the performance of a random model. Missing mouse or human PR-curves are due to missing benchmark data for the corresponding pathway.

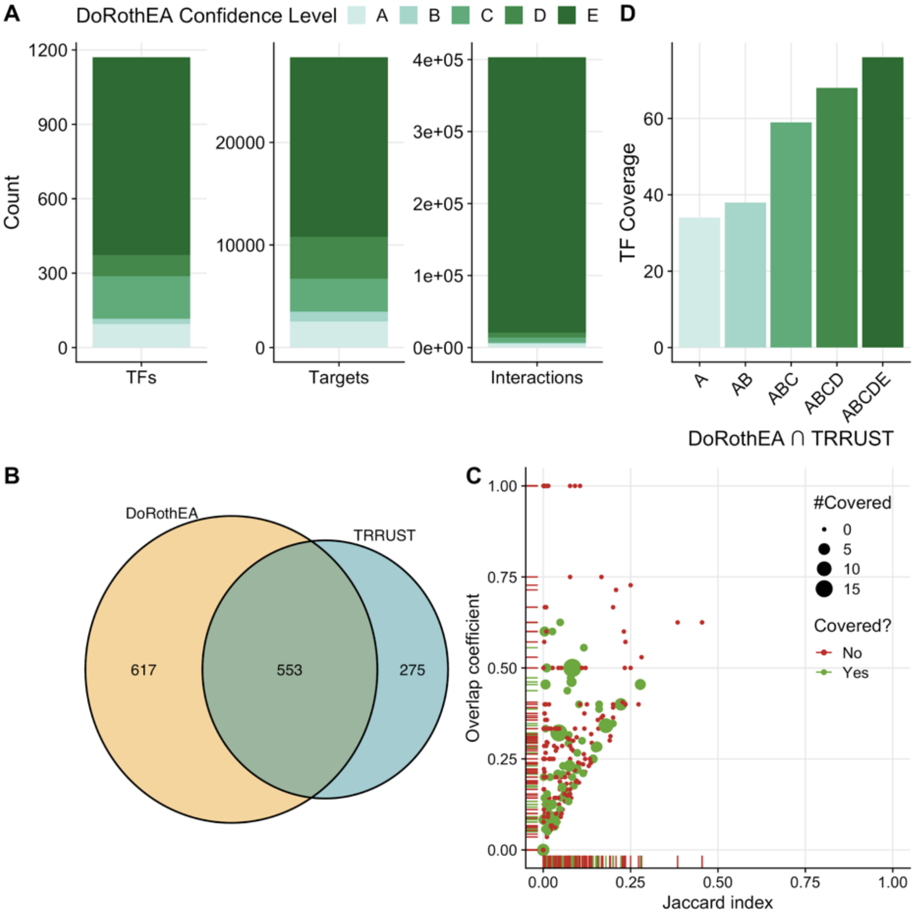


Figure A.4: (A) Mouse-DoRothEA properties showing number of transcription factors (TF), targets, and interactions itemized by confidence level. (B) Overlap of TFs between mouse-DoRothEA and TRRUST. (C) Similarity analysis of target genes for each overlapping TF between mouse-DoRothEA and TRRUST. Jaccard index and overlap coefficient were used to quantify similarity. Color and size indicate if and how often the TF was covered in the benchmark data. (D) Number of TFs covered in the benchmark dataset by intersection of mouse-DoRothEA and TRRUST dependent of the TF-confidence level.

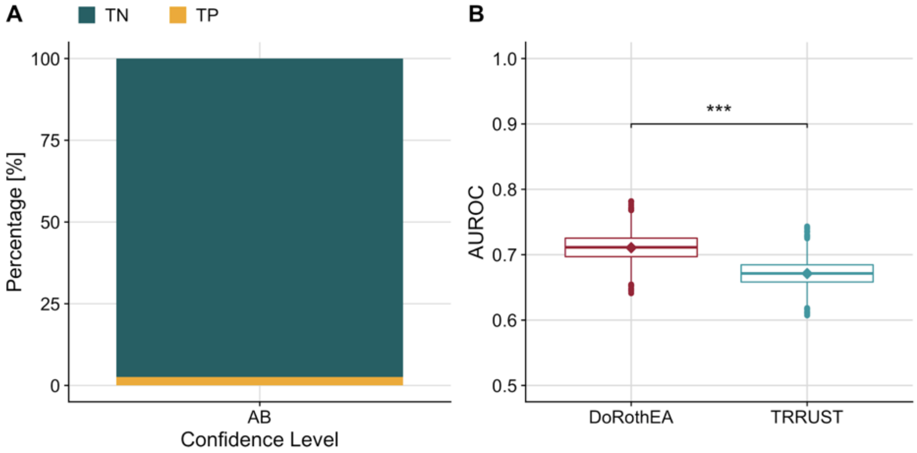
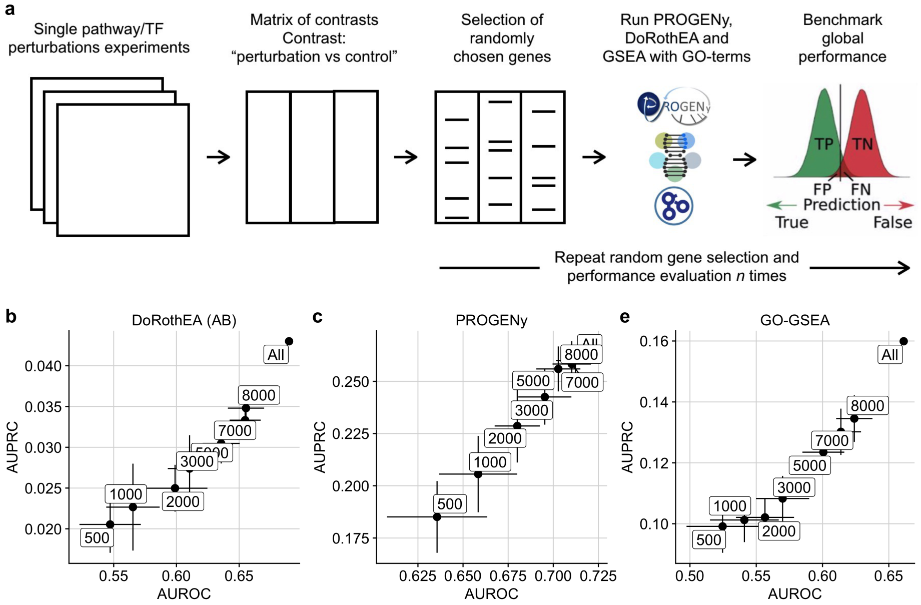


Figure A.5: (A) Barplot showing the imbalance of true negatives (TN) and true positives (TP) in our benchmark dataset for mouse-DoRothEA filtered for confidence level A or B. (B) Distribution of AUROC's computed for DoRothEA and TRRUST separately from a balanced dataset (generated by downsampling the TN to equal the number of TP). The diamonds indicate the AUROC of the unbalanced dataset.

Appendix B

Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data

B.1 Supplementary Figures



d

PROGENy pathway	GO term	GO ID
Androgen	GO_ANDROGEN_RECEPTOR_SIGNALING_PATHWAY	GO:0030521
EGFR	GO_ERBB_SIGNALING_PATHWAY	GO:0038127
Estrogen	GO_INTRACELLULAR_ESTROGEN_RECEPTOR_SIGNALING_PATHWAY	GO:0030520
Hypoxia	GO_REGULATION_OF_CELLULAR_RESPONSE_TO_HYPOXIA	GO:1900037
JAK-STAT	GO_JAK_STAT_CASCADE_INVOLVED_IN_GROWTH_HORMONE_SIGNALING_PATHWAY	GO:0060397
MAPK	GO_REGULATION_OF_MAPK_CASCADE	GO:0043408
NFkB	GO_NIK_NF_KAPPAB_SIGNALING	GO:0038061
p53	GO_REGULATION_OF_DNA_DAMAGE_RESPONSE_SIGNAL_TRANSDUCTION_BY_P53_CLASS_MEDIATOR	GO:0043516
PI3K	GO_PHOSPHATIDYLINOSITOL_3_KINASE_SIGNALING	GO:0014065
TGFb	GO_TRANSFORMING_GROWTH_FACTOR_BETA_RECEPTOR_SIGNALING_PATHWAY	GO:0007179
TNFa	GO_TUMOR_NECROSIS_FACTOR_MEDIATED_SIGNALING_PATHWAY	GO:0033209
Trail	GO_EXTRINSIC_APOPTOTIC_SIGNALING_PATHWAY_VIA_DEATH_DOMAIN_RECEPTORS	GO:0008625
VEGF	GO_VASCULAR_ENDOTHELIAL_GROWTH_FACTOR_SIGNALING_PATHWAY	GO:0038084
WNT	GO_WNT_SIGNALING_PATHWAY	GO:0016055

Figure B.1: a Workflow to induce low gene coverage with a subsequent benchmark of the tools PROGENy, DoRothEA and GO-GSEA against low gene coverage. b,c,e Scatterplot showing how well AUROC and AUPRC of b DoRothEA (AB), c PROGENy with 100 footprint genes per pathway, and GO-GSEA are correlated. The labels correspond to the gene coverages. d Mapping table between PROGENy pathways and GO terms/GO IDs.

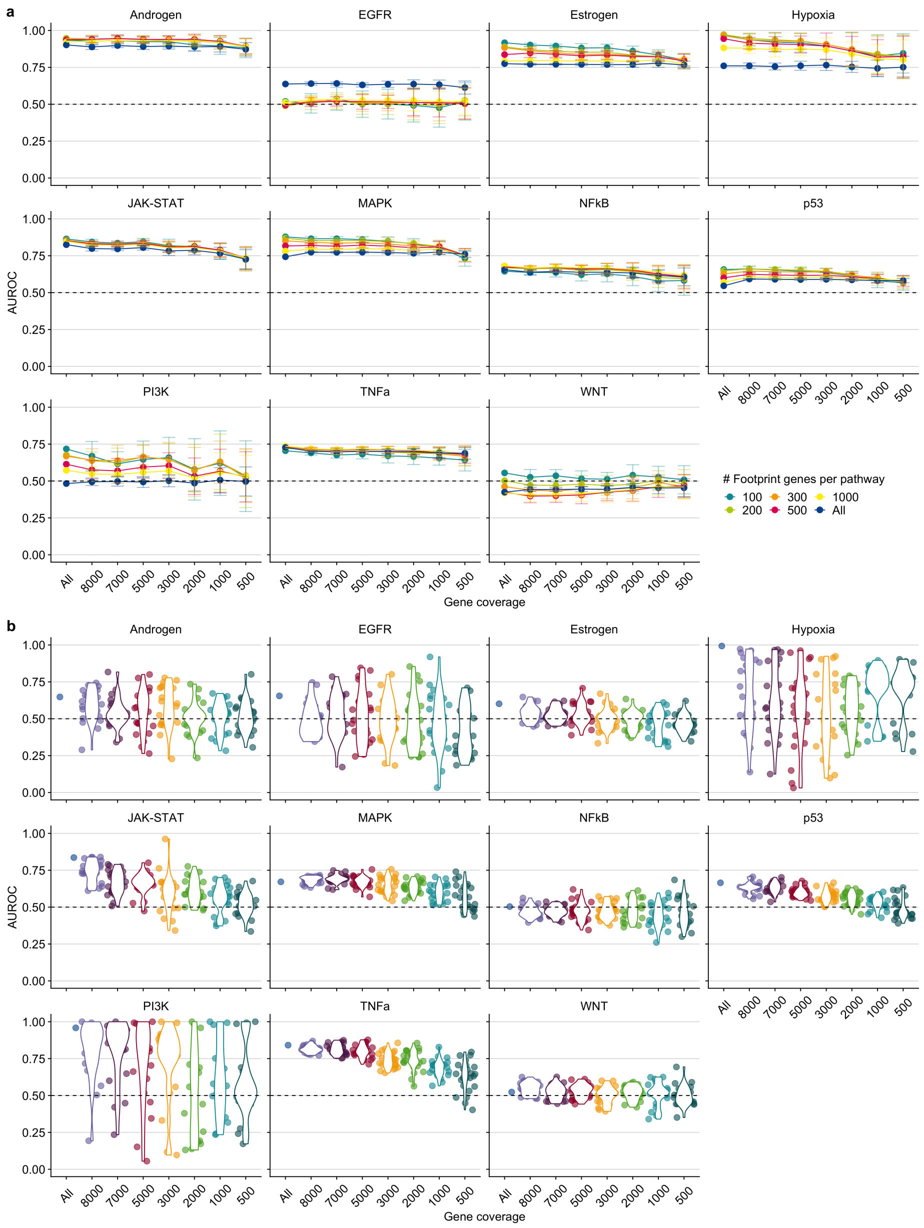


Figure B.2: Pathway-wise evaluation of a PROGENy and b GO-GSEA at different gene coverages. Performance is measured as Area under the ROC curve (AUROC). The dashed line indicates the performance of a random model. The colors in b are meant only as a visual support to distinguish between the individual violin plots and jittered points.

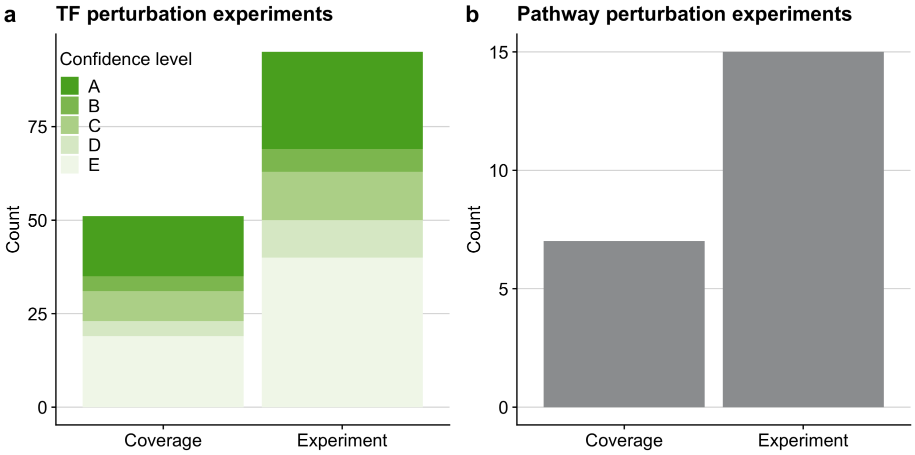


Figure B.3: Overview of the benchmark dataset of the in silico study for a TF and b pathway analysis tools. The term coverage denotes the number of distinct perturbed TFs and pathways in the benchmark dataset covered by the respective gene set resource. As individual pathways/TFs can be perturbed several times in independent experiments we also provide the total number of perturbation experiments. In the case of TF perturbation experiments we also provide DoRothEA's confidence class for each perturbed TF indicating the quality of its regulon within DoRothEA (A - high quality to E - low quality).

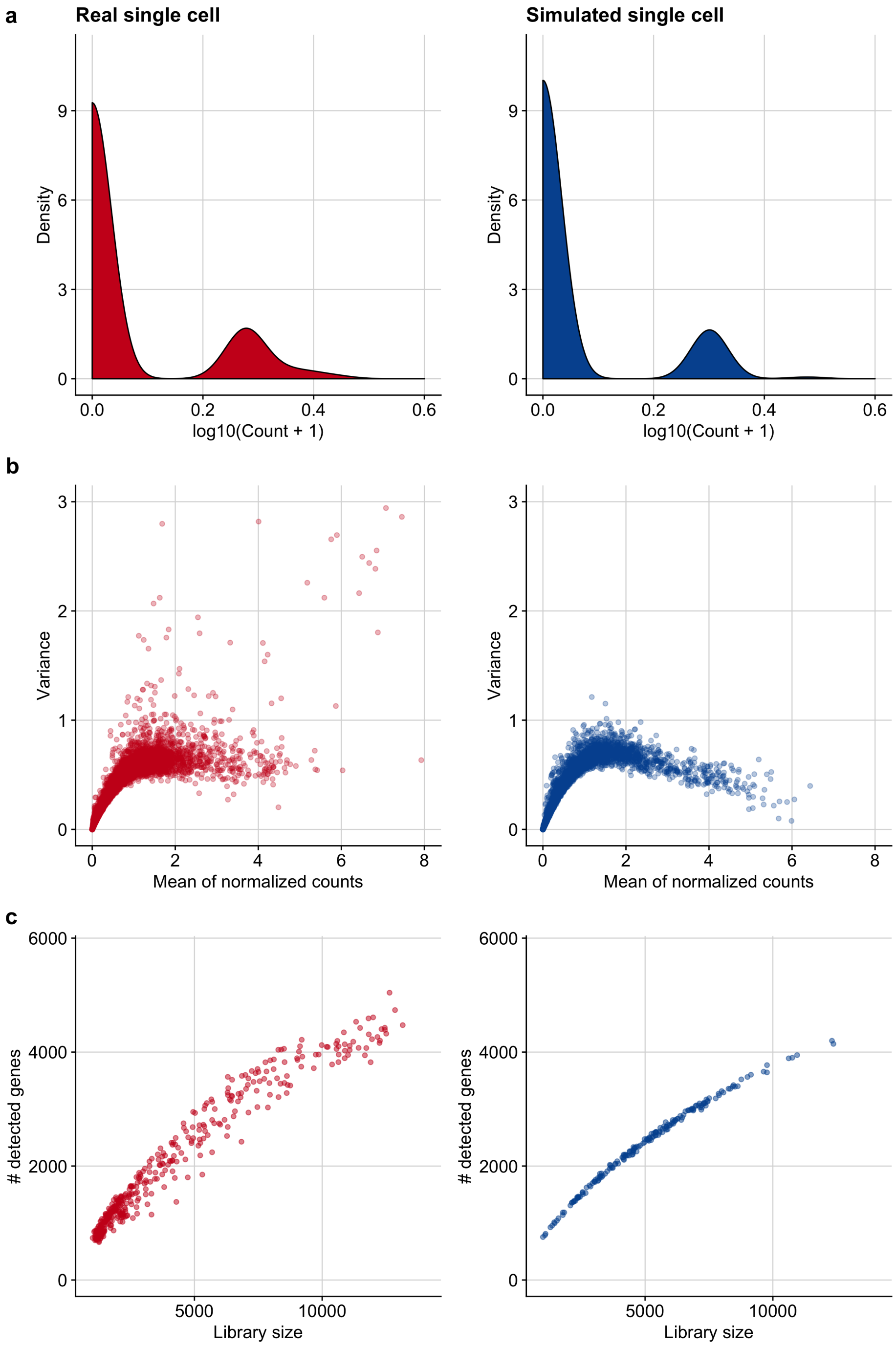


Figure B.4: Comparison of single-cell-specific properties between real and simulated single cells. a Count distribution of a representative gene for a real and a simulated single cell. b Mean-variance relationship of gene expression of a representative data set for a real and a simulated single cell. c The dependence of the number of detected genes in a real and a simulated single cell on the library size.

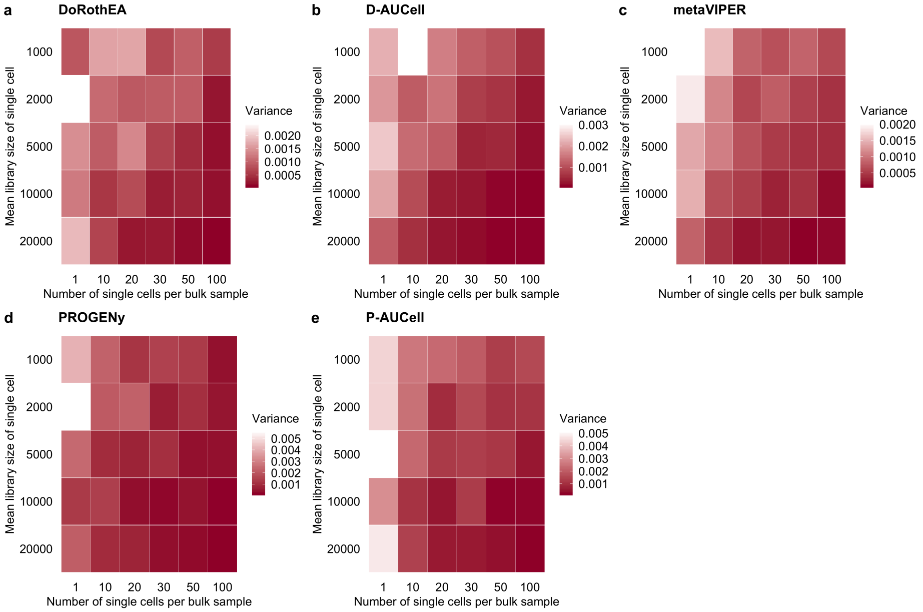


Figure B.5: Variance in the performance (measured as AUROC) of a DoRothEA, b D-AUCell, c metaVIPER, d PROGENy and e P-AUCell on single cells for different combinations of simulation parameters. The variance is calculated by repeating the simulation of each single-cell for each parameter combination 25 times.

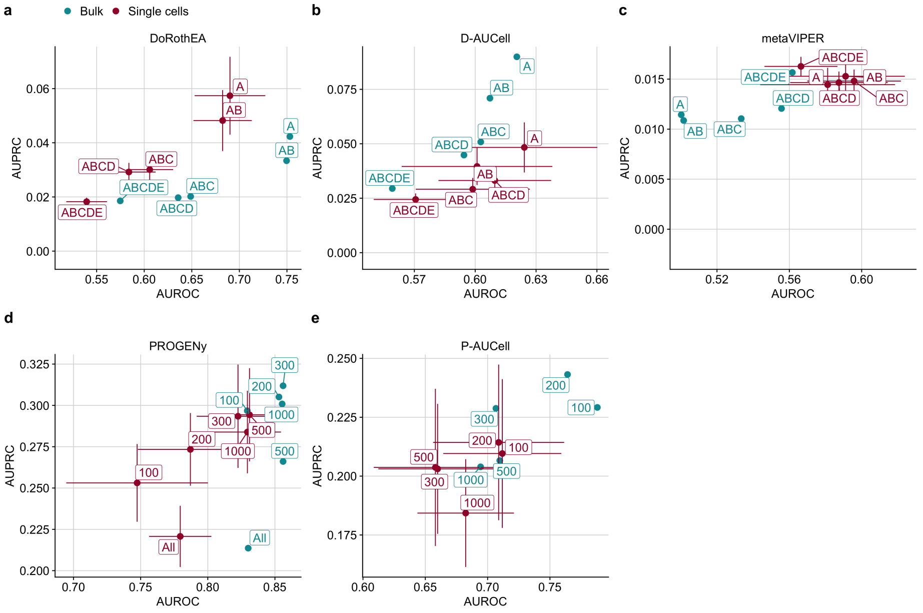


Figure B.6: Scatterplot comparing the performance of a DoRothEA, b D-AUCell, c metaVIPER, d PROGENy, and e P-AUCell on single cells and bulk, measured with AUROC and AUPRC with respect to different combinations of a,b,c DoRothEA's confidence levels or d,e different number of footprint genes per pathway.

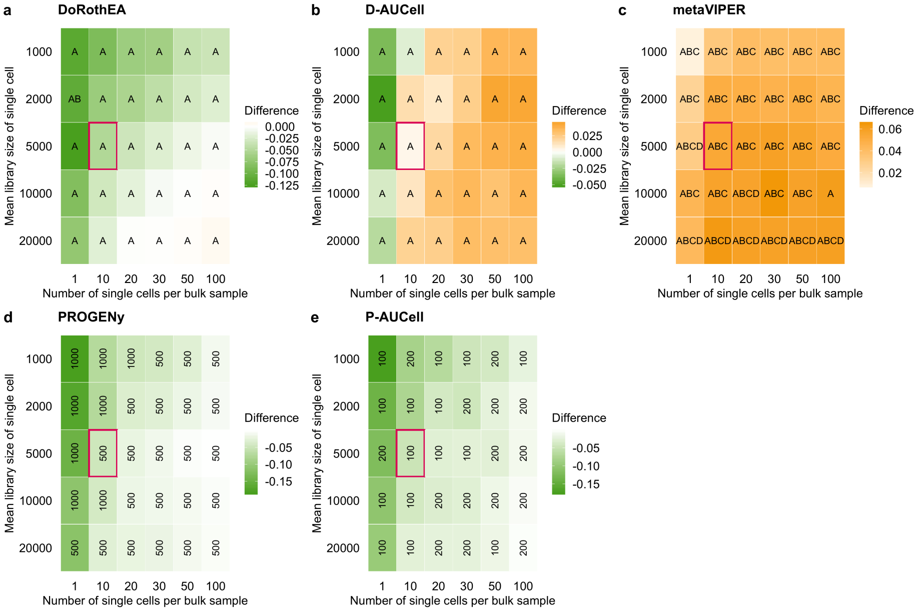


Figure B.7: Effect of the simulation parameters on the performance of TF and pathway analysis tools. The tile plots show the difference in performance of a DoRothEA, b D-AUCell, c metaVIPER, d PROGENy, and e P-AUCell between single cells and corresponding bulk samples, a,b,c across all confidence level combinations or d,e different number of footprint genes per pathway. A negative value indicates that the performance on bulk was better than on the simulated single cells and vice versa. The letters/numbers within the tiles indicates which confidence level combination/number of footprint genes per pathway performed the best on the single-cell data for the given parameter combination. The tile marked in red corresponds to the parameter setting used for previous plots in the main manuscript.

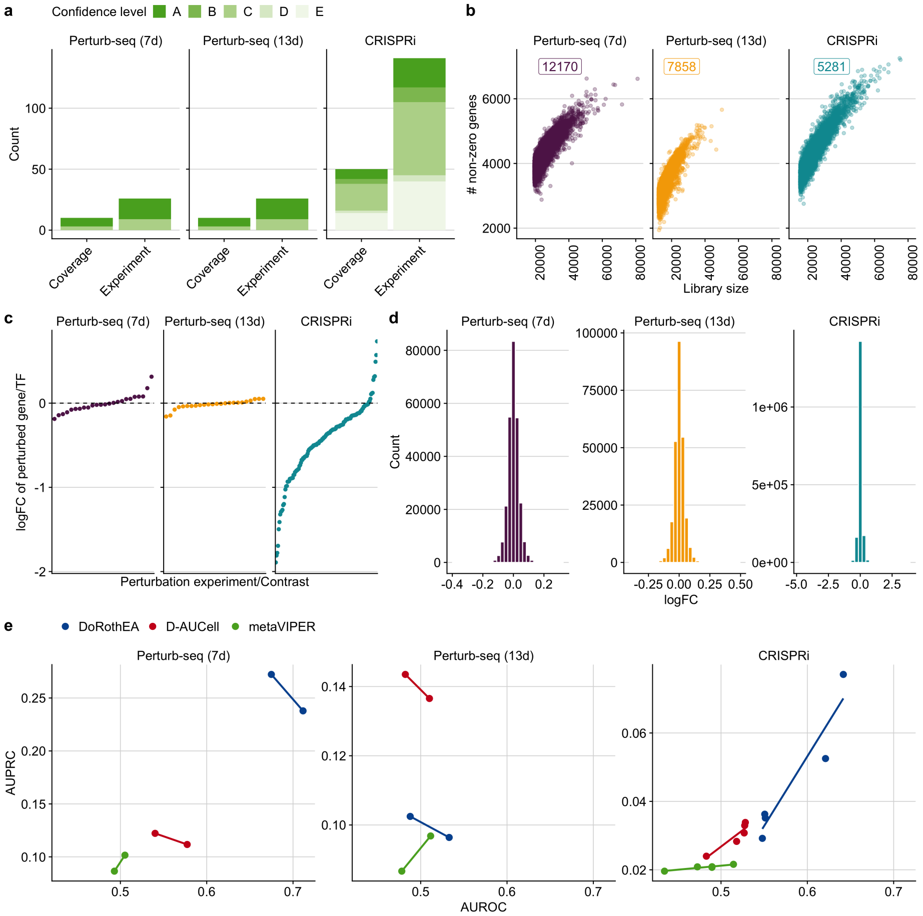


Figure B.8: a Overview of the in-vitro benchmark dataset. The term coverage denotes the number of distinct perturbed TFs in the benchmark datasets. As individual TFs can be perturbed several times in independent experiments we also provide the total number of perturbation experiments. We also provide DoRothEA's confidence class for each TF indicating the quality of its regulon (A - high quality to E - low quality). b The dependence of the number of detected genes on the library size for all benchmark datasets. The number of corresponding cells are displayed as well. c logFC of perturbed target/TF for the corresponding perturbation experiment for all benchmark datasets. d Distribution of logFC of all genes for each benchmark dataset. e Relationship between AUROC and AUPRC for DoRothEA, D-AUCell and metaVIPER with respect to different combinations of DoRothEA's confidence levels for each benchmark dataset.

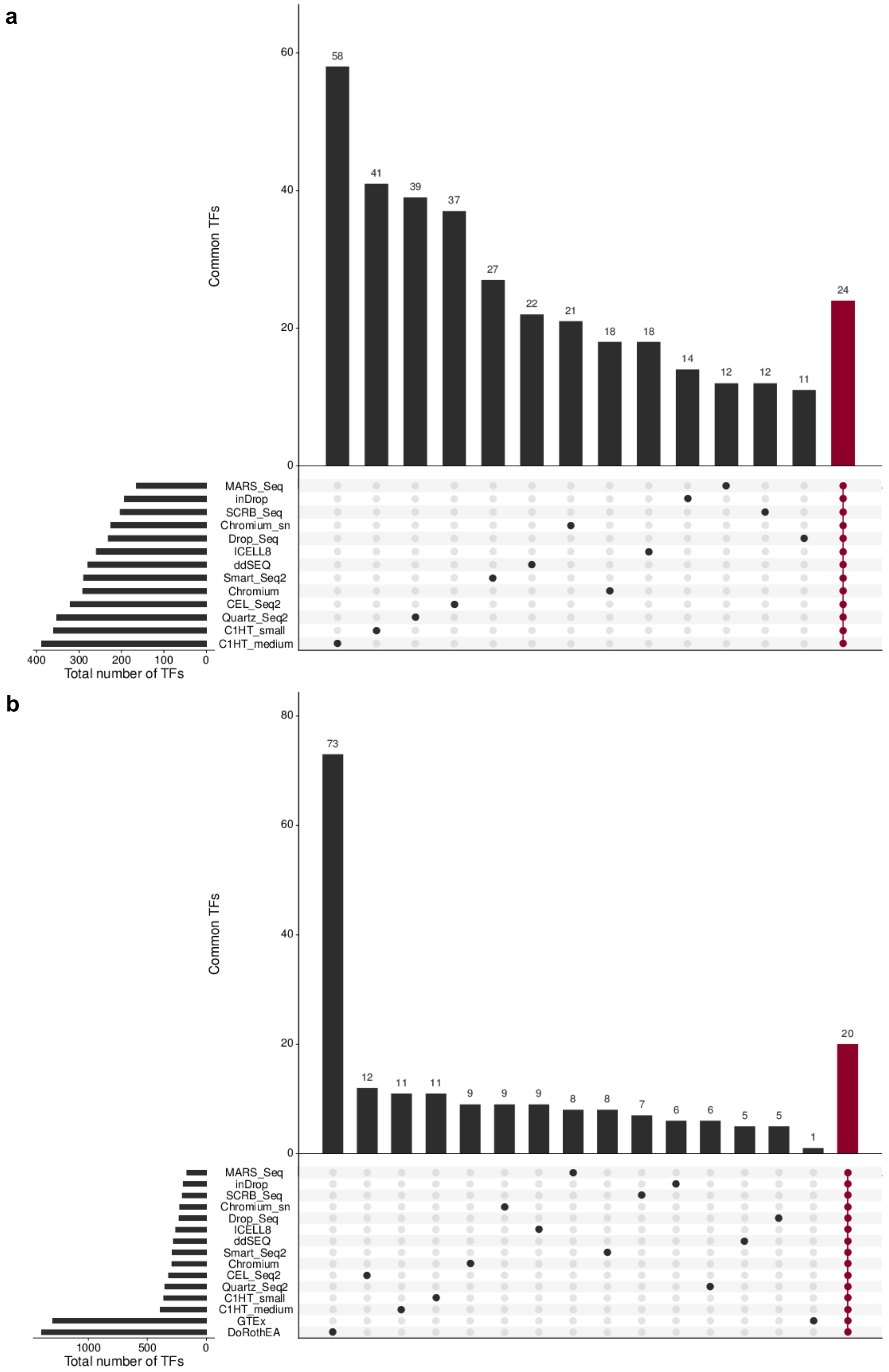


Figure B.9: Overlap of TF regulon resources. a Overlapping TFs of protocol-specific SCENIC regulatory networks. All 13 networks share 24 TFs. b Overlapping TFs between protocol-specific SCENIC regulatory networks, GTEx regulons and DoRothEA. All resources share 20 TFs. The remaining vertical bar plots indicate the number of TFs that are exclusive for the respective regulon resource. The horizontal bar plots indicate the total number of TFs for the regulon resource.

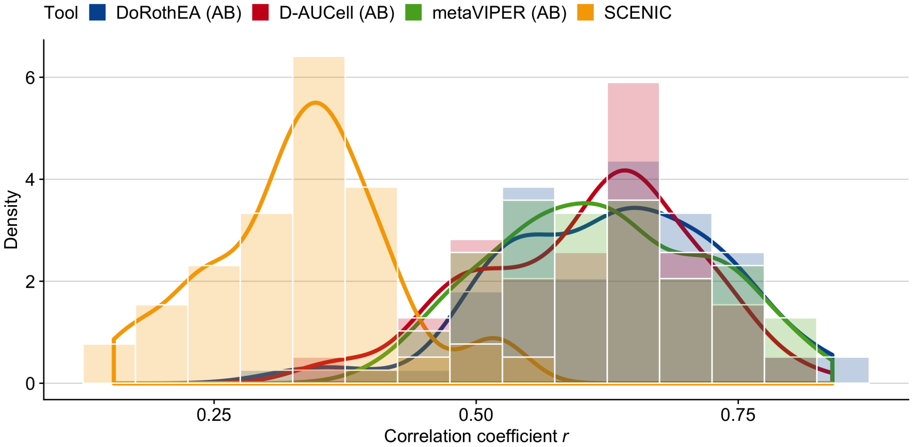


Figure B.10: Pairwise (Pearson) correlations of TF activities between the scRNA-technologies for each TF analysis tool.

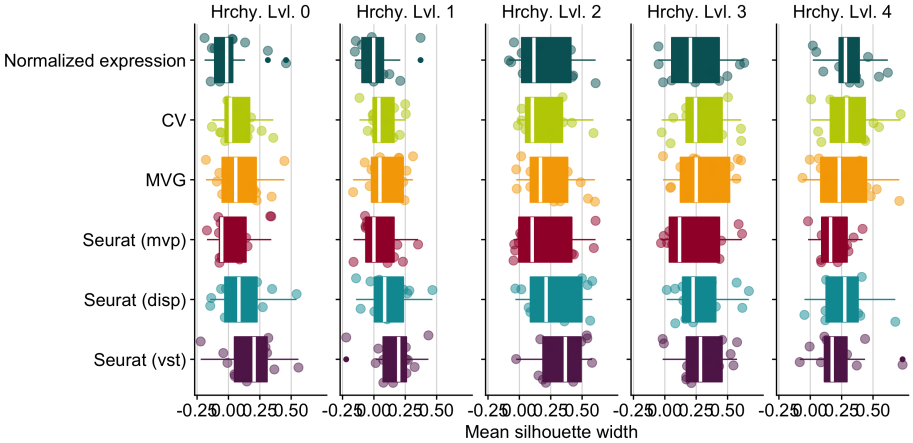


Figure B.11: Identification of the best method to determine the top 2000 highly variable genes to be considered for dimensionality reduction. We tested three different selection methods implemented in Seurat (disp = dispersion, mvp = mean.var.plot, vst). We also included CV (squared coefficient of variation - (sd/mean)**2) and MVG (most variable genes - genes with the highest variance). Those methods are compared to the case of considering the full gene expression matrix for dimensionality reduction, indicated here as ‘Normalized expression’.

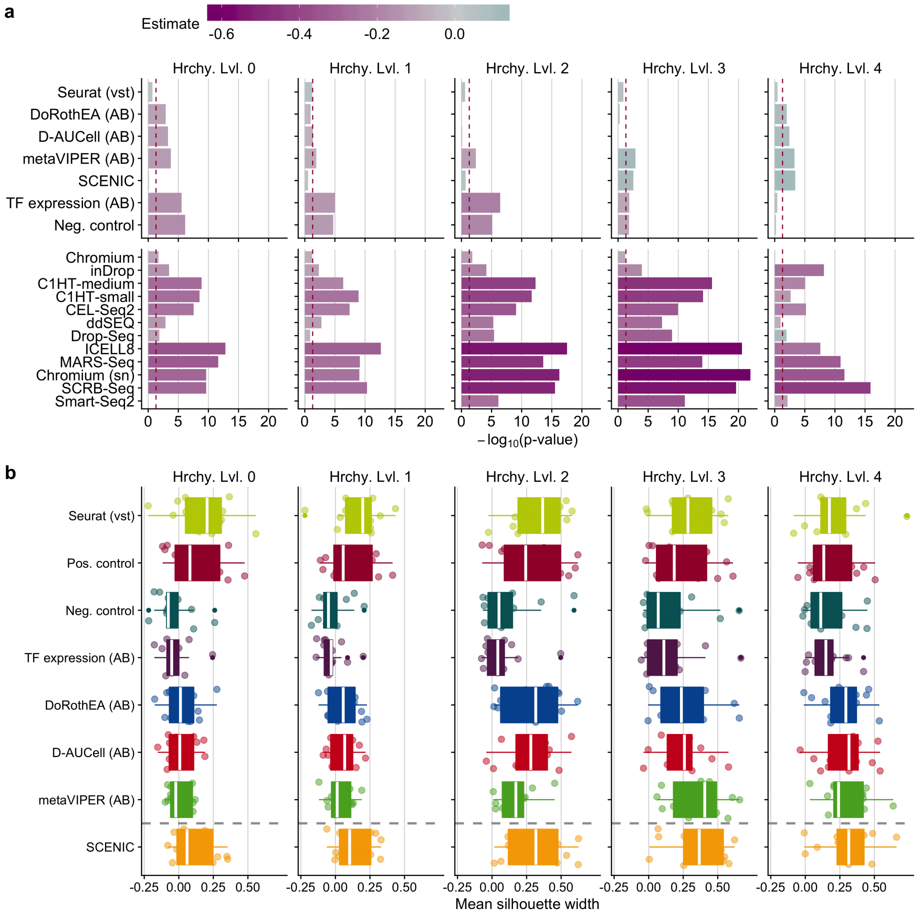


Figure B.12: a Statistical analysis of cell cluster purity in the reduced space: i) differences in the quality of clustering with respect to the positive control and ii) quality of clustering of scRNA-seq protocols in contrast to Quartz-Seq2 for TF activity tools. This analysis was performed independently for all hierarchy levels (Hrchy. Lvl.). The legend key 'estimate' corresponds to the estimated coefficients of the linear model. A negative value indicates a worse performance than the reference level (positive control for input matrices and Quartz-Seq2 for protocols) and vice versa.The dashed line indicates a p-value of 0.05. b Comparison of cluster purity measured by the silhouette widths obtained when considering highly variable genes identified by Seurat, TF analysis tools and controls for all hierarchy levels.

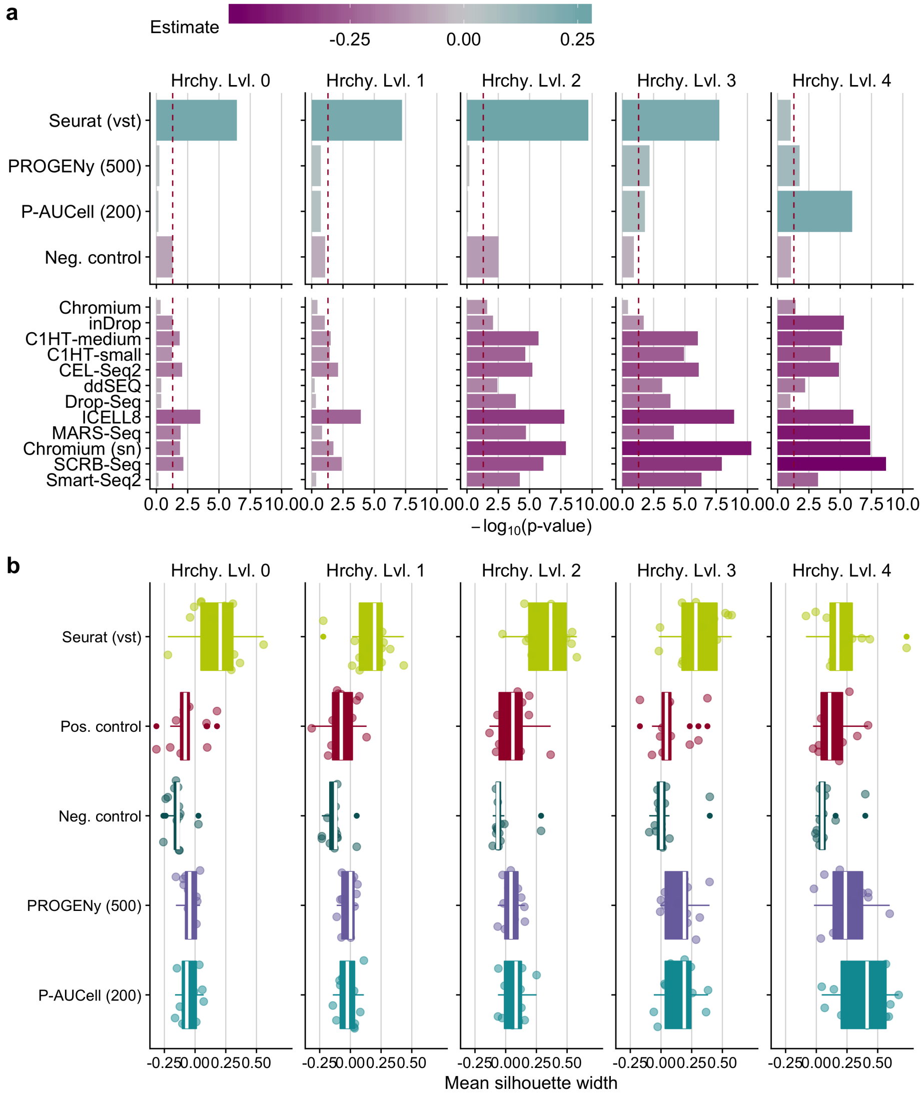


Figure B.13: a Statistical analysis of cell cluster purity in the reduced space : i) differences in the quality of clustering using different input matrices with respect to positive control and ii) quality of clustering of scRNA-seq protocol in contrast to Quartz-Seq2 for pathway activity inference tools. This analysis was performed independently for all hierarchy levels (Hrchy. Lvl.). The legend key 'estimate' corresponds to the estimated coefficients of the linear model. A negative value indicates a worse performance than the reference level (positive control for input matrices and Quartz-Seq2 for protocols) and vice versa. The dashed line indicates a p-value of 0.05. b Comparison of cluster purity measured by the silhouette widths obtained when considering highly variable genes defined by Seurat, pathway analysis tools and controls for all hierarchy levels.

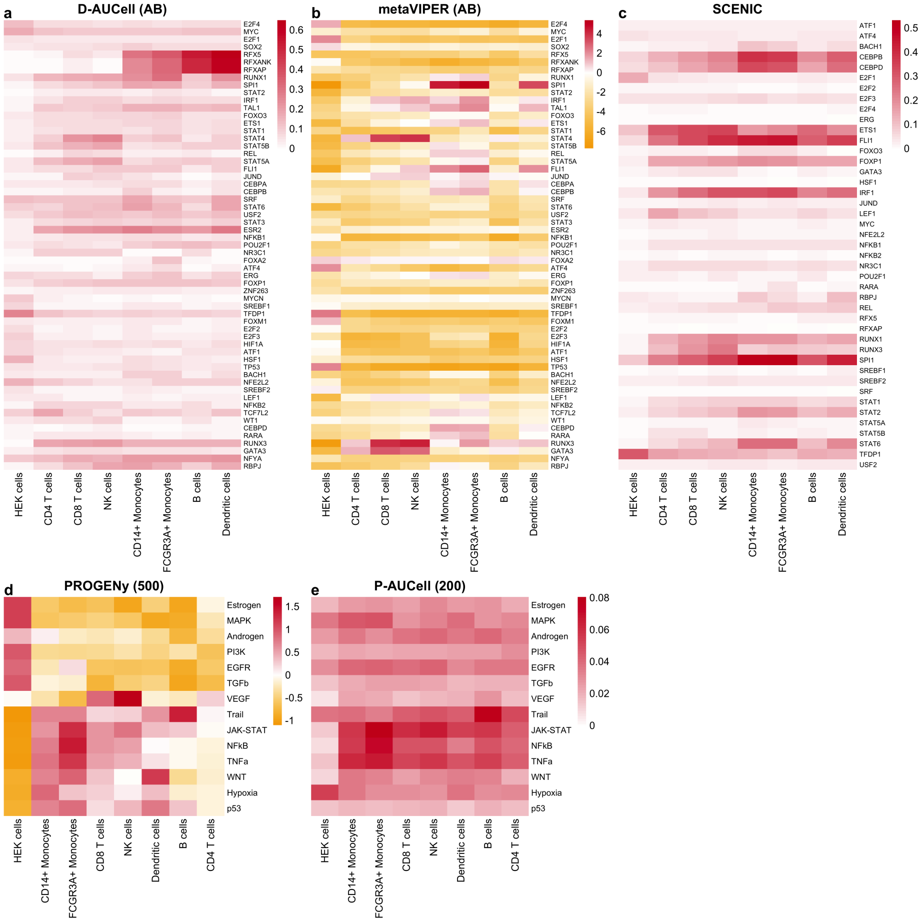


Figure B.14: Selected TF and pathway activities of various tools inferred from the Quartz-Seq2 gene expression data summarized for each cell type/cell line separately. The letters in the brackets correspond to DoRothEA's confidence levels and the numbers in brackets correspond to the number of footprint genes per pathway.

Appendix C

Transcriptomic cross-species analysis of chronic liver disease reveals consistent regulation between humans and mice

C.1 Supplementary Material and Methods

Collecting, curating, and processing publicly available genome-wide transcriptomics data

The raw data of the publicly available genome-wide transcriptomic studies (.CEL files for microarrays and count matrices for RNA-seq) were downloaded from Gene Expression Omnibus (GEO; Edgar, Domrachev, & Lash (2002)) and ArrayExpress (Kauffmann et al., 2009). For the respective accession, identifier see Table 4.1. Associated metadata was manually collected or accessed via the R/Bioconductor package GEOquery (version 2.56.0; S. Davis & Meltzer (2007)). To further process the raw data of microarrays and RNA-seq a suite of different Bioconductor packages (Gentleman et al., 2004) was used.

Processing of microarray data

First, a probe-level model was fitted to the raw data to subsequently control the array quality based on the relative log expression values (RLE) and the normalized unscaled standard errors (NUSE) using the R/Bioconductor package oligo (version 1.52.0; Carvalho & Irizarry (2010)). Arrays that deviated more than 0.1 from 0 for RLE and from 1 for NUSE were discarded due to expected poor quality. Subsequently, the raw data was normalized with the RMA algorithm, also implemented within the oligo package. Dependent on the studied organisms probe identifiers were mapped either to mouse/MGI or human/HGNC symbols using the R/Bioconductor package annotate (version 1.66.0) in combination with the individual annotation package of the used array type (e.g. hugene11sttranscriptcluster.db). If several probes matched the same gene symbol the expression was averaged. Genes with a constant expression across all samples were manually removed.

Processing of RNA-seq data

The preprocessing started by filtering out lowly expressed genes using the R/Bioconductor package edgeR (version 3.30.0; Robinson, McCarthy, & Smyth (2010)) to increase the power of downstream statistical tests. Subsequently, the expression data was normalized by correcting for differences in library composition using also edgeR. Finally, the normalized data was transformed to log2-counts per million with the R/Bioconductor package limma (version 3.44.1; Ritchie et al. (2015))

Differential gene expression analysis

The differential gene expression analysis was performed using the R/Bioconductor package limma (version 3.44.1; Ritchie et al. (2015)). Unless otherwise stated a gene is considered differentially expressed with a false discovery rate (FDR) ≤ 0.05 and $|\text{logFC}| \geq 1$. Experiments with a time course or case-control design were handled differently. The resulting gene signatures of a differential gene expression analysis is referred to as contrast.

Time course design

For the acute CCl_4, Acetaminophen (APAP), and partial hepatectomy (PH) experiment, each time point is compared vs time point 0 to extract the effect of the respective intoxication or treatment. The chronic CCl_4 experiment was handled differently since time-matched oil controls were available for month 2 and month 12, but not for month 6 (Figure 4.1A). As the oil effect was assumed to be constant across time the expression values for the oil sample at month 6 were imputed by averaging the expression of the oil samples at month 2 and month 12. To regress out the effect of the oil the treated samples were compared against their time-matched oil controls. Also, the Bile duct ligation (BDL) experiment was handled differently because there were time-matched sham surgery samples available for days 1, 3, and 7 but not for day 21 (Supplementary Figure C.8A). For this reason, the expression values for the sham surgery sample at day 21 were again imputed by averaging the expression of the remaining sham surgery samples. Finally, the effect of BDL was extracted by comparing BDL samples vs time-matched sham surgery samples.

Case-Control design

For the experiments following the classical case-control design which is true for the Lipopolysaccharide (LPS) and Tunicamycin mouse models and all human studies, treated samples/patients with CLD were compared against untreated/healthy samples/patients or samples of patients with lower disease stages.

Clustering of time-series gene expression data

Time-series gene expression data were clustered with the software program STEM (Short Time-series Expression Miner, version 1.3.12; Ernst, Nau, & Bar-Joseph (2005); Ernst & Bar-Joseph (2006)). As input, log-fold changes were provided from preceded differential gene expression analysis, where each time point was compared against time point 0. Accordingly, within STEM the normalization strategy “No normalization/add 0” was selected. Expression profiles were clustered using the default STEM clustering method with a specified maximal change between time points of 10 to not exclude drastic changes in the gene expression program, given partially large time spans between individual time points. Up to 20 different model profiles were returned by STEM. All other STEM parameters were left as default.

Comparing the similarity of gene sets via the Jaccard Index and Overlap Coefficient

The similarity or overlap between two gene sets was summarized either as Jaccard Index or Overlap Coefficient. Jaccard Index was used if the tested gene set had the same size, while for unbalanced gene set size the Overlap Coefficient was used. For the similarity analysis among the different acute mouse models or patient cohorts, the gene sets were composed of the top 500 differentially expressed genes based on limma’s moderated t-value.

Testing direction of regulation with Gene Set Enrichment Analysis

To test whether the differentially expressed genes of an arbitrary study A are consistently regulated in an also arbitrary but independent study B Gene Set Enrichment Analysis (GSEA; Subramanian et al. (2005)) was performed. From study A the top 500 up-and downregulated genes were extracted based on the moderated t-statistic. If the upregulated genes of study A are enriched among the upregulated genes of study B GSEA returns a positive enrichment score (ES). For enriched downregulated genes among the downregulated genes of study B a negative ES. GSEA was performed with the R/Bioconductor package fgsea (version 1.14.0; Sergushichev (2016)). If the enrichment was significant ($FDR \leq 0.05$) it is concluded that either up- and/or downregulated genes of study A are consistently regulated also in study B.

Selection of the time point of the strongest altered expression profile

For the reversible acute time-course experiments (CCl_4, APAP, and BDL) the time point with the strongest deregulated expression profile was determined by comparing the mean distance of the individual time points to their respective control time point along with the principal component 1 (PC1) axis in the principal component (PC) space. The time point with the largest mean distance was considered as the time point of strongest deregulation. For the bile duct ligation experiment this strategy was not applicable due to the irreversible damage. We selected 24 hours as this was comparable to the selected time points of the other acute experiments.

Functional characterization of transcriptomic data with various gene set resources

To functionally characterize the mouse and human transcriptomic data sets different gene set resources were used, such as biological processes from Gene Ontology (GO (Ashburner et al., 2000)), DoRothEA's regulons of confidence level A, B, and C (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019; Christian H. Holland, Szalai, & Saez-Rodriguez, 2020) and PROGENy's top 100 pathway responsive-genes (Christian H. Holland, Szalai, & Saez-Rodriguez, 2020; Schubert et al., 2018). GO gene sets were accessed with the R package msigdf (version 7.1; `https://github.com/ToledoEM/msigdf`), which queries itself the Molecular Signatures Database (MSigDB; Liberzon et al. (2011)). DoRothEA and PROGENy gene sets were accessed via their corresponding R/Bioconductor packages dorothea (version 1.0.1) and progeny (version 1.11.3).

As the statistical method overrepresentation analysis (ORA) was applied via the R function `stats::fisher.test()`. The number of background genes was set to 20,000 as this reflects a reasonable number of covered genes in transcriptomic studies. The minimum required gene set size was set to 10. The false discovery rate (FDR) is computed for each gene set resource individually to account for multiple hypothesis testing.

Manual clustering and text analysis of GO-terms

Significantly overrepresented GO-terms (FDR ≤ 0.05) were further analyzed by manual classification into more coarse-grained biological processes and text analysis to identify the most common keywords. Text analysis was performed with the R package tidytext (version 0.2.6; Silge & Robinson (2016)). Numbers and highly unspecific words such as “process” or “regulation” were removed. The frequencies of the most common keywords were visualized as word cloud with the R package ggwordcloud (version 0.5.0; `https://lepennec.github.io/ggwordcloud/`).

Construction and ranking of exclusive chronic, exclusive acute, and common gene sets

To identify a set of genes that is exclusive for chronic or acute liver damage in mouse models the acute and chronic mouse studies were integrated.

Construction of gene sets

First, the union of all differentially expressed genes was built across all the contrasts derived from the chronic or acute mouse experiment. A gene was considered differentially expressed with a FDR $\leq$ 1e-4 and $|\text{logFC}| \geq 1$. This conservative FDR threshold was opted for to include only the most reliable deregulated genes in this analysis. Indeed all chronic contrasts were considered. In the acute setup, however, the last time point of the bile duct ligation experiment (21 days after surgery) was removed as this damage is no longer considered to be acute. A consensus gene-level statistic was assigned for each gene in the unified chronic and acute gene set by computing the median of the t-statistics. Based on the sign of this consensus gene-level statistic genes were classified as up or downregulated. Genes that are listed only in the chronic or acute gene set were considered as exclusive chronic or exclusive acute genes, respectively. Accordingly, genes that were covered by both gene sets were deregulated in the liver irrespective of the nature of the damage. Therefore this gene set is referred to as the "common" gene set.

Ranking of gene sets

For these three different gene sets, a metric was computed per gene i integrating statistics from the chronic and selected acute contrasts. Regarding the acute mouse models, the number of considered acute contrasts was expanded. Next to the time points with the strongest deregulated profile, all time points between 8 and 48 hours were integrated to take into account the dynamic process of liver injury and recovery.

Metric for exclusive chronic gene set: i) consensus chronic gene-level statistic $c_{consensus-chronic_i}$, ii) median t-statistic of selected acute contrasts μ_{acute_i} , and iii) variance of selected acute contrasts $\sigma^2_{acute_i}$.

$$rank\left(\left|c_{consensus-chronic_i} \cdot \frac{1}{\mu_{acute_i}} \cdot \sqrt{\frac{1}{\sigma^2_{acute_i}}}\right|\right)$$

This metric prioritizes genes that have a high consensus chronic gene-level statistic and at the same time are consistently not deregulated in selected acute contrasts.

Metric for exclusive acute gene set: i) consensus acute gene-level statistic $a_{consensus-acute_i}$, ii) median t-statistic of selected chronic contrasts $\mu_{chronic_i}$ and iii) variance of selected chronic contrasts $\sigma^2_{chronic_i}$.

$$rank\left(\left|a_{consensus-acute_i} \cdot \frac{1}{\mu_{chronic_i}} \cdot \sqrt{\frac{1}{\sigma^2_{chronic_i}}}\right|\right)$$

Similar to above this metric ranks genes high that have a high consensus acute gene-level statistic and at the same time are consistently not deregulated in the chronic contrasts.

Metric for common gene set: same variables as for the exclusive chronic gene set.

$$rank\left(c_{consensus-chronic_i} \cdot \frac{1}{\mu_{acute_i}} \cdot \sqrt{\frac{1}{\sigma^2_{acute_i}}}\right)$$

This metric will rank genes high that have a high consensus chronic gene-level statistic and simultaneously are consistently and strongly regulated in the same direction as in the chronic scenario in selected acute contrasts.

Identification of consistently deregulated genes in patients and chronic CCl_4 mouse model

To identify consistently deregulated genes in patients of CLD and chronic CCl_4 mouse model the top 500 up and top 500 downregulated genes selected by the absolute value of the moderated t-statistic were extracted from each human contrast. Subsequently, those gene sets were enriched in the three different contrasts from the chronic CCl_4 mouse model corresponding to the three different time points: 2, 6, and 12 months. For the enrichment, the R/Bioconductor package fgsea (version 1.14.0; Sergushichev (2016)) was used with 1000 permutations. Afterward, the leading-edge genes were extracted from each enrichment if it met two criteria: i) FDR ≤ 0.05 and ii) consistent regulation between human and mouse data, i.e. when the upregulated human genes were enriched in the upregulated mouse genes indicated by a positive enrichment score (ES)) and vice versa, indicated by a negative ES. As some human studies have multiple contrasts the leading-edge genes were unified per study and chronic time point. Subsequently, those leading-edge genes were filtered for appearing per chronic time point in at least three studies.

Mapping of consistently deregulated genes in human and mouse to cell types using single-cell RNA-sequencing data

To identify whether the consistently deregulated genes in mice and humans are deregulated in a specific cell type single-cell RNA-sequencing (scRNA-seq) data was required. A preprocessed scRNA-seq data set of cirrhotic patients and healthy controls generated by Ramachandran et al. (2019) was downloaded from `https://datashare.is.ed.ac.uk/handle/10283/3433`. The single-cells were annotated by 11 different cell types: MPs, T Cells, ILCs, endothelia, B cells, pDCs, plasma B cells, mast cells, mesenchyme, cholangiocytes, hepatocytes. This data set was stored in a

Seurat version 2 object and was manually transformed to a Seurat version 3 object, to make it compatible with the latest version of the R package Seurat (version 3.2.3; Butler, Hoffman, Smibert, Papalexi, & Satija (2018)) Subsequently, a differential gene expression analysis was performed for each cell type individually between cirrhotic and healthy individuals using the function `Seurat::FindAllMarkers()`. A gene was considered differentially expressed with an absolute logFC ≥ 0.25 and an FDR ≤ 0.05.

Cross-species mapping of gene symbols

To map MGI gene symbols to HGNC symbols or vice versa the R/Bioconductor package biomaRt (version 2.44.0; Durinck, Spellman, Birney, & Huber (2009)) was used, which itself queries the Ensembl Archive Release 99 from January 2020. Gene that did not match were discarded and in case of an ambiguous mapping, the gene with the highest absolute log-fold change was selected.

Accessing differentially expressed genes from published mouse models of CLD

The differentially expressed genes of 9 published mouse models of CLD were extracted from Supplementary Table 2 of (Teufel et al., 2016). The 9 mouse models are defined as: WTD = Western-type diet; HF12/18/30 = High fat diet for 12, 18 and 30 weeks; STZ12/18 = Streptozocin diet for 12 and 18 weeks ; MCD4/8 = Methionine- and choline-deficient diet for 4 and 8 weeks; PTEN = Phosphatase and tensin homologue deleted on chromosome 10 knockout mice

Teufel et al. (2016) defined a gene differentially expressed with FDR ≤ 0.05 and logFC $\geq log_2(1.5)$. Hence, this cutoff is also applied for the other mouse models and patients cohorts to make them compared to these chronic mouse models.

Gene browser for comparison of human and mouse liver disease

The online gene browser was built with the R package shiny (version 1.6.0; `https://shiny.rstudio.com`). The code to generate the application is freely available on GitHub.

C.2 Supplementary Figures

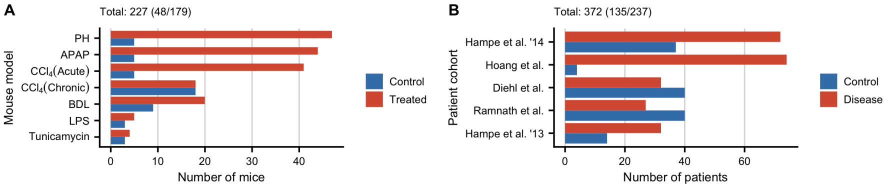


Figure C.1: Overview of the study cohorts. A. Number of analyzed mice per mouse model (control/treated). B. Number of analyzed patients per cohort (control/disease).

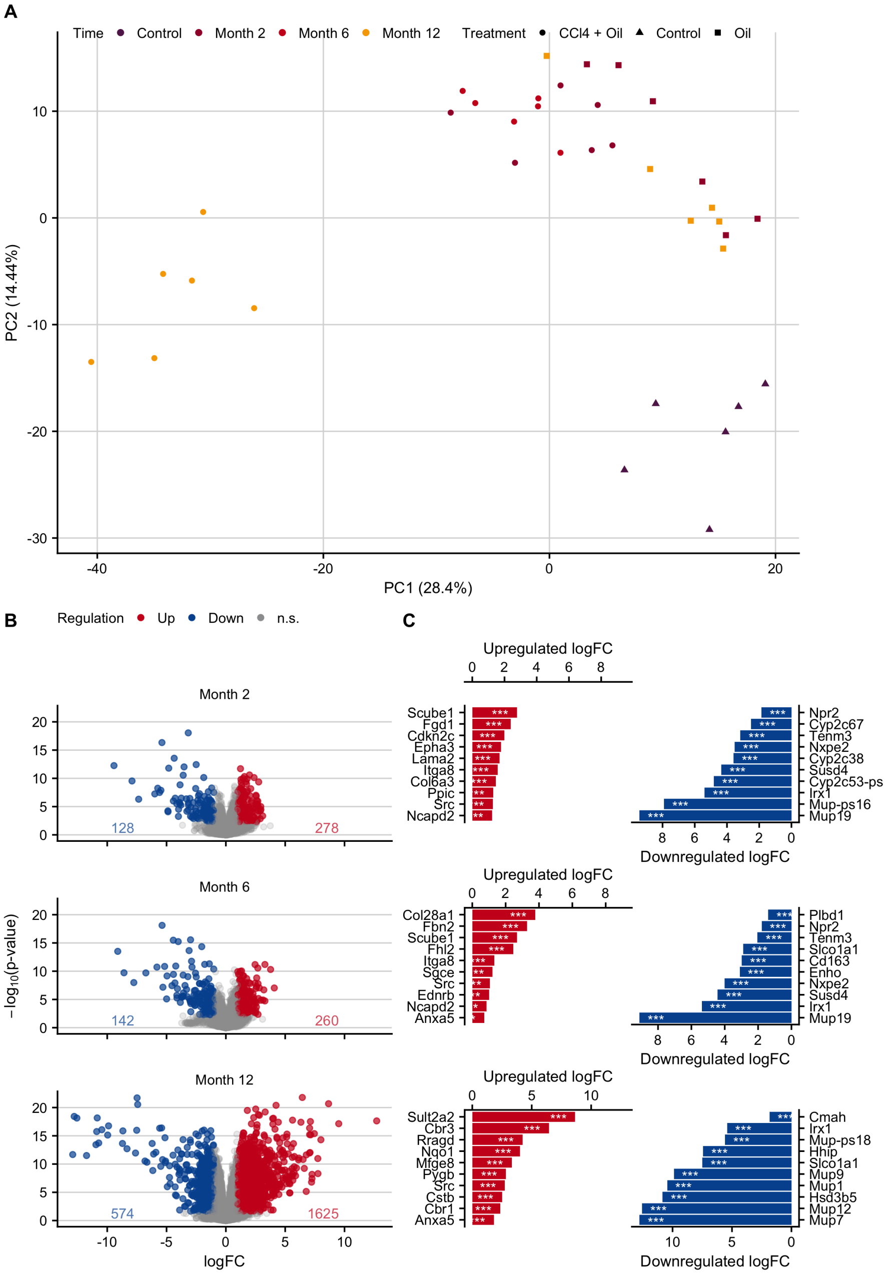


Figure C.2: Gene expression changes after induction of chronic liver damage in mice by repeated administration of 1 g/kg b.w. CCl 4 twice weekly for up to 12 months. A. PCA plot contextualized by time points and treatments. B. Volcano plots of months 2, 6 and 12. C. Genes with the highest log-fold changes.

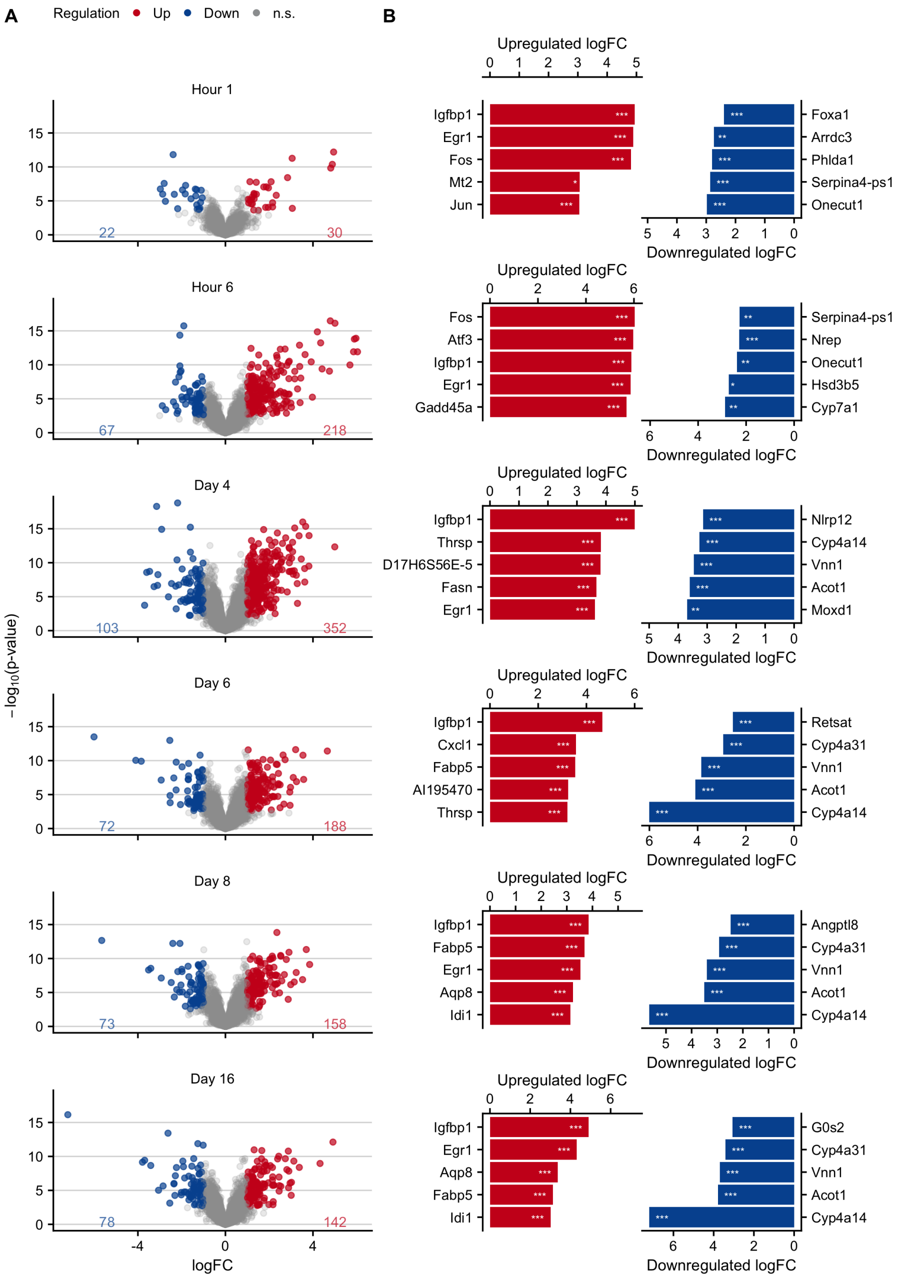


Figure C.3: Expression changes in an acute liver damage mouse model following a single administration of 300 mg/kg b.w. acetaminophen (APAP). A. Volcano plots at hours 1 and 6, and on days 4, 6, 8, 16 after APAP administration. B. Genes with the highest log-fold changes.

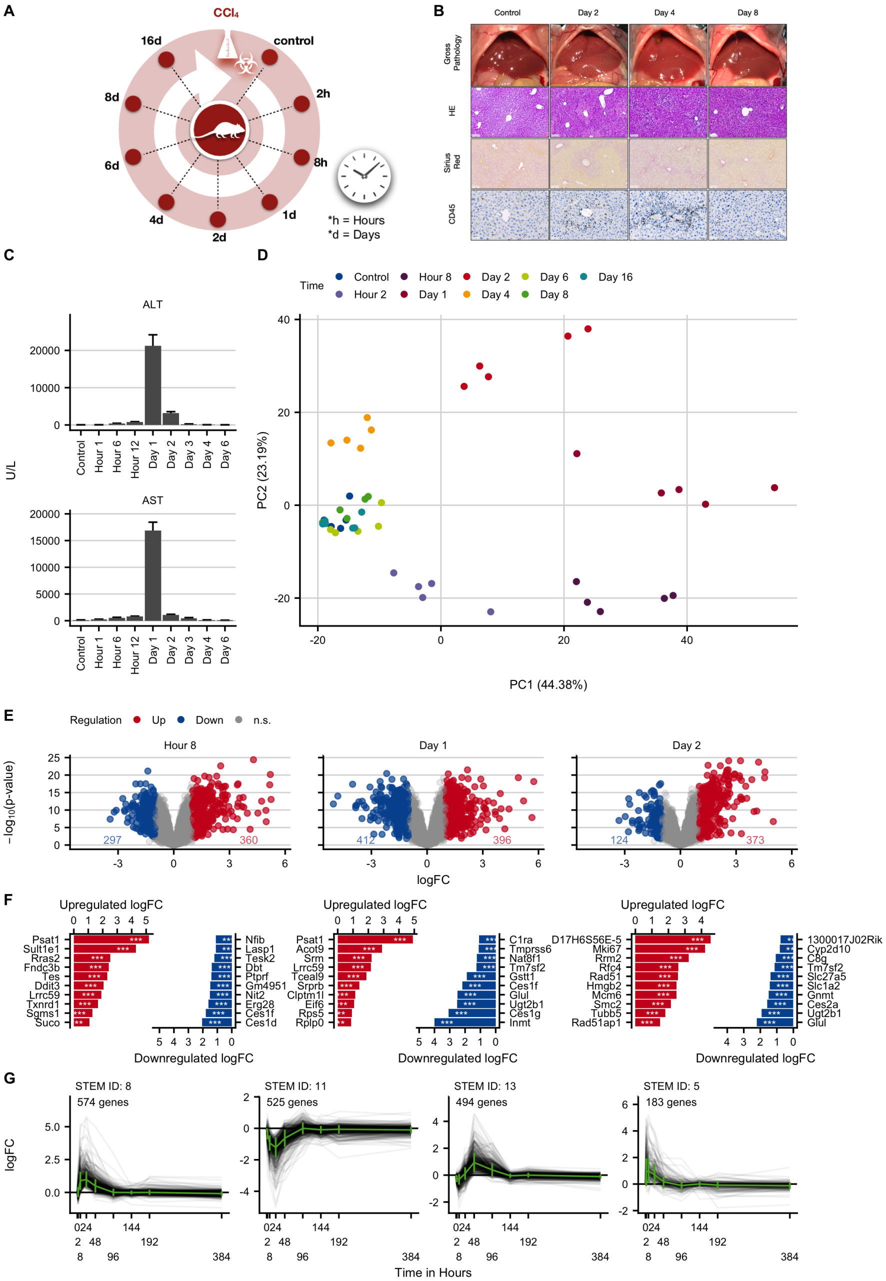


Figure C.4: Expression changes in an acute liver damage mouse model induced by administration of a single administration of CCl 4 . A. Experimental design. B. Histological analysis with hematoxylin and eosin (HE) staining, fibrosis grade visualized by Sirius red, and infiltration of immune cells by CD45. The images show induction of pericentral liver lobule damage but without fibrosis development. Scale bars: 100 µm (HE; Sirius red) and 50 µm (CD45). C. Clinical chemistry with alanine transaminase (ALT), aspartate transaminase (AST) activities in plasma. D. PCA analysis of global expression changes. E. Volcano plots at12 hours, and on days 1 and 2 after CCl 4 administration. F. Genes with the highest log-fold changes. G. Time-resolved clustering of deregulated genes. The panels B and C were provided by Ahmed Ghallab.

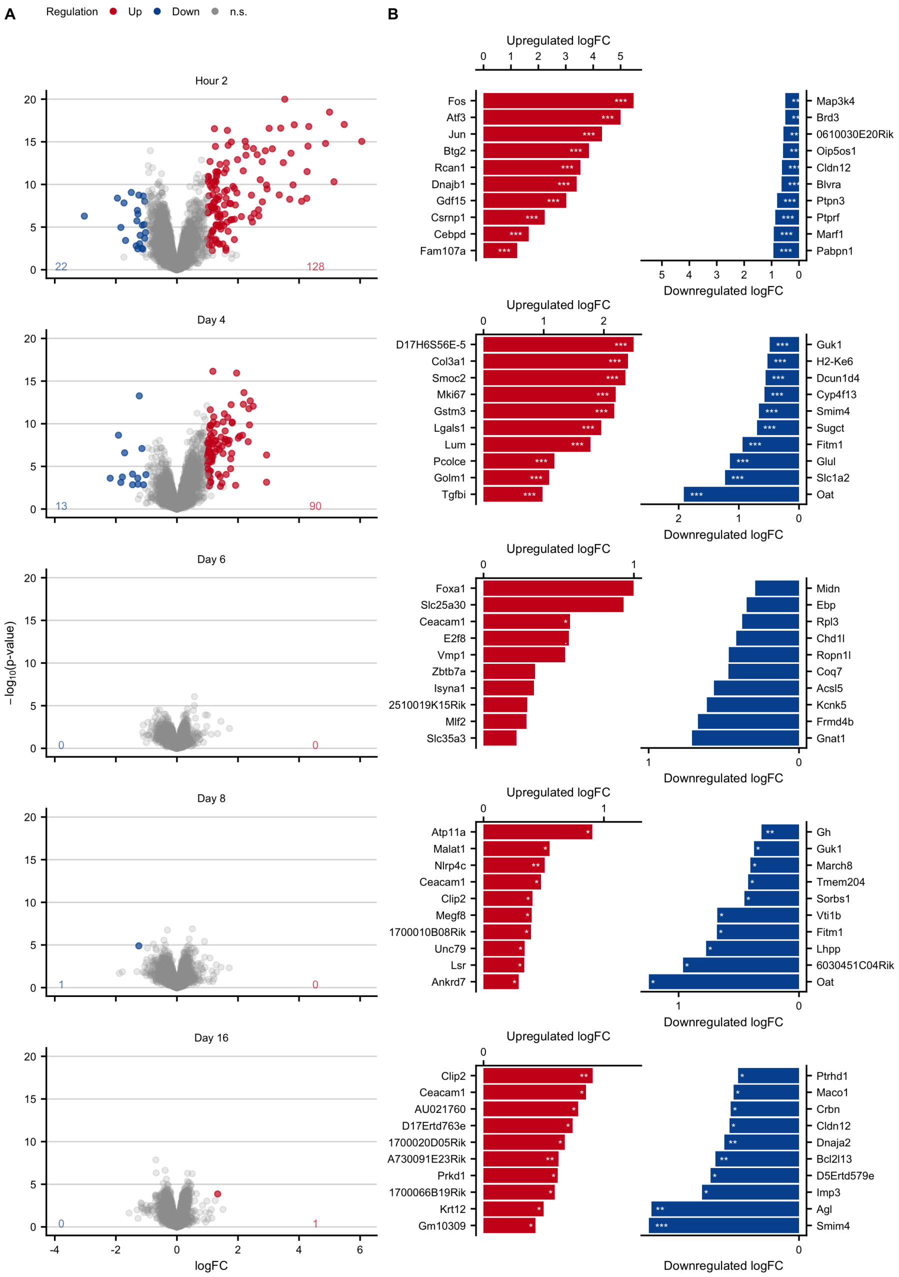


Figure C.5: Expression changes in an acute liver damage mouse model following a single administration of CCl 4 . A. Volcano plots at 2 hours and on days 4, 6, 8, 16 after CCl 4 administration. B. Genes with the highest log-fold changes.

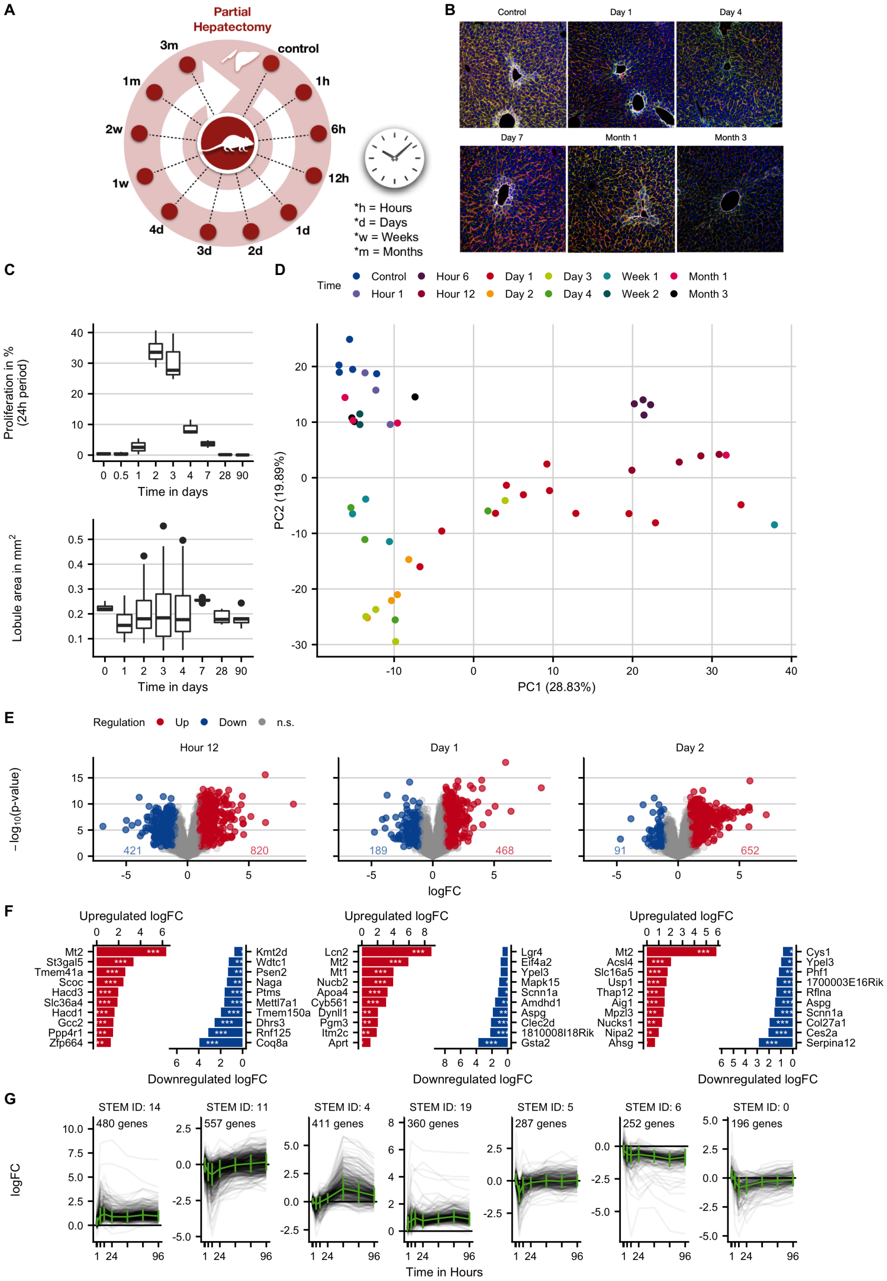


Figure C.6: Characterization and expression changes of mouse liver tissue after two-third hepatectomy. A. Experimental design. B. Transient increase in cell proliferation after hepatectomy based on BrDU staining and lobule area at different time periods after hepatectomy. C. Tissue morphology visualized by co-staining of bile canaliculi (DPPIV; green), pericentral hepatocytes (glutamine synthetase, white) and sinusoidal endothelial cells (yellow). D. PCA analysis of global expression changes. E. Volcano plots at 12 hours, and on days 1 and 2 after partial hepatectomy. F. Genes with the highest log-fold changes. G. Time-resolved clustering of deregulated genes. The panels B and C were provided by Ahmed Ghallab.

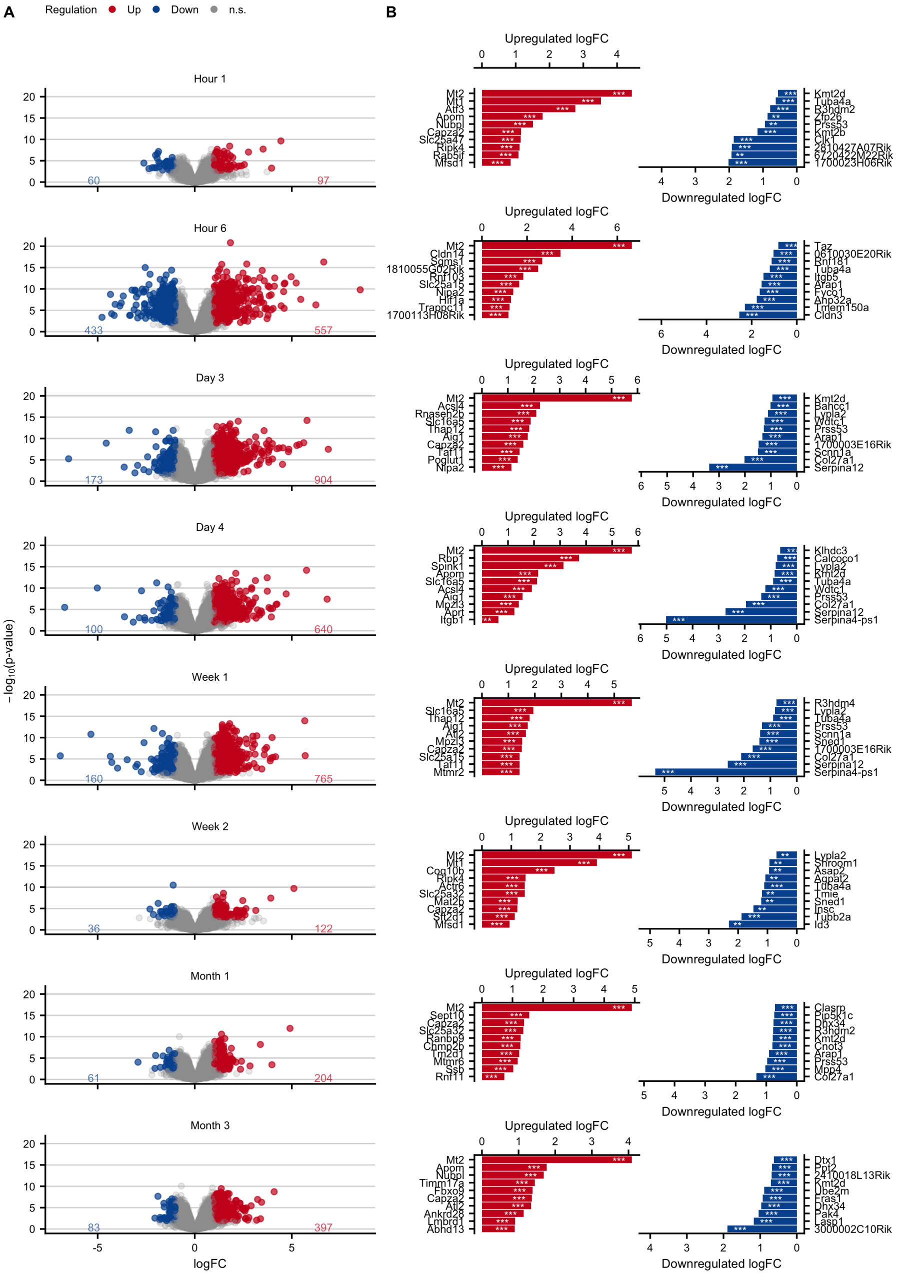


Figure C.7: Characterization and expression changes of mouse liver tissue after two-third hepatectomy. A. Volcano plots at hours 1 and 6, and on days 3 and 4, weeks 1 and 2, and months 1 and 3 after partial hepatectomy. B. Genes with the highest log-fold changes.

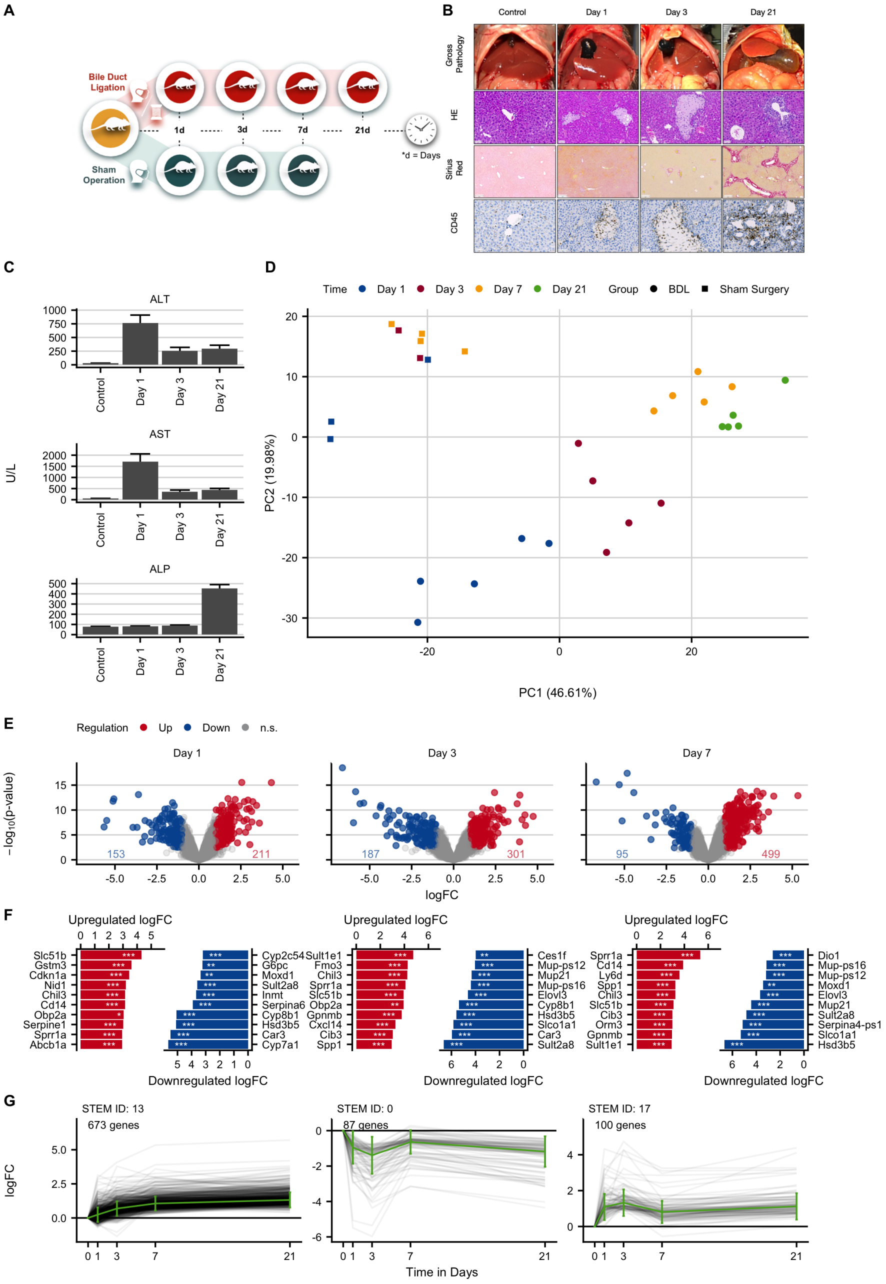


Figure C.8: Gene expression changes at different time intervals after induction of obstructive cholestasis by bile duct ligation (BDL). A. Experimental design. B. Histological analysis with hematoxylin and eosin (HE) staining, fibrosis grade visualized by Sirius red, and infiltration of immune cells by CD45. The images show bile infarct formation on days 1 and 3, and periportal fibrosis on day 21. Scale bars: 50 µm (HE; CD45) and 200 µm (Sirius red). C. Clinical chemistry with alanine transaminase (ALT), aspartate transaminase (AST) and alkaline phosphatase activities in plasma. D. PCA analysis of global expression changes. E. Volcano plots on days 1, 3 and 7 after BDL. F. Genes with the highest log-fold changes. G. Time-resolved clustering of deregulated genes. The panels B and C were provided by Ahmed Ghallab.

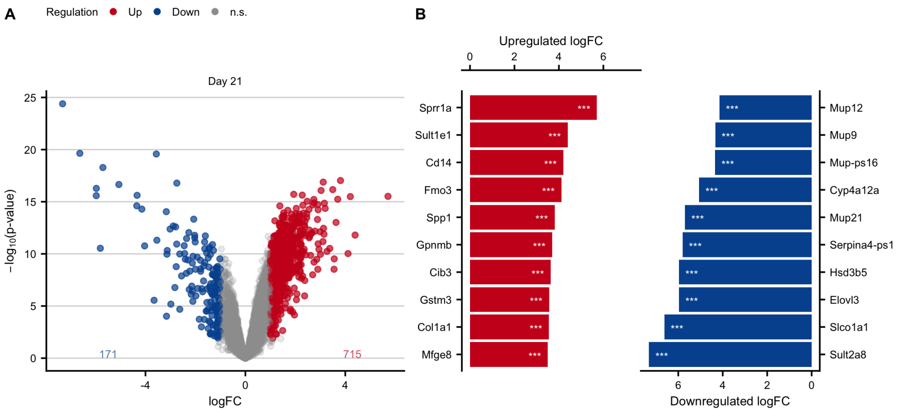


Figure C.9: Expression changes in the chronic stage after induction of obstructive cholestasis by bile duct ligation (BDL). A. Volcano plots on day 21 after BDL. B. Genes with the highest log-fold changes.

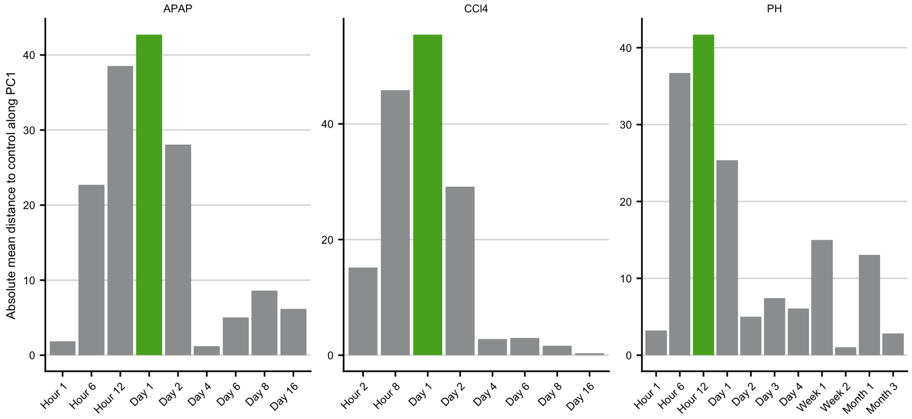


Figure C.10: Identification of the time point with the most deregulated expression profile after induction of acute liver injury based on the distance to the respective controls in PCA space along principle component 1 (PC1). Identified time point is colored in green.

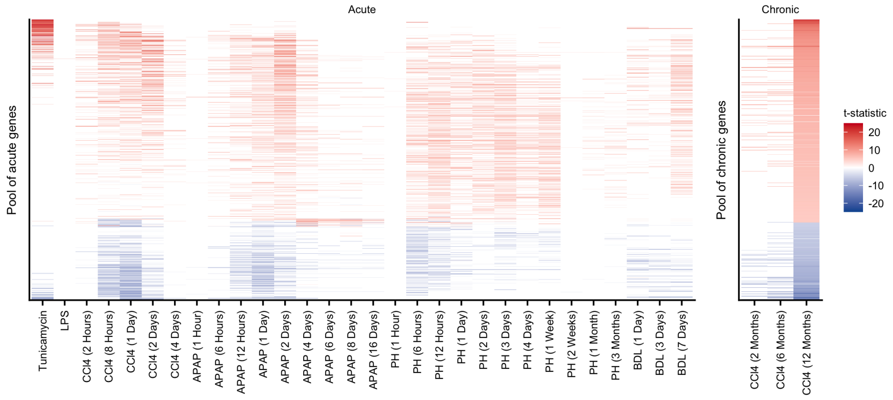


Figure C.11: Pool of unified acute and chronic genes demonstrating their respective consistent direction of regulation as indicated by t-statistic.

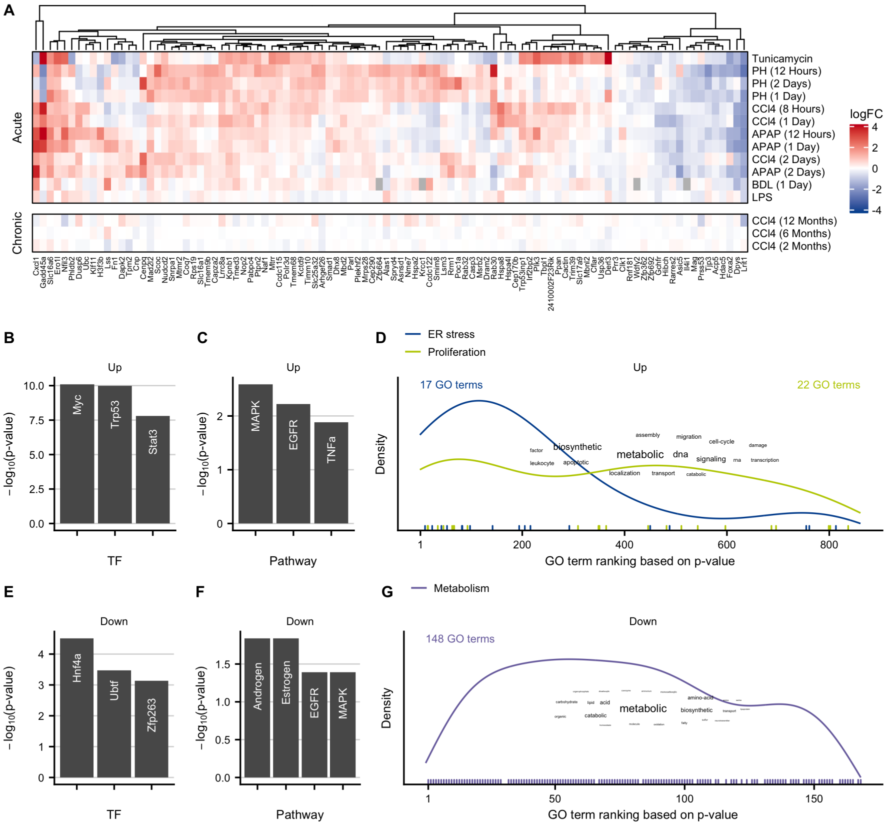


Figure C.12: Characterization of exclusive acute genes. A. Heatmap of top 100 exclusive acute genes. B-D. Overrepresented transcription factors identified by DoRothEA (B), pathways obtained by PROGENy (C), and GO terms (D) in the upregulated exclusive acute genes. E-F. Same as B-D but for the downregulated exclusive acute genes.

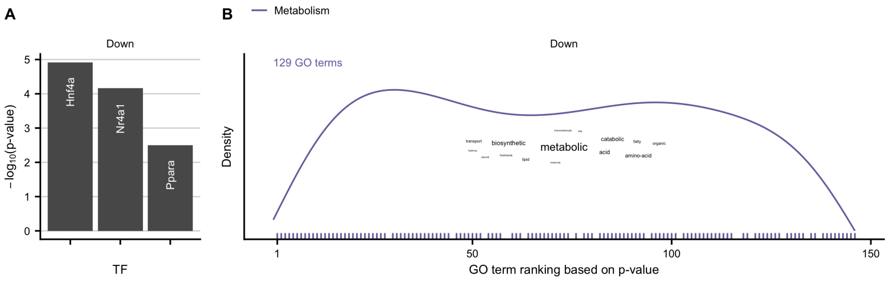
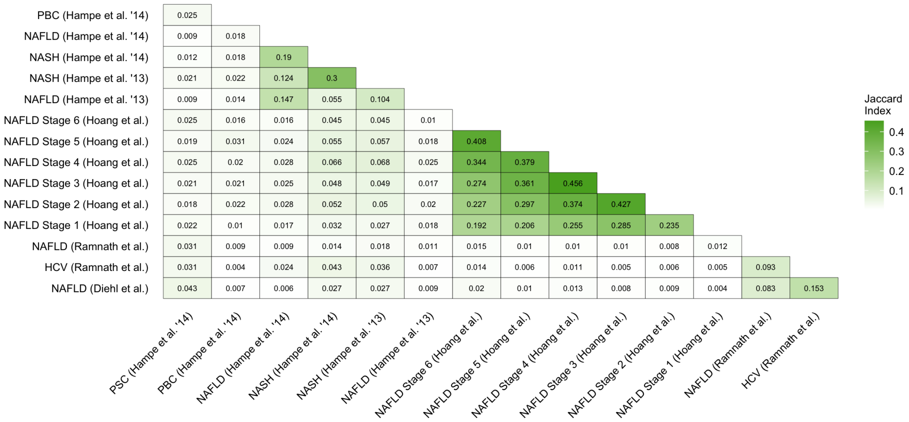


Figure C.13: A-B Overrepresented transcription factors identified by DoRothEA (A), and GO terms (B) in the set of commonly downregulated genes in acute and chronic mouse models.



	PSC (Hampe et al. '14)	PBC (Hampe et al. '14)	NAFLD (Hampe et al. '14)	NASH (Hampe et al. '14)	NASH (Hampe et al. '13)	NAFLD (Hampe et al. '13)	NAFLD Stage 6 (Hoang et al.)	NAFLD Stage 5 (Hoang et al.)	NAFLD Stage 4 (Hoang et al.)	NAFLD Stage 3 (Hoang et al.)	NAFLD Stage 2 (Hoang et al.)	NAFLD Stage 1 (Hoang et al.)	NAFLD (Ramnath et al.)	HCV (Ramnath et al.)
PBC (Hampe et al. '14)	0.025													
NAFLD (Hampe et al. '14)	0.009	0.018												
NASH (Hampe et al. '14)	0.012	0.018	0.19											
NASH (Hampe et al. '13)	0.021	0.022	0.124	0.3										
NAFLD (Hampe et al. '13)	0.009	0.014	0.147	0.055	0.104									
NAFLD Stage 6 (Hoang et al.)	0.025	0.016	0.016	0.045	0.045	0.01								
NAFLD Stage 5 (Hoang et al.)	0.019	0.031	0.024	0.055	0.057	0.018	0.408							
NAFLD Stage 4 (Hoang et al.)	0.025	0.02	0.028	0.066	0.068	0.025	0.344	0.379						
NAFLD Stage 3 (Hoang et al.)	0.021	0.021	0.025	0.048	0.049	0.017	0.274	0.361	0.456					
NAFLD Stage 2 (Hoang et al.)	0.018	0.022	0.028	0.052	0.05	0.02	0.227	0.297	0.374	0.427				
NAFLD Stage 1 (Hoang et al.)	0.022	0.01	0.017	0.032	0.027	0.018	0.192	0.206	0.255	0.285	0.235			
NAFLD (Ramnath et al.)	0.031	0.009	0.009	0.014	0.018	0.011	0.015	0.01	0.01	0.01	0.008	0.012		
HCV (Ramnath et al.)	0.031	0.004	0.024	0.043	0.036	0.007	0.014	0.006	0.011	0.005	0.006	0.005	0.093	
NAFLD (Diehl et al.)	0.043	0.007	0.006	0.027	0.027	0.009	0.02	0.01	0.013	0.008	0.009	0.004	0.083	0.153

Figure C.14: Pairwise comparison of the similarity of the 500 deregulated genes per human contrast. Similarity is computed with the Jaccard index.

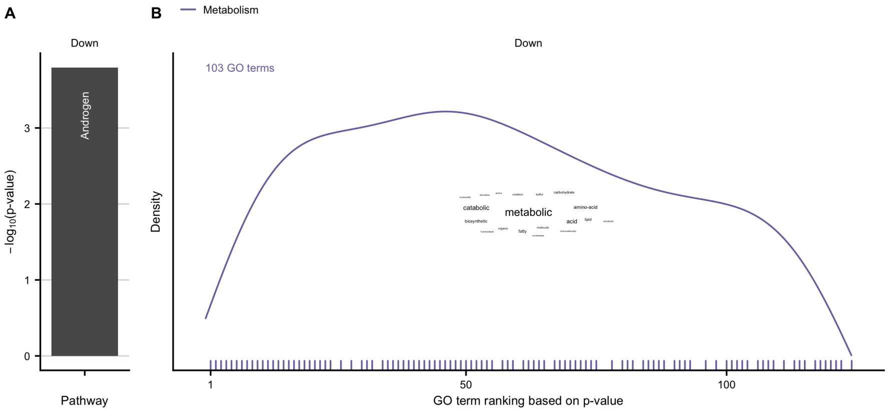


Figure C.15: A-B Overrepresented pathways obtained by PROGENy (A), and GO terms (B) in the set of consistently downregulated genes between human and mouse.

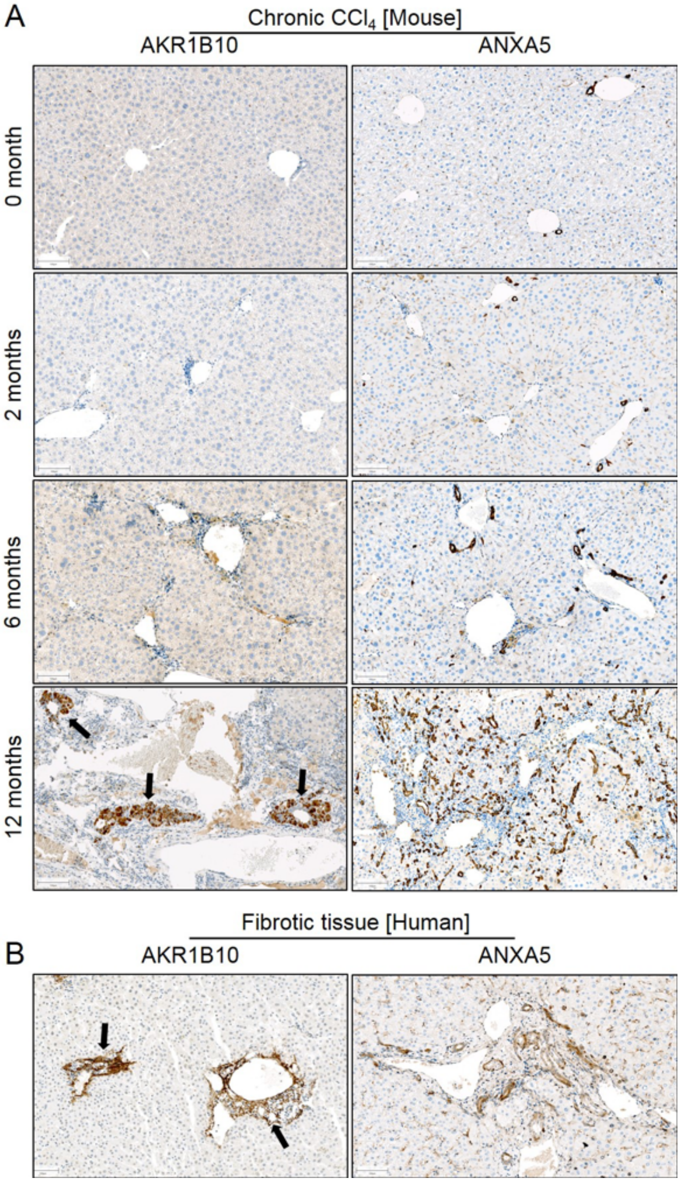


Figure C.16: Aldo-keto reductase (AKR1B10) and annexin V (ANXA5) increase in CLD of mice and humans. A. Immunostaining of AKR1B10 and ANXA5 at different stages during CLD progression induced by chronic CCl 4 administration in mice; AKR1B10 shows clusters of positive signals at 12 months (arrows); ANXA5 stained positive in the progressive ductular reaction. B. AKR1B10 and ANXA5 immunostaining in fibrotic tissues of human patients showing similar expression patterns as in mice. Scale bars: 100 µm. The entire figure was provided by Ahmed Ghallab.

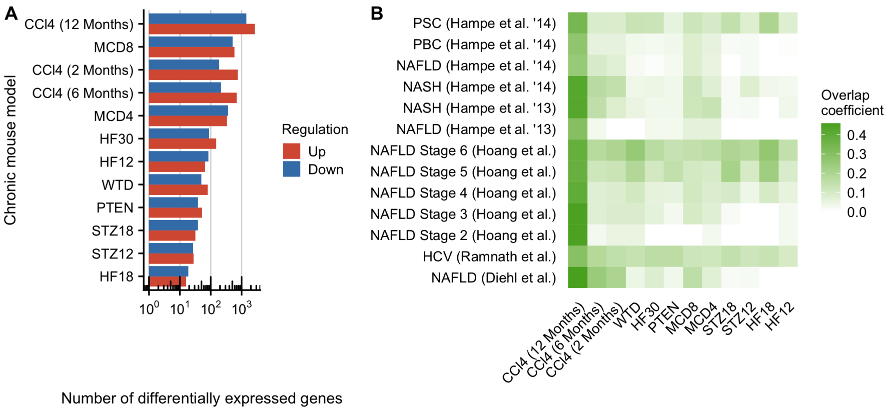


Figure C.17: A. Number of deregulated genes for the chronic CCl 4 mouse models complemented by 9 additional publicly available mouse models of chronic liver diseases. B Similarity of the significantly deregulated genes between all chronic mouse models and the human data. Similarity is computed as overlap coefficient.

References

Ahrens, M., Ammerpohl, O., Schönfels, W. von, Kolarova, J., Bens, S., Itzel, T., … Hampe, J. (2013). DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery. *Cell Metabolism*, *18*(2), 296–302. http://doi.org/10.1016/j.cmet.2013.07.004

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., … Aerts, S. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, *14*(11), 1083–1086. http://doi.org/10.1038/nmeth.4463

Altelaar, A. F. M., Munoz, J., & Heck, A. J. R. (2013). Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nature Reviews. Genetics*, *14*(1), 35–48. http://doi.org/10.1038/nrg3356

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., & Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, *48*(8), 838–847. http://doi.org/10.1038/ng.3593

Alvarez, M. J., Subramaniam, P. S., Tang, L. H., Grunn, A., Aburi, M., Rieckhof, G., … Califano, A. (2018). A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nature Genetics*, *50*(7), 979–989. http://doi.org/10.1038/s41588-018-0138-4

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. http://doi.org/10.1186/s13059-020-1935-5

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. http://doi.org/10.1038/75556

Aytes, A., Mitrofanova, A., Lefebvre, C., Alvarez, M. J., Castillo-Martin, M., Zheng, T., … Abate-Shen, C. (2014). Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell*, *25*(5), 638–651.

http://doi.org/10.1016/j.ccr.2014.03.017

Banerjee, S., Biehl, A., Gadina, M., Hasni, S., & Schwartz, D. M. (2017). JAK-STAT signaling as a target for inflammatory and autoimmune diseases: Current and future prospects. *Drugs*, *77*(5), 521–546. http://doi.org/10.1007/s40265-017-0701-9

Baran-Gale, J., Chandra, T., & Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Briefings in Functional Genomics*, *17*(4), 233–239. http://doi.org/10.1093/bfgp/elx035

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., … Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, *462*(7269), 108–112. http://doi.org/10.1038/nature08460

Barbosa, S., Niebel, B., Wolf, S., Mauch, K., & Takors, R. (2018). A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints. *Bio Systems*, *174*, 37–48. http://doi.org/10.1016/j.biosystems.2018.10.008

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *57*, 289–300.

Blischak, J. D., Carbonetto, P., & Stephens, M. (2019). Creating and sharing reproducible research code the workflowr way. *F1000Research*, *8*, 1749. http://doi.org/10.12688/f1000research.20843.1

Blouin, A., Bolender, R. P., & Weibel, E. R. (1977). Distribution of organelles and membranes between hepatocytes and nonhepatocytes in the rat liver parenchyma. A stereological study. *The Journal of Cell Biology*, *72*(2), 441–455. http://doi.org/10.1083/jcb.72.2.441

Boyer, J. L. (2013). Bile formation and secretion. *Comprehensive Physiology*, *3*(3), 1035–1078. http://doi.org/10.1002/cphy.c120027

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. http://doi.org/10.3322/caac.21492

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. http://doi.org/10.1038/nbt.3519

Brubaker, D. K., Proctor, E. A., Haigis, K. M., & Lauffenburger, D. A. (2019). Computational translation of genomic responses from experimental model systems to humans. *PLoS Computational Biology*, *15*(1), e1006286.

http://doi.org/10.1371/journal.pcbi.1006286

Burd, A. L., Ingraham, R. H., Goldrick, S. E., Kroe, R. R., Crute, J. J., & Grygon, C. A. (2004). Assembly of major histocompatibility complex (MHC) class II transcription factors: Association and promoter recognition of RFX proteins. *Biochemistry*, *43*(40), 12750–12760. http://doi.org/10.1021/bi030262o

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. http://doi.org/10.1038/nbt.4096

Campos, G., Schmidt-Heck, W., De Smedt, J., Widera, A., Ghallab, A., Pütter, L., ... Godoy, P. (2020). Inflammation-associated suppression of metabolic gene networks in acute and chronic liver disease. *Archives of Toxicology*, *94*(1), 205–217. http://doi.org/10.1007/s00204-019-02630-3

Cantini, L., Calzone, L., Martignetti, L., Rydenfelt, M., Blüthgen, N., Barillot, E., & Zinovyev, A. (2018). Classification of gene signatures for their information value and functional redundancy. *NPJ Systems Biology and Applications*, *4*, 2. http://doi.org/10.1038/s41540-017-0038-8

Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., ... Shendure, J. (2020). A human cell atlas of fetal gene expression. *Science*, *370*(6518). http://doi.org/10.1126/science.aba7721

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., ... Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, *357*(6352), 661–667. http://doi.org/10.1126/science.aam8940

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., ... Consortium, G. (2015). A novel approach to high-quality postmortem tissue procurement: The GTEx project. *Biopreservation and Biobanking*, *13*(5), 311–319. http://doi.org/10.1089/bio.2015.0032

Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, *26*(19), 2363–2367. http://doi.org/10.1093/bioinformatics/btq431

Chen, J., Gao, G., Wang, H., Ye, X., Zhou, J., & Lin, J. (2019). Expression and clinical significance of latent-transforming growth factor beta-binding protein 2 in primary hepatocellular carcinoma. *Medicine*, *98*(39), e17216. http://doi.org/10.1097/{MD}.0000000000017216

Consortium, G. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. http://doi.org/10.1038/ng.2653

Costache, M. I., Ioana, M., Iordache, S., Ene, D., Costache, C. A., & Săftoiu, A. (2015). VEGF expression in pancreatic cancer and other malignancies: A review of the literature. *Romanian Journal of Internal Medicine = Revue Roumaine de Medecine Interne*, *53*(3), 199–208. http://doi.org/10.1515/rjim-2015-0027

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. http://doi.org/10.1038/227561a0

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning - ICML '06* (pp. 233–240). New York, New York, USA: ACM Press. http://doi.org/10.1145/1143844.1143874

Davis, S., & Meltzer, P. S. (2007). GEOquery: A bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics*, *23*(14), 1846–1847. http://doi.org/10.1093/bioinformatics/btm254

Ding, H., Douglass, E. F., Sonabend, A. M., Mela, A., Bose, S., Gonzalez, C., … Califano, A. (2018). Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nature Communications*, *9*(1), 1471. http://doi.org/10.1038/s41467-018-03843-3

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., … Regev, A. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, *167*(7), 1853–1866.e17. http://doi.org/10.1016/j.cell.2016.11.038

Dobie, R., Wilson-Kanamori, J. R., Henderson, B. E. P., Smith, J. R., Matchett, K. P., Portman, J. R., … Henderson, N. C. (2019). Single-cell transcriptomics uncovers zonation of function in the mesenchyme during liver fibrosis. *Cell Reports*, *29*(7), 1832–1847.e8. http://doi.org/10.1016/j.celrep.2019.10.024

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., … Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, *17*(10), 1537–1545. http://doi.org/10.1101/gr.6202607

Dugourd, A., & Saez-Rodriguez, J. (2019). Footprint-based functional analysis of multiomic data. *Current Opinion in Systems Biology*, *15*, 82–90. http://doi.org/10.1016/j.coisb.2019.04.002

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. http://doi.org/10.1038/nprot.2009.97

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. http://doi.org/10.1093/nar/30.1.207

Ernst, J., & Bar-Joseph, Z. (2006). STEM: A tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, *7*, 191. http://doi.org/10.1186/1471-2105-7-191

Ernst, J., Nau, G. J., & Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, *21 Suppl 1*, i159–68. http://doi.org/10.1093/bioinformatics/bti1022

Essaghir, A., Toffalini, F., Knoops, L., Kallin, A., Helden, J. van, & Demoulin, J.-B. (2010). Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Research*, *38*(11), e120. http://doi.org/10.1093/nar/gkq149

Feng, X., Wang, H., Takata, H., Day, T. J., Willen, J., & Hu, H. (2011). Transcription factor Foxp1 exerts essential cell-intrinsic regulation of the quiescence of naive t cells. *Nature Immunology*, *12*(6), 544–550. http://doi.org/10.1038/ni.2034

Fisher, R. A. (1992). Statistical methods for research workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 66–70). New York: Springer. http://doi.org/10.1007/978-1-4612-4380-9_6

Foroutan, M., Bhuva, D. D., Lyu, R., Horan, K., Cursons, J., & Davis, M. J. (2018). Single sample scoring of molecular phenotypes. *BMC Bioinformatics*, *19*(1), 404. http://doi.org/10.1186/s12859-018-2435-4

Fox, J. G., Barthold, S., Newcomer, C. E., Smith, A., & Quimby, F. W. (2006). *The mouse in biomedical research* (2nd ed., p. 2192). Amsterdam: Academic Pr.

Fruman, D. A., & Rommel, C. (2014). PI3K and cancer: Lessons, challenges and opportunities. *Nature Reviews. Drug Discovery*, *13*(2), 140–156. http://doi.org/10.1038/nrd4204

Fry, E. A., & Inoue, K. (2018). Aberrant expression of ETS1 and ETS2 proteins in cancer. *Cancer Reports and Reviews*, *2*(3). http://doi.org/10.15761/{CRR}.1000151

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, *29*(8), 1363–1375. http://doi.org/10.1101/gr.240663.118

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., … Saez-Rodriguez, J. (2018). Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Research*, *78*(3), 769–780. http://doi.org/10.1158/0008-5472.{CAN}-17-1679

Genga, R. M. J., Kernfeld, E. M., Parsi, K. M., Parsons, T. J., Ziller, M. J., & Maehr, R. (2019). Single-cell RNA-sequencing-based CRISPRi screening resolves molecular drivers of early human endoderm development. *Cell Reports*, *27*(3), 708–718.e10. http://doi.org/10.1016/j.celrep.2019.03.076

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. http://doi.org/10.1186/gb-2004-5-10-r80

Ghallab, A., Hofmann, U., Sezgin, S., Vartak, N., Hassan, R., Zaza, A., … Reif, R. (2019). Bile microinfarcts in cholestasis are initiated by rupture of the apical hepatocyte membrane and cause shunting of bile to sinusoidal blood. *Hepatology*, *69*(2), 666–683. http://doi.org/10.1002/hep.30213

Ghallab, A., Myllys, M., Holland, C. H., Zaza, A., Murad, W., Hassan, R., … Hengstler, J. G. (2019). Influence of liver fibrosis on lobular zonation. *Cells*, *8*(12). http://doi.org/10.3390/cells8121556

Godoy, P., Widera, A., Schmidt-Heck, W., Campos, G., Meyer, C., Cadenas, C., … Hengstler, J. G. (2016). Gene network activity in cultivated primary hepatocytes is highly similar to diseased mammalian liver tissue. *Archives of Toxicology*, *90*(10), 2513–2529. http://doi.org/10.1007/s00204-016-1761-4

Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: Computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, *31*(15), 2595–2597. http://doi.org/10.1093/bioinformatics/btv153

Greenbaum, D., Colangelo, C., Williams, K., & Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, *4*(9), 117. http://doi.org/10.1186/gb-2003-4-9-117

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., … Lee, I. (2018). TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1), D380–D386. http://doi.org/10.1093/nar/gkx1013

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, *14*, 7. http://doi.org/10.1186/1471-2105-14-7

Hegde, M., Strand, C., Hanna, R. E., & Doench, J. G. (2018). Uncoupling of sgRNAs from their associated barcodes during PCR amplification of combinatorial CRISPR screens. *Plos One*, *13*(5), e0197547. http://doi.org/10.1371/journal.pone.0197547

Henderson, N. C., Rieder, F., & Wynn, T. A. (2020). Fibrosis: From mechanisms to medicines. *Nature*, *587*(7835), 555–566. http://doi.org/10.1038/s41586-020-2938-9

Hernandez-Armenta, C., Ochoa, D., Gonçalves, E., Saez-Rodriguez, J., & Beltrao, P. (2017). Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, *33*(12), 1845–1851. http://doi.org/10.1093/bioinformatics/btx082

Hernansaiz-Ballesteros, R., Holland, C. H., Dugourd, A., & Saez-Rodriguez, J. (2021). FUNKI: Interactive functional footprint-based analysis of omics data. Retrieved from http://arxiv.org/abs/2109.05796

Hidalgo, M. R., Cubuk, C., Amadoz, A., Salavert, F., Carbonell-Caballero, J., & Dopazo, J. (2017). High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget*, *8*(3), 5160–5178. http://doi.org/10.18632/oncotarget.14107

Hoang, S. A., Oseini, A., Feaver, R. E., Cole, B. K., Asgharpour, A., Vincent, R., … Sanyal, A. J. (2019). Gene expression predicts histological severity and reveals distinct molecular profiles of nonalcoholic fatty liver disease. *Scientific Reports*, *9*(1), 12541. http://doi.org/10.1038/s41598-019-48746-5

Hoheisel, J. D. (2006). Microarray technology: Beyond transcript profiling and genotype analysis. *Nature Reviews. Genetics*, *7*(3), 200–210. http://doi.org/10.1038/nrg1809

Holland, Christian H., Ramirez Flores, R. O., Myllys, M., Hassan, R., Edlund, K., Hofmann, U., … Ghallab, A. (2021). Transcriptomic cross-species analysis of chronic liver disease reveals consistent regulation between humans and mice. *Hepatology Communications*. http://doi.org/10.1002/hep4.1797

Holland, Christian H., Szalai, B., & Saez-Rodriguez, J. (2020). Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms*, *1863*(6), 194431. http://doi.org/10.1016/j.bbagrm.2019.194431

Holland, Christian H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M. P., Mereu, E., … Saez-Rodriguez, J. (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biology*, *21*(1), 36. http://doi.org/10.1186/s13059-020-1949-z

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., Schönfels, W. von, Ahrens, M., … Hampe, J. (2014). Obesity accelerates epigenetic aging of human liver. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(43), 15538–15543. http://doi.org/10.1073/pnas.1412759111

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. http://doi.org/10.1038/nprot.2008.211

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., & DeLisi, C. (2012). Gene set enrichment analysis: Performance evaluation and usage guidelines. *Briefings in*

Bioinformatics, *13*(3), 281–291. http://doi.org/10.1093/bib/bbr049

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., … Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, *292*(5518), 929–934. http://doi.org/10.1126/science.292.5518.929

J, S., B, T., L, C., & Parkinson, H. (2015). A new ontology lookup service at EMBL-EBI. In *SWAT4LS*.

Jansen, P. L. M., Ghallab, A., Vartak, N., Reif, R., Schaap, F. G., Hampe, J., & Hengstler, J. G. (2017). The ascending pathophysiology of cholestatic liver disease. *Hepatology*, *65*(2), 722–738. http://doi.org/10.1002/hep.28965

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. http://doi.org/10.1093/nar/gkz1031

Johnson, D. G. (2000). The paradox of E2F1: Oncogene and tumor suppressor gene. *Molecular Carcinogenesis*, *27*(3), 151–157. http://doi.org/10.1002/({SICI})1098-2744(200003)27:3\textless151::{AID}-{MC1\textgreater3}.0.{CO};2-C

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. http://doi.org/10.1093/nar/28.1.27

Kauffmann, A., Rayner, T. F., Parkinson, H., Kapushesky, M., Lukk, M., Brazma, A., & Huber, W. (2009). Importing ArrayExpress datasets into r/bioconductor. *Bioinformatics*, *25*(16), 2092–2094. http://doi.org/10.1093/bioinformatics/btp354

Keenan, A. B., Torre, D., Lachmann, A., Leong, A. K., Wojciechowicz, M. L., Utti, V., … Ma'ayan, A. (2019). ChEA3: Transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Research*, *47*(W1), W212–W224. http://doi.org/10.1093/nar/gkz446

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, *11*(7), 740–742. http://doi.org/10.1038/nmeth.2967

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2), e1002375. http://doi.org/10.1371/journal.pcbi.1002375

Kim, A., Wu, X., Allende, D. S., & Nagy, L. E. (2021). Gene deconvolution reveals aberrant liver regeneration and immune cell infiltration in alcohol-associated hepatitis. *Hepatology*. http://doi.org/10.1002/hep.31759

Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J., & Smith, V. A. (2015). Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Scientific Reports*, *5*, 10775.

http://doi.org/10.1038/srep10775

Krämer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, *30*(4), 523–530. http://doi.org/10.1093/bioinformatics/btt703

Krenkel, O., Hundertmark, J., Ritz, T. P., Weiskirchen, R., & Tacke, F. (2019). Single cell RNA sequencing identifies subsets of hepatic stellate cells and myofibroblasts in liver fibrosis. *Cells*, *8*(5). http://doi.org/10.3390/cells8050503

Kuppe, C., Ramirez Flores, R. O., Li, Z., Hannani, M. T., Tanevski, J., Halder, M., … Kramann, R. (2020). Spatial multi-omic map of human myocardial infarction. *BioRxiv*. http://doi.org/10.1101/2020.12.08.411686

Kwon, A. T., Arenillas, D. J., Worsley Hunt, R., & Wasserman, W. W. (2012). oPO-3: Advanced analysis of regulatory motif over-representation across genes or ChIP-seq datasets. *G3 (Bethesda, Md.)*, *2*(9), 987–1002. http://doi.org/10.1534/g3.112.003202

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., … Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, *9*(1), 1366. http://doi.org/10.1038/s41467-018-03751-6

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., … Golub, T. R. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, *313*(5795), 1929–1935. http://doi.org/10.1126/science.1132939

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., & Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, *4*(11), e1000217. http://doi.org/10.1371/journal.pcbi.1000217

Leist, M., & Hartung, T. (2013). Inflammatory findings on species extrapolations: Humans are definitely no 70-kg mice. *Archives of Toxicology*, *87*(4), 563–567. http://doi.org/10.1007/s00204-013-1038-0

Li, L., & Clevers, H. (2010). Coexistence of quiescent and active adult stem cells in mammals. *Science*, *327*(5965), 542–545. http://doi.org/10.1126/science.1180794

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. http://doi.org/10.1093/bioinformatics/btr260

Liu, A., Trairatphisan, P., Gjerga, E., Didangelos, A., Barratt, J., & Saez-Rodriguez, J. (2019). From expression footprints to causal pathways: Contextualizing large signaling networks with CARNIVAL. *NPJ Systems Biology and Applications*, *5*, 40. http://doi.org/10.1038/s41540-019-0118-z

Liu, T., Zhang, L., Joo, D., & Sun, S.-C. (2017). NF- signaling in inflammation. *Signal Transduction and Targeted Therapy*, *2*. http://doi.org/10.1038/sigtrans.2017.23

Llovet, J. M., Kelley, R. K., Villanueva, A., Singal, A. G., Pikarsky, E., Roayaie, S., … Finn, R. S. (2021). Hepatocellular carcinoma. *Nature Reviews. Disease Primers*, *7*(1), 6. http://doi.org/10.1038/s41572-020-00240-3

Lopez-Dominguez, R., Toro-Dominguez, D., Martorell-Marugan, J., Garcia-Moreno, A., Holland, C. H., Saez-Rodriguez, J., … Carmona-Saez, P. (2021). Transcription factor activity inference in systemic lupus erythematosus. *Life (Chicago, Ill. : 1978)*, *11*(4), 299. http://doi.org/10.3390/life11040299

Lun, A. T. L., McCarthy, D. J., & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Research*, *5*, 2122. http://doi.org/10.12688/f1000research.9501.2

Mahi, N. A., Najafabadi, M. F., Pilarczyk, M., Kouril, M., & Medvedovic, M. (2019). GREIN: An interactive web platform for re-analyzing GEO RNA-seq data. *Scientific Reports*, *9*(1), 7580. http://doi.org/10.1038/s41598-019-43935-8

Mann, M., & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology*, *21*(3), 255–261. http://doi.org/10.1038/nbt0303-255

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7 Suppl 1*, S7. http://doi.org/10.1186/1471-2105-7-S1-S7

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Álvarez-Varela, A., … Heyn, H. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology*, *38*(6), 747–755. http://doi.org/10.1038/s41587-020-0469-4

Method of the year 2013. (2014). *Nature Methods*, *11*(1), 1. http://doi.org/10.1038/nmeth.2801

Method of the year 2020: Spatially resolved transcriptomics. (n.d.). *Nature Methods*, *18*(1), 1. http://doi.org/10.1038/s41592-020-01042-x

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*(D1), D419–D426. http://doi.org/10.1093/nar/gky1038

Michel, K., Roth, S., Trautwein, C., Gong, W., Flemming, P., & Gressner, A. M. (1998). Analysis of the expression pattern of the latent transforming growth factor beta binding protein isoforms in normal and diseased human liver reveals a new splice variant missing the proteinase-sensitive hinge region. *Hepatology*, *27*(6), 1592–1599. http://doi.org/10.1002/hep.510270619

Mitchell, C., & Willenbring, H. (2008). A reproducible and well-tolerated method for 2/3 partial hepatectomy in mice. *Nature Protocols*, *3*(7), 1167–1170. http://doi.org/10.1038/nprot.2008.80

Mohs, A., Otto, T., Schneider, K. M., Peltzer, M., Boekschoten, M., Holland, C. H., … Trautwein, C. (2020). Hepatocyte-specific NRF2 activation controls fibrogenesis and carcinogenesis in steatohepatitis. *Journal of Hepatology*. http://doi.org/10.1016/j.jhep.2020.09.037

Moylan, C. A., Pang, H., Dellinger, A., Suzuki, A., Garrett, M. E., Guy, C. D., … Diehl, A. M. (2014). Hepatic gene expression profiles differentiate presymptomatic patients with mild versus severe nonalcoholic fatty liver disease. *Hepatology*, *59*(2), 471–482. http://doi.org/10.1002/hep.26661

Network, C. G. A. R., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., … Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. http://doi.org/10.1038/ng.2764

Nguyen, T.-M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: A comprehensive review and assessment. *Genome Biology*, *20*(1), 203. http://doi.org/10.1186/s13059-019-1790-4

Normand, R., Du, W., Briller, M., Gaujoux, R., Starosvetsky, E., Ziv-Kenet, A., … Shen-Orr, S. S. (2018). Found in translation: A machine learning model for mouse-to-human inference. *Nature Methods*, *15*(12), 1067–1073. http://doi.org/10.1038/s41592-018-0214-9

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., … Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, *22*(16), 2028–2036. http://doi.org/10.1093/bioinformatics/btl344

Parikh, J. R., Klinger, B., Xia, Y., Marto, J. A., & Blüthgen, N. (2010). Discovering causal signaling pathways through gene-expression patterns. *Nucleic Acids Research*, *38*(Web Server issue), W109–17. http://doi.org/10.1093/nar/gkq424

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: The apogee of the omics trilogy. *Nature Reviews. Molecular Cell Biology*, *13*(4), 263–269. http://doi.org/10.1038/nrm3314

Pellicoro, A., Ramachandran, P., Iredale, J. P., & Fallowfield, J. A. (2014). Liver fibrosis and repair: Immune regulation of wound healing in a solid organ. *Nature Reviews. Immunology*, *14*(3), 181–194. http://doi.org/10.1038/nri3623

Peng, T., Zhu, Q., Yin, P., & Tan, K. (2019). SCRABBLE: Single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biology*, *20*(1), 88. http://doi.org/10.1186/s13059-019-1681-8

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., & Smyth, G. K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics*, *10*(2), 946–963. http://doi.org/10.1214/16-{AOAS920}

Puente-Santamaria, L., Wasserman, W. W., & Del Peso, L. (2019). TFEA.ChIP: A tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets. *Bioinformatics*, *35*(24), 5339–5340. http://doi.org/10.1093/bioinformatics/btz573

Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, *11*(1), 1169. http://doi.org/10.1038/s41467-020-14976-9

Quiñonez-Flores, C. M., González-Chávez, S. A., & Pacheco-Tena, C. (2016). Hypoxia and its implications in rheumatoid arthritis. *Journal of Biomedical Science*, *23*(1), 62. http://doi.org/10.1186/s12929-016-0281-0

R Core Team, R. C. T. (2020). R: A language and environment for statistical computing. {SOFTWARE}.{COMPUTER_SOFTWARE}, R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Ramachandran, P., Dobie, R., Wilson-Kanamori, J. R., Dora, E. F., Henderson, B. E. P., Luu, N. T., … Henderson, N. C. (2019). Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature*, *575*(7783), 512–518. http://doi.org/10.1038/s41586-019-1631-3

Ramirez Flores, R. O., Lanzer, J. D., Holland, C. H., Leuschner, F., Most, P., Schultz, J., … Saez-Rodriguez, J. (2021). Consensus transcriptional landscape of human end-stage heart failure. *JOURNAL OF ANCIENT HISTORY AND ARCHAEOLOGY*. http://doi.org/10.1161/{JAHA}.120.019667

Ramnath, D., Irvine, K. M., Lukowski, S. W., Horsfall, L. U., Loh, Z., Clouston, A. D., … Sweet, M. J. (2018). Hepatic expression profiling identifies steatosis-independent and steatosis-driven advanced fibrosis genes. *JCI Insight*. Retrieved from `https://insight.jci.org/articles/view/120274`

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., … Participants, H. C. A. M. (2017). The human cell atlas. *eLife*, *6*. http://doi.org/10.7554/{eLife}.27041

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. http://doi.org/10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. http://doi.org/10.1186/1471-2105-12-77

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. http://doi.org/10.1093/bioinformatics/btp616

Robrahn, L., Dupont, A., Jumpertz, S., Zhang, K., Holland, C. H., Guillaume, J., … Cramer, T. (2021). Conditional deletion of HIF-1provides new insight regarding the murine response to gastrointestinal infection with salmonella typhimurium. *BioRxiv*. http://doi.org/10.1101/2021.01.16.426940

Roopra, A. (2020). MAGIC: A tool for predicting transcription factors and cofactors driving gene sets using ENCODE data. *PLoS Computational Biology*, *16*(4), e1007800. http://doi.org/10.1371/journal.pcbi.1007800

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., … Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176–182. http://doi.org/10.1126/science.aam8999

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. http://doi.org/10.1016/0377-0427(87)90125-7

Rydenfelt, M., Klinger, B., Klünemann, M., & Blüthgen, N. (2020). SPEED2: Inferring upstream pathway activity from differential gene expression. *Nucleic Acids Research*, *48*(W1), W307–W312. http://doi.org/10.1093/nar/gkaa236

Salviato, E., Djordjilović, V., Chiogna, M., & Romualdi, C. (2019). SourceSet: A graphical model approach to identify primary genes in perturbed biological pathways. *PLoS Computational Biology*, *15*(10), e1007357. http://doi.org/10.1371/journal.pcbi.1007357

Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., … Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications*, *9*(1), 20. http://doi.org/10.1038/s41467-017-02391-6

Segal, J. M., Kent, D., Wesche, D. J., Ng, S. S., Serra, M., Oulès, B., … Rashid, S. T. (2019). Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nature Communications*, *10*(1), 3350. http://doi.org/10.1038/s41467-019-11266-x

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv*. http://doi.org/10.1101/060012

Sezgin, S., Hassan, R., Zühlke, S., Kuepfer, L., Hengstler, J. G., Spiteller, M., & Ghallab, A. (2018). Spatio-temporal visualization of the distribution of acetaminophen as well as its metabolites and adducts in mouse livers by MALDI MSI. *Archives of Toxicology*, *92*(9), 2963–2977. http://doi.org/10.1007/s00204-018-2271-3

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., … Regev, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, *510*(7505), 363–369. http://doi.org/10.1038/nature13437

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software*, *1*(3). http://doi.org/10.21105/joss.00037

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., … Willighagen, E. L. (2018). WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, *46*(D1), D661–D667. http://doi.org/10.1093/nar/gkx1064

Smits, A. H., Ziebell, F., Joberty, G., Zinn, N., Mueller, W. F., Clauder-Münster, S., … Huber, W. (2019). Biological plasticity rescues target activity in CRISPR knock outs. *Nature Methods*, *16*(11), 1087–1093. http://doi.org/10.1038/s41592-019-0614-5

Sousa Abreu, R. de, Penalva, L. O., Marcotte, E. M., & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular Biosystems*, *5*(12), 1512–1526. http://doi.org/10.1039/b908315d

Staniek, J., Lorenzetti, R., Heller, B., Janowska, I., Schneider, P., Unger, S., … Rizzi, M. (2019). TRAIL-R1 and TRAIL-R2 mediate TRAIL-dependent apoptosis in activated primary human b lymphocytes. *Frontiers in Immunology*, *10*, 951. http://doi.org/10.3389/fimmu.2019.00951

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews. Genetics*, *16*(3), 133–145. http://doi.org/10.1038/nrg3833

Stricker, S. H., Köferle, A., & Beck, S. (2017). From profiles to function in epigenomics. *Nature Reviews. Genetics*, *18*(1), 51–66. http://doi.org/10.1038/nrg.2016.138

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., … Golub, T. R. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, *171*(6), 1437–1452.e17. http://doi.org/10.1016/j.cell.2017.10.049

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. http://doi.org/10.1073/pnas.0506580102

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*(4), 599–604. http://doi.org/10.1038/nprot.2017.149

Szalai, B., & Saez-Rodriguez, J. (2020). Why do pathway methods work better than they should? *FEBS Letters*, *594*(24), 4189–4200. http://doi.org/10.1002/1873-3468.14011

Szalai, B., Subramanian, V., Holland, C. H., Alföldi, R., Puskás, L. G., & Saez-Rodriguez, J. (2019). Signatures of cell death and proliferation in perturbation transcriptomics data-from confounding factor to effective prediction. *Nucleic Acids Research*, *47*(19), 10010–10026. http://doi.org/10.1093/nar/gkz805

Tajti, F., Kuppe, C., Antoranz, A., Ibrahim, M. M., Kim, H., Ceccarelli, F., … Saez-Rodriguez, J. (2020). A functional landscape of CKD entities from public transcriptomic data. *Kidney International Reports*, *5*(2), 211–224. http://doi.org/10.1016/j.ekir.2019.11.005

Tak, P. P., & Firestein, G. S. (2001). NF-kappaB: A key role in inflammatory diseases. *The Journal of Clinical Investigation*, *107*(1), 7–11. http://doi.org/10.1172/{JCI11830}

Tanevski, J., Ramirez Flores, R. O., Gabor, A., Schapiro, D., & Saez-Rodriguez, J. (2020). Explainable multi-view framework for dissecting inter-cellular signaling from highly multiplexed spatial data. *BioRxiv*. http://doi.org/10.1101/2020.05.08.084145

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … Surani, M. A. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. http://doi.org/10.1038/nmeth.1315

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-S., … Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, *25*(1), 75–82. http://doi.org/10.1093/bioinformatics/btn577

Tenenbaum, J. D., Walker, M. G., Utz, P. J., & Butte, A. J. (2008). Expression-based pathway signature analysis (EPSA): Mining publicly available microarray data for insight into human disease. *BMC Medical Genomics*, *1*, 51. http://doi.org/10.1186/1755-8794-1-51

Teske, B. F., Wek, S. A., Bunpo, P., Cundiff, J. K., McClintick, J. N., Anthony, T. G., & Wek, R. C. (2011). The eIF2 kinase PERK and the integrated stress response facilitate activation of ATF6 during endoplasmic reticulum stress. *Molecular Biology of the Cell*, *22*(22), 4390–4405. Retrieved from `http://dx.doi.org/10.1091/mbc.E11-06-0510`

Teufel, A., Itzel, T., Erhart, W., Brosch, M., Wang, X. Y., Kim, Y. O., … Hampe, J. (2016). Comparison of gene expression patterns between mouse models of nonalcoholic fatty liver disease and liver tissues from patients. *Gastroenterology*, *151*(3), 513–525.e0. http://doi.org/10.1053/j.gastro.2016.05.051

Tomfohr, J., Lu, J., & Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, *6*, 225. http://doi.org/10.1186/1471-2105-6-225

Torre, D., Lachmann, A., & Ma'ayan, A. (2018). BioJupies: Automated generation of interactive notebooks for RNA-seq data analysis in the cloud. *Cell Systems*, *7*(5), 556–561.e3. http://doi.org/10.1016/j.cels.2018.10.007

Trescher, S., Münchmeyer, J., & Leser, U. (2017). Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. *BMC Systems Biology*, *11*(1), 41. http://doi.org/10.1186/s12918-017-0419-z

Väremo, L., Nielsen, J., & Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*, *41*(8), 4378–4391. http://doi.org/10.1093/nar/gkt111

Wang, Zhenjia, Civelek, M., Miller, C. L., Sheffield, N. C., Guertin, M. J., & Zang, C. (2018). BART: A transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics*, *34*(16), 2867–2869. http://doi.org/10.1093/bioinformatics/bty194

Wang, Zhong, Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. http://doi.org/10.1038/nrg2484

Wang, Zichen, Monteiro, C. D., Jagodnik, K. M., Fernandez, N. F., Gundersen, G. W., Rouillard, A. D., … Ma'ayan, A. (2016). Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nature Communications*, *7*, 12846. http://doi.org/10.1038/ncomms12846

Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R., Plouffe, D., Deciu, C., … Yates, J. R. (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(6), 3107–3112. http://doi.org/10.1073/pnas.0634629100

Wenk, M. R. (2005). The emerging field of lipidomics. *Nature Reviews. Drug Discovery*, *4*(7), 594–610. http://doi.org/10.1038/nrd1776

Wickham, H. (2016). *ggplot2 - elegant graphics for data analysis* (2nd ed.). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-24277-4

Wilhelm, B. T., & Landry, J.-R. (2009). RNA-seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, *48*(3), 249–257. http://doi.org/10.1016/j.ymeth.2009.03.016

Wingender, E., Schoeps, T., Haubrock, M., Krull, M., & Dönitz, J. (2018). TFClass: Expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research*, *46*(D1), D343–D347. http://doi.org/10.1093/nar/gkx987

Wiredja, D. D., Koyutürk, M., & Chance, M. R. (2017). The KSEA app: A web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics*, *33*(21), 3489–3491. http://doi.org/10.1093/bioinformatics/btx415

Younossi, Z. M., Stepanova, M., Ong, J., Trimble, G., AlQahtani, S., Younossi, I., … Henry, L. (2020). Nonalcoholic steatohepatitis is the most rapidly increasing indication for liver transplantation in the united states. *Clinical Gastroenterology and Hepatology*. http://doi.org/10.1016/j.cgh.2020.05.064

Zakrzewska, A., Cui, C., Stockhammer, O. W., Benard, E. L., Spaink, H. P., & Meijer, A. H. (2010). Macrophage-specific gene functions in Spi1-directed innate immunity. *Blood*, *116*(3), e1–11. http://doi.org/10.1182/blood-2010-01-262873

Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, *18*(1), 174. http://doi.org/10.1186/s13059-017-1305-0

Zardi, E. M., Navarini, L., Sambataro, G., Piccinni, P., Sambataro, F. M., Spina, C., & Dobrina, A. (2013). Hepatic PPARs: Their role in liver physiology, fibrosis and treatment. *Current Medicinal Chemistry*, *20*(27), 3370–3396. http://doi.org/10.2174/09298673113209990136

Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., … Ren, B. (2021). A cell atlas of chromatin accessibility across 25 adult human tissues. *BioRxiv*. http://doi.org/10.1101/2021.02.17.431699

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-seq and microarray in transcriptome profiling of activated t cells. *Plos One*, *9*(1), e78644. http://doi.org/10.1371/journal.pone.0078644