

Development of a Multi-Species scRNA-seq Atlas of Metabolic
Dysfunction–Associated Steatotic Liver Disease (MASLD)

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
MOLECULAR BIOLOGY AND GENETICS

BY
Mert Demirdizen
JULY 2024

Development of a Multi-Species scRNA-seq Atlas of Metabolic Dysfunction–Associated Steatotic Liver Disease (MASLD)

By Mert Demirdizen

July 2024

We certify that we have read this thesis and that, in our opinion, it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Özlen Konu KARAKAYALI

Assoc. Prof. Ceren Sucularlı

Asst. Prof. Serkan İsmail Göktuna

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Orhan Arıkan

Director of the Graduate School of Engineering and Science

ABSTRACT

Development of a Multi-Species scRNA-seq Atlas of Metabolic Dysfunction–Associated Steatotic Liver Disease (MASLD)

Mert Demirdizen

M.Sc. in Molecular Biology and Genetics

Advisor: Özlen Konu Karakayalı

July 2024

Metabolic dysfunction-associated steatotic liver disease (MASLD) affects a significant portion of the human population, potentially leading to severe secondary conditions. Despite extensive research on patients and animal models to understand the disease's initiation and progression, and the development of various therapeutic strategies, current non-invasive techniques remain inefficient, and only one drug has been approved. This underscores the critical need to elucidate the cellular mechanisms driving MASLD to develop new diagnostic and therapeutic approaches. Single-cell RNA sequencing (scRNA-seq) enables the investigation of transcriptomic profiles at the cellular level. Various studies have utilized scRNA-seq to explore MASLD in patient samples and animal models. Although there are numerous publicly available scRNA-seq datasets from these studies, no comprehensive tool exists to explore and analyse them collectively.

In this study, publicly available scRNA-seq datasets related to MASLD were collected and reanalysed, adhering to best practices suggested in the literature, and a web tool was developed for their visualisation and custom downstream analyses. Thirteen datasets obtained from the livers of human MASLD spectrum patients, and three datasets different animal models, two mouse and one zebrafish, were uniformly preprocessed. The two human datasets containing cirrhotic samples and two fibrotic datasets, one from mouse and another from zebrafish models, were integrated to reveal the conserved mechanisms of liver fibrosis among human patients and animal models. A web tool was developed in R Shiny to enable visualizations, cell cluster selection, differential expression analysis and gene set enrichment of genes and gene sets. The utility of the app was demonstrated by conducting a case study with the integrated dataset. The results included an endothelial cluster that had cells from each of the four datasets, and differential gene expression analysis on the integrated dataset provided significantly modulated known, e.g., SYNPO and TIMD4 liver fibrosis-associated genes across the three species. As a result, the custom analysis workflow provided in this thesis can allow the detection of conserved markers in different MASDL stages by further analysing this and other integrated datasets in the future.

Keywords: Metabolic Dysfunction–Associated Steatotic Liver Disease, MASLD, single-cell RNA sequencing, scRNA-seq, Shiny App

ÖZET

Metabolik Disfonksiyon İlişkili Yağlı Karaciğer Hastalığı (MASLD) için Çok Türü scRNA-seq Atlasının Geliştirilmesi

Mert Demirdizen

Moleküler Biyoloji ve Genetik Yüksek Lisans Programı

Danışman: Özlen Konu Karakayalı

Temmuz 2024

Metabolik disfonksiyon ilişkili steatotik karaciğer hastalığı (MASLD), insan nüfusunun önemli bir kısmını etkileyerek ciddi ikincil durumlara yol açabilir. Hastalığın başlangıcı ve ilerlemesini anlamak ve çeşitli tedavi stratejileri geliştirmek amacıyla yapılan kapsamlı hasta ve deney hayvanı araştırmalarına rağmen, mevcut non-invaziv teknikler yetersiz kalmakta ve sadece bir ilaç onaylanmıştır. Bu durum, MASLD'nin hücresel mekanizmalarını ortaya çıkarmak ve yeni teşhis ve tedavi yaklaşımları geliştirmek için kritik bir ihtiyacı vurgulamaktadır. Tek hücre RNA dizilemesi (scRNA-seq), hücresel düzeyde transkriptomik profilleri incelemeyi mümkün kılar. Çeşitli çalışmalar, hasta örneklerinde ve hayvan modellerinde MASLD'yi keşfetmek için scRNA-seq'i kullanmıştır. Bu çalışmalardan elde edilen çok sayıda kamuya açık scRNA-seq veri seti olmasına rağmen, bunları topluca incelemek ve analiz etmek için kapsamlı bir araç bulunmamaktadır.

Bu çalışmada, MASLD ile ilgili halka açık scRNA-seq veri setleri toplanmış ve literatürde önerilen en iyi uygulamalara uygun olarak yeniden analiz edilmiştir. Ayrıca, bu veri setlerinin görselleştirilmesi ve özel aşağı yönlü analizler için bir web aracı geliştirilmiştir. İnsan MASLD spektrum hastalarının karaciğerlerinden, iki farklı fare modelinden ve bir zebrafish modelinden elde edilen toplamda 16 veri seti, homojen bir şekilde ön işleme tabi tutulmuştur. Sirotik örnekler içeren iki insan veri seti ve bir fare ve bir zebrafish modelinden elde edilen fibrotik veri setleri, insan hastaları ve hayvan modelleri arasında karaciğer fibrozunun korunan mekanizmalarını ortaya çıkarmak amacıyla entegre edilmiştir. R Shiny'de geliştirilen web aracı, görselleştirmeler, hücre kümesi seçimi, farklı gen ekspresyonu analizi ve gen setleri ve gen setlerinin zenginleştirilmesini sağlamaktadır. Uygulamanın faydası, entegre veri setiyle gerçekleştirilen bir vaka çalışması ile gösterilmiştir. Sonuçlar, dört veri setinin her birinden hücreler içeren bir endotel kümesini içermekte olup, entegre veri seti üzerindeki farklı gen ekspresyonu analizi, üç tür arasında anlamlı şekilde modüle edilmiş, örneğin SYNPO ve TIMD4 gibi karaciğer fibrozisi ile ilişkili bilinen genler sağlamıştır. Sonuç olarak, bu tezde sunulan özel analiz iş akışı, bu ve diğer entegre veri setlerinin gelecekte daha fazla analiz edilmesiyle farklı MASLD aşamalarında korunan belirteçlerin tespit edilmesine olanak sağlayabilir.

Anahtar Kelimeler: Metabolik disfonksiyon ilişkili yağlı karaciğer hastalığı, MASLD, tek hücre RNA dizilemesi, scRNA-seq, Shiny Uygulaması



To my mother and brother...

Acknowledgements

I would like to express my gratitude to my advisor, Dr. Özlen Konu Karakayalı, for her supervision and invaluable support since my undergraduate years. I am grateful for the opportunity to be a part of Konu Lab.

I sincerely thank my committee members, Dr. Ceren Sucularlı and Dr. Serkan Göktuna, for their time reviewing my thesis, and I appreciate their valuable remarks.

I would like to especially thank Konu Lab members and alumni, namely Dr. Merve Vural Özdeniz, not only for being a great supervisor and providing me with influential guidance but also for her valuable friendship; Siber Güneş Kılınç for all her support and for making me gain new points of view; Seda Aktürk for all her caring and generous attitudes; and Abdul Moiz Aftab for cheering up my master's years.

I also want to thank my friends, Ahmet Göktuğ Özkurt and Naciye Bozkurt, for their consolations and all the memorable times we shared. Special thanks are extended to Hüseyin Atamtürk, who has greatly impacted me and has always been there when I needed.

I want to express my gratitude to Rana Acar for her continuous love, tolerance, and support. Her belief in me has been a continual source of strength, and her support has served as my anchor throughout this journey. I could not have completed my thesis or many other things in my life without her.

I would like to thank my family, my brother, Berkay Demirdizen, and my mother, Güler Egritağ, for their unconditional and endless support of my decisions. They never expect anything from me except my happiness. Anything good I have done or will do, they are the architects of it.

I thank Bilkent University, Molecular Biology and Genetics Department for providing scholarship throughout my master's studies. I acknowledge that I am a recipient of the fellowship of 2210-A National MSc/MA Scholarship Program of Directorate of Science Fellowships and Grant Programmes (BİDEB) of Scientific and Technological Research Institution of Türkiye (TÜBİTAK). This study was conducted as part of the Halt-RONIN consortium funded by EU HORIZON (HORIZON-HLTH-2022-STAYHLTH-02/101095679).

Table of Contents

1. Introduction.....	1
1.1 Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD)	1
1.1.1 Simple Steatosis	5
1.1.2 Metabolically Associated Steatohepatitis	5
1.1.3 Cirrhosis	6
1.2 Model Animals for MASLD	6
1.2.1 Mouse Models for MASLD	7
1.2.1 Zebrafish Models for MASLD	7
1.3 Single Cell RNA Sequencing	8
1.3.1 Use of Single-Cell RNA Sequencing in Different Scopes	8
1.3.2 Single-Cell RNA Sequencing Techniques	8
1.3.3 Single-Cell RNA Sequencing Analysis.....	9
1.3.3.1 Preprocessing	10
1.3.3.1.1 Generation of Gene Expression Matrix	11
1.3.3.1.2 Ambient RNA Removal.....	13
1.3.3.1.3 Doublet Removal	13
1.3.3.1.4 Quality Control	15
1.3.3.1.5 Normalization	15
1.3.3.1.6 Feature Selection.....	16
1.3.3.1.7 Integration	17
1.3.3.1.8 Dimensionality Reduction	18
1.3.3.2 Downstream Analyses	18
1.3.3.2.1 Clustering.....	19
1.3.3.2.2 Differential Gene Expression Analysis	20
1.3.3.2.3 Cell Type Annotation.....	21
1.3.3.2.4 Gene Set Enrichment Analysis	22
1.3.3.2.5 Cell-Cell Communication Analysis	22
1.3.3.2.6 Trajectory Analysis.....	23
1.4 Use of Single-Cell RNA Sequencing in Liver Diseases	24
1.4.1 Use of Single-Cell RNA Sequencing in MASLD	24
1.4.2 Use of Single-Cell RNA Sequencing in MASLD with Mouse Models	26
1.4.3 Use of Single-Cell RNA Sequencing in MASLD with Zebrafish Models.....	27
1.5 Web Tools for Single-Cell RNA Sequencing Analysis	27
1.5.1 Web Tools for MASLD.....	28

2. Aims and Rationale.....	29
3. Material and Methods	31
3.1 scRNA-seq Data Analyses	31
3.1.1 Data Acquisition.....	30
3.1.2 Preprocessing.....	33
3.2 Multi-Species scRNA-seq Atlas of MASLD Shiny App Development.....	38
4. Results	41
4.1 Preprocessing of Individual and Integrated Datasets	41
4.2 Multi-species scRNA-seq Atlas of MASLD Shiny App.....	46
5. Conclusion and Discussion and Future Perspectives.....	59
5.1 Preprocessing	59
5.2 Multi-Species scRNA-seq Atlas of MASLD Shiny App	61
5.3 Multi-Species Integration of Fibrotic Samples	63
References.....	67
Appendix.....	92

List of Figures

Figure 1.1 Disease Progression of MASLD (Created with BioRender.com)	5
Figure 1.2 scRNA-seq preprocessing workflow with a list of available and widely used tools or methods for each step.	11
Figure 1.3 Approaches of scRNA-seq downstream analyses	19
Figure 3.1 Conducted scRNA-seq data preprocessing workflow	35
Figure 3.2 Workflow of <i>Multi-Species scRNA-seq Atlas of MASLD</i>	39
Figure 4.1 UMAP visualizations of preprocessed datasets used in dataset integration A) Categorized by conditions, B) Categorized by cell type annotations.	42
Figure 4.2 UMAP visualisation of cell type annotation, condition, and dataset identifiers of integrated datasets and distribution of samples per cell type A) Conditions of the cells on UMAP, B) Cell type annotations on UMAP, C) Datasets on UMAP, D) Proportions of sample conditions per cell type.....	44
Figure 4.3 Import Dataset tab of Multi-Species scRNA-seq Atlas of MASLD	47
Figure 4.4 Cluster Selection Tab, Visualize Expression of Gene Section of <i>Multi-Species scRNA-seq Atlas of MASLD</i>	48
Figure 4.5 Cluster Selection Tab, Visualize Expression of Geneset Section of Multi-Species scRNA-seq Atlas of MASLD.	50
Figure 4.6 Cluster Selection Tab, Calculate Co-Expression Section of <i>Multi-Species scRNA-seq Atlas of MASLD</i>	51
Figure 4.7 Differential Expression Tab of <i>Multi-Species scRNA-seq Atlas of MASLD</i> . A) Tab Overview, B) Term enrichment results.....	52
Figure 4.8 Pseudobulk Tab of Multi-Species scRNA-seq Atlas of MASLD.	54
Figure 4.9 Comparison of Differentially Expressed Genes of Individually Preprocessed Datasets Against Integrated Dataset. A) Scatter plots of log fold changes of individual datasets on the x-axis and log fold changes of integrated datasets on the y-axis with Spearman correlation scores. B) Categorical comparison of significantly regulated genes ($ \log FC > 1$ and adjusted p-val < 0.05) in the individual datasets and their regulation status in the integrated dataset where the first bar plot shows the number of significantly regulated genes and the second bar plot their categorical comparison with the integrated dataset.....	56

List of Tables

Table 3.1. Human scRNA-seq Datasets Used in This Study	32
Table 3.2. Animal Models scRNA-seq Datasets Used in This Study	33
Table 3.3. Libraries Used in Preprocessing.....	37
Table 3.4. Libraries Used in Web Tool Development	40
Table 4.1. Significantly Differentially Regulated Genes in Individual Datasets Used in Integration	58



List of Appendixes

Appendix Figure 1 UMAP visualisation of cell type annotation of GSE212735 A) Conditions of the cells, B) Cell type annotations.	92
Appendix Figure 2 UMAP visualisation of cell type annotation of GSE212837 A) Conditions of the cells, B) Cell type annotations.	92
Appendix Figure 3 UMAP visualisation of cell type annotation of GSE202379 A) Conditions of the cells, B) Cell type annotations.	92
Appendix Figure 4 UMAP visualisation of cell type annotation of GSE159977 A) Conditions of the cells, B) Cell type annotations.	93
Appendix Figure 5 UMAP visualisation of cell type annotation of GSE168933 A) Conditions of the cells, B) Cell type annotations.	93
Appendix Figure 6 UMAP visualisation of cell type annotation of GSE169446 A) Conditions of the cells, B) Cell type annotations	93
Appendix Figure 7 UMAP visualisation of cell type annotation of GSE174748 A) Conditions of the cells, B) Cell type annotations.	94
Appendix Figure 8 UMAP visualisation of cell type annotation of GSE189600 A) Conditions of the cells, B) Cell type annotations.	94
Appendix Figure 9 UMAP visualisation of cell type annotation of GSE190487 A) Conditions of the cells, B) Cell type annotations.	94
Appendix Figure 10 UMAP visualisation of cell type annotation of GSE192740 A) Conditions of the cells, B) Cell type annotations.	95
Appendix Figure 11 UMAP visualisation of cell type annotation of GSE156059 A) Conditions of the cells, B) Cell type annotations.	95
Appendix Figure 12 Cluster Selection Tab: EGFL7 Expression in Integrated Datasets	96
Appendix Table 1. Gene List Used as Cell Type Markers Used for Enrichment	97

Abbreviations

Aquaporin 1 : AQP1
Cardiovascular Diseases : CVD
Continuous-state Hidden Markov Models: CSHMM

Cysteine And Glycine-Rich Protein 2 : CSRP2
Deoxyribonuclease Gamma : DNASE1L3
Differential Gene Expression : DGE
EGF-Like Domain-Containing Protein 7 : EGFL7
Food And Drug Administration : FDA
Galectin-1 : LGALS1
Gene Set Enrichment Analysis : GSE
Genome-Wide Association Studies : GWAS
Haptoglobin : HP
Hydroxylase : PAH
Insulin-Like Growth Factor-Binding Protein 2 : IGFBP2
Insulin-Like Growth Factor-Binding Protein 7 : IGFBP7
Leucine Rich Repeat and Fibronectin Type III Domain Containing 4 : LRFN4
Mannose Receptor C-type 1 : MRC1
Median Absolute Deviations : Mads
Member 4 : TNFRSF4
Metabolic Dysfunction-Associated Steatotic Liver Disease : MASLD
Metabolically Associated Steatohepatitis : MASH
Muscle A Actin : ACTA2
Non-Alcoholic Fatty Liver Disease : NAFLD
Non-Alcoholic Steatohepatitis : NASH
Platelet-Derived Growth Factor Receptors : PDGFR
Polymerase Chain Reaction : PCR
Principal Component Analysis : PCA
Receptor-Type Tyrosine-Protein Phosphatase Beta : PTPRB
RNA Sequencing : RNA-Seq
Single-Cell RNA Sequencing : scRNA-seq
SPARC-Like Protein 1 : SPARCL1
Synaptopodin : SYNPO
T-cell Immunoglobulin and Mucin Domain Containing 4 : TIMD4
T-Distributed Stochastic Neighborhood : T-SNE
TSC22 Domain Family Protein 1 : TSC22D1
Tumor Necrosis Factor Receptor Superfamily
Tumor Necrosis Factor Receptor Superfamily Member 12A : TNFRSF12A
Type 2 Diabetes Mellitus : T2DM
Unconventional Myosin-Va : MYO5A
Uniform Manifold Approximation And Projection : UMAP
Unique Molecular Identifiers : UMIs
Vimentin : VIM

1. Introduction

1.1 Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD)

Metabolic dysfunction-associated steatotic liver disease (MASLD), hitherto known as non-alcoholic fatty liver disease (NAFLD), is thought to be induced by obesity, type 2 diabetes mellitus (T2DM), and the consumption of a considerably lower amount of alcohol than the definition of alcoholism (Rinella et al., 2023). It was previously considered a western disease; however, recent growing epidemiological data suggest it is prevalent worldwide (Wong & Chan, 2021; Z. Younossi et al., 2019). It is estimated that 25.24-29.8% of the world population is affected by the disease (Le et al., 2022; Z. M. Younossi et al., 2016).

Diagnosis of MASLD can be made by non-invasive techniques such as tests for elevated liver enzymes in the serum and imaging of the liver in routine clinical care (Alexander et al., 2018). However, liver biopsy is still considered a gold standard method to conduct differential diagnosis and determine the severity of the disease, and other methods do not meet regulatory requirements to be used as diagnostic tests (Berzigotti et al., 2021; Sanyal et al., 2023). However, liver biopsy has potential complications as it is an invasive technique and error-prone method due to its sampling step (Rockey et al., 2009). Radiological imaging of the liver provides information about the disease; however, a liver biopsy is needed to confirm changes in the liver with higher sensitivity (Tapper & Lok, 2017). Recent research aims to use gathered data on patients with MASLD to develop more accurate non-invasive techniques with the use of imaging and serum enzyme levels (Kouvari et al., 2023; McTeer et al., 2024; Stähli et al., 2023).

Recently, the Food and Drug Administration (FDA) approved resmetirom under the commercial name of Rezzdiffra as the first drug to treat MASLD (Harrison et al., 2024). It is

a liver-directed, thyroid hormone receptor beta-selective agonist that effectively reduces fat accumulation in the liver. It was approved by the FDA in an accelerated approval pathway, and it is under another trial study (U.S. Food and Drug Administration, 2024). The trial will be applied to subjects whose inflammation due to MASLD was confirmed by liver biopsy. In the initial stages, weight loss is shown to have a reversibility effect, and weight loss medications, bariatric surgery, and diabetes medications are advised to be used for patients (Cusi et al., 2022). For advanced stages, liver transplantation is required and is effective for increasing survival (Cervantes-Alvarez et al., 2022). Several drugs are being investigated for repurposing to stop the progression of the disease and reverse its effects on the liver; however, they are still under trial (Eeda et al., 2024).

There are various risk factors for MASLD, including metabolic, genetic, epigenetic, environmental, demographic, and gut dysbiosis factors (Juanola et al., 2021). The metabolic risk factors include diabetes, obesity, and dyslipidemia (Ampuero et al., 2018). New insights into genetic factors were gained with genome-wide association studies (GWAS) studies with multi-centre studies and different populations, including both the Western and Eastern parts of the world (Chambers et al., 2011; Lee et al., 2024; J. Liu et al., 2023; Speliotes et al., 2011; Yuan et al., 2008). However, contradictory results show no significant association with any foci on MASLD (X. Cheng et al., 2024; Trépo & Valenti, 2020). Results show that different populations have common and distinguishing loci for predisposition to MASLD; however, more research is needed using advanced sequencing techniques and more diverse population settings (Trépo & Valenti, 2020). Epigenome studies show significant patterns for DNA methylation, histone modification, and non-coding RNAs playing a role during the initiation and advancement of MASLD (Bayoumi et al., 2020). Although associated histone modifications are still vague, regulations of miRNAs and DNA methylations were well

justified in both human samples and mouse models (Jonas & Schürmann, 2021). Environmental factors, including Western diet, smoking, high body weight, and low physical activity, were associated with an increased risk of MASLD (Jahromi et al., 2023). Also, the changed nomenclature includes the definition of light and moderate alcohol intake as a risk factor (Eslam et al., 2020). In addition to these established risk factors, current research identifies new associations with oral health (Pischke et al., 2023). The demographic studies show that the prevalence of MASLD increases with age, and male patients have a higher risk than female patients (Fresneda et al., 2022). Gut dysbiosis is another risk factor that was depicted to cause or worsen the development of the disease with well-characterized mechanistic pathways (Fang et al., 2022).

MASLD and its risk factors have the potential to cause mild to severe complications, from obstructive sleep apnea to the risk of hepatic cancers, and various epidemiological studies have shown that they increase mortality rates (Mantovani et al., 2020). Advancement of the disease increases the mortality rate, and most contributing mortality causes are by order: cardiovascular diseases (CVD), extrahepatic malignancies, and liver-related complications (Konyn et al., 2023; Lekakis & Papatheodoridis, 2024). Even in the intermediate stages of MASLD, CVD mortality rates were shown to significantly increased (Ma et al., 2022). Although some of these CVDs, for example, atherosclerotic cardiovascular disease, have common risk factors with MASLD, causal relationships independent from them were established (Driessen et al., 2023). Meta-analysis of cohort studies depicts an increased risk of extrahepatic cancers due to MASLD. While gastrointestinal cancers are the most prevalent among them, lung, breast, and gynaecological or urinary system cancers were also significantly increased independent of common risk factors (Mantovani et al., 2022). Liver-related mortality

is mainly associated with progressed disease presenting cirrhosis and cirrhosis-related complications such as chronic liver failure and hepatocellular carcinoma (Konyn et al., 2023).

Pathogenesis of MASLD is explained by the parallel, multiple-hit hypothesis, where the first hit is the hepatic accumulation of lipids followed by liver damage due to various factors as second multiple hits (Bessone et al., 2019). Hepatic accumulation of lipids, i.e., the first hit, mainly consists of changes in four molecular mechanisms: uptake of circulating lipids, de novo lipogenesis, fatty acid oxidation, and export of lipids within very low-density lipoproteins (Ipsen et al., 2018). An increase in the uptake of circulating lipids and de novo lipogenesis causes increased amounts of lipids and reduced capabilities of fatty acid oxidation and export of lipids in very low-density lipoproteins, thus leading to fat accumulation in the liver. Subsequent liver damage, i.e., the second hit, can be caused by various factors; however, lipotoxicity is the main result of fat accumulation, which is well characterised by changes in endoplasmic reticulum stress, oxidative stress, organelle dysfunction, and ferroptosis (Y. Li et al., 2024). Lipotoxicity, through hypoxia, was shown to worsen the disease, which further increases lipotoxicity (Geng et al., 2021). Other mechanisms contribute to the second hit, such as how gut dysbiosis causes intrahepatic and extrahepatic damage during MASLD (Benedé-Ubieto et al., 2024). Various other defined inducers of hepatic damage mechanisms are also caused by different factors (X. Xu et al., 2022).

The spectrum of MASLD can be investigated in three main categories by order: simple steatosis, metabolically associated steatohepatitis (MASH), and cirrhosis (Rao et al., 2023). Their prevalence estimates are difficult since the staging requires a liver biopsy (Z. M. Younossi et al., 2023).

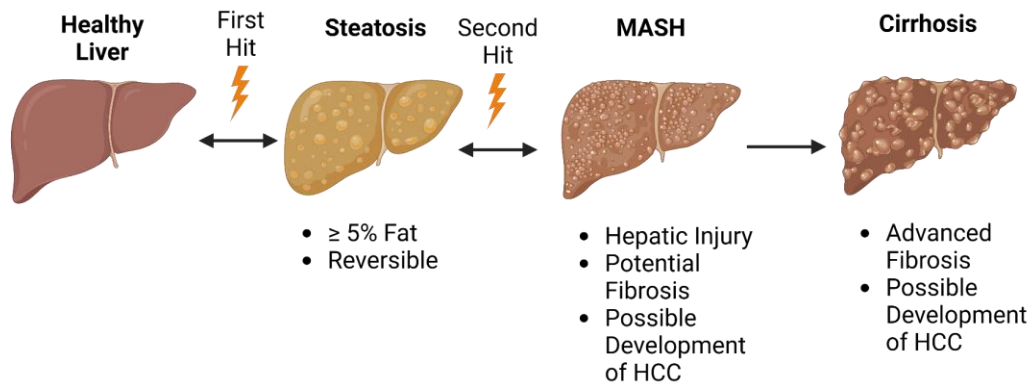


Figure 1.1 Disease Progression of MASLD (Created with BioRender.com).

1.1.1 Simple Steatosis

Simple steatosis is defined by at least 5% fat accumulation in the liver and without any fibrosis or degeneration of hepatocyte cells; however, it provides the basis for hepatic injury due to risk factors (De & Duseja, 2020). This stage does not threaten health and can be reversible by adopting a healthy lifestyle (Koutoukidis et al., 2019).

1.1.2 Metabolically Associated Steatohepatitis

Metabolically Associated Steatohepatitis (MASH) is when the hepatic injury is seen and characterised by lobular inflammation, ballooning, and potentially fibrosis (Loomba et al., 2021). MASH has almost 20% potential to advance to cirrhosis and directly increase hepatocellular carcinoma without cirrhosis (Sheka et al., 2020). So, fibrosis is the most important contributor to mortality rates of MASLD patients, and the focus of drug development is on this stage (S. Wang & Friedman, 2023). Thus, it is the leading reason for liver transplantation for females and second leading, after alcoholic liver disease, for males (Noureddin et al., 2018).

1.1.3 Cirrhosis

Cirrhosis is a secondary disease that is caused by chronic liver injury (Sharma & John, 2018). Cirrhosis, characterized by advanced fibrosis and regenerative nodules in the liver parenchyma, develops due to prolonged inflammation (Ginès et al., 2021). While hepatic parenchymal cells play a central role in the development of fibrosis, the interaction between non-parenchymal cells, especially with endothelial cells, Kupfer cells, and hepatic stellate cells, has a great impact on the development of the fibrotic processes (Kumar et al., 2021).

Cirrhosis is one of the leading causes of death, and MASLD is one of the major causes of it (Ginès et al., 2021). It is that estimated prognosis of MASH-related decompensated cirrhosis is worse than other causes (Lekakis & Papatheodoridis, 2024). The increasing prevalence of MASLD leads to increased cirrhosis; interestingly, the increased ratios for MASLD-caused cirrhosis vary between populations (Zhai et al., 2021). Although symptoms related to cirrhosis can be treated and the fibrosis stage can be lowered by targeting the underlying cause, the annual age-adjusted mortality rates are increasing (Tapper & Parikh, 2023).

1.2 Model Animals for MASLD

To understand underlying mechanisms and develop new therapeutic approaches for MASLD, experimental models have been developed to mimic steatosis, MASH, fibrotic MASH, and cirrhosis (Nevzorova et al., 2020). There are numerous species in MASLD research, but the most common are mice, rats, rabbits, Ossabaw pigs, macaques, and zebrafish (S. Wu et al., 2023). Genetically modified and dietary models are used to develop MASH conditions in the mouse models, and the most used model is diet-induced c57BL/6 mouse models (C.-R. Yang et al., 2024). Also, due to very similar digestive track functions among vertebrates and its

unique advantages, zebrafish has been used in liver diseases since 1952 and in fatty liver since 1990 (Goessling & Sadler, 2015).

1.2.1 Mouse Models for MASLD

Mouse models are widely used for MASLD research by dietary, chemical, and genetic modelling and are used as preclinical models for drug development (Montero-Vallejo et al., 2024). Dietary models have been developed to mimic different stages of MASLD; simple steatosis or advanced fibrotic stages can be obtained with modifications of sugars, fat, cholesterol, and nutrition-deficient diets (Eng & Estall, 2021). Chemical models are primarily used for later stages of the disease since they cause more severe and rapid damage when compared to other methods; the developmental phase of the pathogenesis is not represented compared to other models (H. Wang et al., 2023). Genetically modified models are used to either stimulate the effect of specifically related polymorphisms or the more rapidly developing MASLD stage (Flessa et al., 2022). However, they fail to completely mimic the complex pathogenesis of MASLD, where various risk factors affect the pathogenesis, and the MASLD-related fibrosis mechanisms in currently used models are not complementary (Ipsen et al., 2020). Additional mechanisms should be considered when new models are designed to mimic both the onset and advancement of the disease (Gunes & Estall, 2024).

1.2.2 Zebrafish Models for MASLD

Due to rapid development and transparency during early developmental stages, zebrafish have been used as a model organism for MASLD using genetic, dietary, and chemical models (C. Chang et al., 2023). Dietary factors have been shown to induce fatty liver in zebrafish (B. Chen et al., 2018; X. Li et al., 2023). Chemically induced MASLD models were also shown with a combination of genetically modified zebrafish lines (M. K. Singh et al., 2023). The two-hit

mechanism was also demonstrated in zebrafish using a combination of genetic modification and diet (Kulkarni et al., 2022).

1.3 Single-Cell RNA Sequencing

Single-cell RNA Sequencing (scRNA-seq) allows for the investigation of heterogeneity between different cell types and regulatory mechanisms in single-cell resolution in a sample (Kolodziejczyk et al., 2015). It was first introduced in 2009 to obtain a high throughput transcriptional profile from a limited number of cells (Tang et al., 2009). After that, the scRNA-seq concept emerged, and advancements in the technique were made in different research fields (Stark et al., 2019).

1.3.1 Use of Single-Cell RNA Sequencing in Different Scopes

Over the past ten years, scRNA-seq has been used in various research fields, from microbiology to neurobiology, using samples from cell cultures, tissues, and plants (Jovic et al., 2022). It provides crucial information on heterogeneous profiles of cell types in samples; for example, it enables the characterisation of immune, tumour, and stromal cells for cancer research where bulk RNA-seq could not show their different profiles (Suvà & Tirosh, 2019). It is also considered a new tool for understanding the molecular mechanisms of tumours in clinics (X. Liu et al., 2022). In order to investigate cell types in the human body, initiatives have been proposed to identify the molecular mechanisms of the known cell types and discover novel cell types (Regev et al., 2017).

1.3.2 Single-Cell RNA Sequencing Techniques

There are four main steps in scRNA-seq: isolation of single cells, cDNA amplification, library preparation, and sequencing libraries (Haque et al., 2017). Different isolation techniques exist,

but five are well established: manual cell picking, microfluidic, laser capture microdissection, magnetic-activated cell sorting, and fluorescence-activated cell sorting (P. Hu et al., 2016). They can be used for different requirements; for example, manual cell picking can isolate cells with interest-specific morphology, and microfluidics can isolate cells with specific antigens on their surfaces. For cDNA amplification, polymerase chain reaction (PCR) is the most used method, but in vitro transcription amplification is also used in different protocols (S. Wang, Sun, et al., 2023). In this step, unique molecular identifiers (UMIs), short random sequences added to each mRNA, are used to eliminate duplicated artefacts due to amplification, but it can also be done with alternative normalisation methods (Townes & Irizarry, 2020). Various methods are used for library preparation, but fragmentation and ligation of sequencing adapters are mainly for preparing samples for high-throughput sequencing (Haque et al., 2017). As a final stage, libraries are subjected to high-throughput sequencing that enables tens of thousands of individual cells per sample (Zheng et al., 2017).

1.3.3 Single-Cell RNA Sequencing Analysis

The raw results of the scRNA-seq process must be processed to get the expression profiles of RNAs with corresponding cells and workflows, and their parameters affect the assay outcome (AlJanahi et al., 2018). There are multiple tools for each step of the analysis, and their uses should be optimised and combined, based on the experimental design, metrics of the reads, and results of the tools (Y. Wu & Zhang, 2020). Recent tools enable integrated workflows of scRNA-seq data analysis, such as *Scanpy* and *Seurat*, which can be used for preprocessing and downstream analyses (Wolf et al., 2018).

1.3.3.1 Preprocessing

Preprocessing is an essential stage of the analysis that affects downstream analyses and yields different results when done with different tools (Soneson et al., 2021; You et al., 2021). Amplification bias, cell cycle effects, library size, and low RNA capture rate lead to background noise (Eraslan et al., 2019). During preprocessing, reads should be aligned to the respective genome, and noise should be cleared out via a method selected based on the conducted experimental methods for accurate analyses (Du et al., 2020; Janssen et al., 2023). Preprocessing includes count matrix generation by raw data processing methods such as ambient RNA and doublet removal, filtering cells that have low quality with quality control metrics, feature selection, integration, and dimensionality reduction (Heumos et al., 2023).

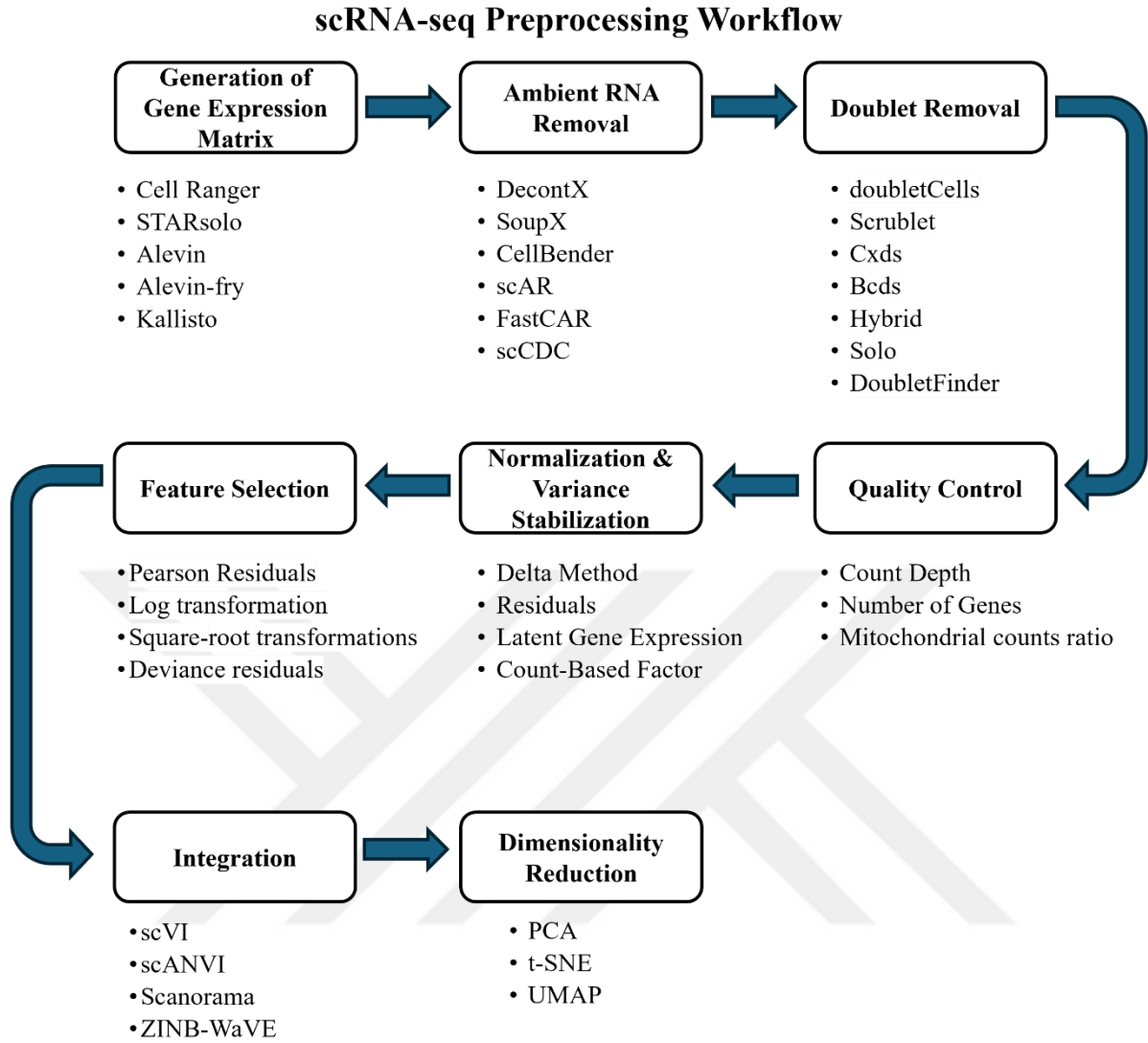


Figure 1.2 scRNA-seq preprocessing workflow with a list of available and widely used tools or methods for each step.

1.3.3.1.1 Generation of Gene Expression Matrix

Various scRNA-seq data preprocessing tools exist for aligning reads on respective genomes, estimation of gene expressions, and barcode and UMI detection (You et al., 2021). Some of the most common tools are *CellRanger*, *STARsolo*, *kallisto*, *alevin*, and *alevin-fry*, and their mapping and barcode and UMI detection performances differ significantly, thus affecting gene expression matrix (Brüning et al., 2022; Du et al., 2020; D. He et al., 2022; Kaminow et al., 2021; Srivastava et al., 2019; Zheng et al., 2017).

CellRanger and *STARsolo* use *STAR alignment*, which yields accurate results but requires a computationally intensive process (Dobin et al., 2013). To achieve better computational efficiency and faster run time, *kallisto*, *alevin*, and *alevin-fry* use pseudo-alignment, which can lead to missing false positive mapping artefacts compared to full alignment methods (Kaminow et al., 2021).

For barcode correction, *CellRanger*, *STARsolo*, and *kallisto* identify and correct sequencing errors using a whitelist, which is a list of barcode sequences in the assay kit, and filter based on hamming distance as first implemented by *CellRanger* (Kaminow et al., 2021; Melsted et al., 2021; Zheng et al., 2017). *Alevin* and *alevin-fry* do not use the provided list of barcodes but generate a permit list by filtering barcodes based on iterative knee thresholds that keep frequently observed barcodes (D. He et al., 2022).

CellRanger and *STARsolo* consider respective barcodes, UMIs, and annotated genes to remove PCR duplicates, allowing one mismatch in UMI and replacing error-prone UMIs with more abundant UMIs (Kaminow et al., 2021). *Alevin* and *alevin-fry* take account of the minimal set of transcripts based on equivalence classes; on the other hand, *kallisto* directly removes different sequences containing the same UMI (Brüning et al., 2022).

To assign reads to genes and their respective cells, *STARsolo*, *CellRanger*, and *kallisto* discard reads that have no unique mapping; however, *alevin* and *alevin-fry* assign reads having more than one equally potential mappings to all possible positions instead of discarding them (Brüning et al., 2022).

1.3.3.1.2 Ambient RNA Removal

Ambient RNAs are cell-free RNAs incorporated into droplets released from cells that have undergone apoptosis or cellular stress, contaminating the cell suspension (S. Yang et al., 2020). The high variability of ambient RNAs across the data can negatively influence downstream analyses; for instance, marker gene detection can be majorly misled by them unless they are removed (Janssen et al., 2023). Tools for discarding ambient RNAs include *DecontX*, *SoupX*, *CellBender*, *scAR*, *FastCAR*, and *scCDC* (Berg et al., 2023; Fleming et al., 2023; Sheng et al., 2022; W. Wang et al., 2024; S. Yang et al., 2020; Young & Behjati, 2020).

SoupX, *CellBender*, and *FastCAR* require count levels in empty droplets for background noise profiling to estimate the noise and remove the corresponding transcripts (Janssen et al., 2023). *DecontX* uses the Bayesian method to remove transcripts by comparing clusters within and with other clusters, and it optionally allows for noise filtering by an empty droplet list, too (Janssen et al., 2023). *scAR* is a deep learning model that first calculates the average expression in all droplets, then uses a multinomial distribution test to determine ambient profiles and discard ambient RNAs with these predicted profiles (Sheng et al., 2022). *scCDC* does not use empty droplets but uses estimates of contamination-causing genes by entropy divergence calculation, comparing expressions in clusters to detect ambient quantity (W. Wang et al., 2024).

1.3.3.1.3 Doublet Removal

Doublets refer to barcodes that include more than one cell, either heterotypic, which contains different cell types, or homotypic, which contains similar cell types and is more challenging to detect (Xi & Li, 2021).

Although methods were tried to prevent doublets during sample preparation, due to their limitations, they have not been optimised for scRNA-seq protocols (Wolock et al., 2019). Therefore, several doublet removal tools have been developed (Neavin et al., 2024). Based on a benchmark that compares nine different tools with multiple datasets, *doubletCells*, *Scrublet*, *cxds*, *bcds*, *hybrid*, *Solo*, and *DoubletFinder* successfully predicted both real and synthetic datasets (Xi & Li, 2021).

doubletCells generate synthetic doublets by summing up counts of two random cells to mimic a doublet, compare it with other cells in a principal component space and give doublet scores based on the distance (Xi & Li, 2021). *Scrublet* also uses the same approach but uses k-nearest neighbours to score them (Wolock et al., 2019). *scds* offers both *cxds*, co-expression-based doublet scoring, and *bcds*, binary classification-based doublet scoring, while *hybrid* combines both algorithms (Bais & Kostka, 2020). *cxds* use binarised expression data and binomial model for co-expression of genes to predict doublets, assuming that heterotypic doublets show unusual co-expressions. For *bcds*, synthetic cells are created and added to the count matrix, and after preprocessing, binary classification algorithms are used to score doublets based on synthetic cells. In the *hybrid* method, both algorithms are combined upon getting their results independently. *Solo* uses synthetic doublets for variational autoencoders to learn low-dimensional representations of gene expression and add another layer, the classification layer, to predict singlets and doublets (Bernstein et al., 2020). *DoubletFinder* generates synthetic doublets by getting random pairs' average gene expressions and using k-nearest neighbours in principal component space to predict doublets (McGinnis et al., 2019).

Also, some recent tools developed to analyse doublets specifically based on the possibility of persistent doublets in scRNA-seq protocols may imply there is juxtacrine

signalling between cells that do not dissociate easily and may hinder critical biological information about their interaction (Schiebout et al., 2023).

1.3.3.1.4 Quality Control

Quality control metrics, including count depth, the number of genes, and a fraction of counts from mitochondrial genes per barcode, are used to filter out low-quality barcodes (Heumos et al., 2023). Different problems can cause outlier barcodes; for instance, low count depth, low number of mapped genes, and a high fraction of mitochondrial gene counts may represent cells with broken membranes that cause a leak of cytoplasmic mRNAs, or very high counts may represent doublets (Luecken & Theis, 2019).

Due to the heterogeneity of transcriptional profiles from different samples and sequencing techniques, there is no definitive threshold for the metrics; thus, filtering should be done considering variations per sample; however, adaptive thresholding can be used based on median absolute deviation (MAD) (Subramanian et al., 2022). It was suggested that barcodes with metrics differing by five MADs could be considered outliers (Germain et al., 2020; Heumos et al., 2023).

1.3.3.1.5 Normalisation

Normalisation based on read depth and following variance stabilisation across cells is essential in comparing transcription levels within and across cells by reducing biological and technical variability (Vallejos et al., 2017). It was shown in a benchmark that the normalisation effects vary in the downstream analysis and shifted logarithm transformation, without counts per million, showed the best performance, but another method analytic approximation of Pearson residuals was also similarly efficient, particularly for low expressed genes (Ahlmann-Eltze &

Huber, 2023; Heumos et al., 2023). The shifted logarithm transformation approach uses logarithmic transformation by considering size factor based on cell variation, and shifted logarithm is applied to normalise data. The analytic approximation of Pearson residuals calculates expected counts of cells based on library size and average expression and calculates Pearson residual for each gene and use residuals that are normalised to differences in sequencing depths; yet it does not require pseudo count addition or log normalisation (Lause et al., 2021).

1.3.3.1.6 Feature Selection

High-throughput sequencing technologies allow for sequencing thousands of transcripts; however, it comes with noise and technical variations (P. Yang et al., 2021). To overcome this issue, several feature selection methods have been developed to determine genes that represent individual cells' characteristic features to make dimension reduction and downstream analyses more precise (Su et al., 2021).

Most methods use predetermined thresholding to select highly variable genes across data, but other approaches have also been developed to determine the number of genes as threshold and calculate variance by different modalities (C. Cheng et al., 2023; Su et al., 2021). A benchmarking showed that using Pearson residuals performed well compared to other common methods, such as log or square-root transformations after depth normalisation, deviance residuals, in terms of sensitivity in distinguishing small sets of cell types; however, it was noted that no selection of features also performed well as selecting features (Lause et al., 2021).

1.3.3.1.7 Integration

Handling, capturing times, equipment, sequencing techniques, and various other factors lead to significant variations or batch effects between samples of scRNA-seq data, and to address this issue, several tools for integration by removing batch effects were proposed (Tran et al., 2020). In addition, to investigate cellular mechanisms across species, batch effects can be reduced to a level where similar transcriptomic profiles could be extracted (Y. Song et al., 2023). Different methods use linear decomposition, similarity-based batch correction, and generative models using a variational autoencoder (Ryu et al., 2023).

Linear decomposition methods, such as *ComBat*, which uses an empirical Bayes approach in microarrays, relied on batch effect correction efforts for other sequencing techniques and were used in scRNA-seq analyses, too (Johnson et al., 2007; Ryu et al., 2023). Then, this approach was explicitly adopted for scRNA-seq data, such as *ZINB-WaVE*, by using the zero-inflated negative binomial model and taking account of scRNA-seq data features, dropouts, overdispersion and count distributions (Risso et al., 2018). However, compared to similarity-based batch corrections such as *Scanorama* and *scMerge*, it was ineffective and could not integrate some datasets (Y. Liu et al., 2021). Similarity-based approaches use reduced dimension spaces to determine similar cell types by distances of individual cells or clusters, which are used to determine variation due to batches (Ryu et al., 2023). Autoencoder-based models outperform similarity-based approaches in complex integration tasks (Luecken et al., 2022). These methods use neural networks to model and reduce the nonlinear characteristics of batch effects. Autoencoder-based models *scANVI* and *scVI* were shown to integrate complex datasets effectively (Hie et al., 2024). *scVI* uses stochastic optimisation and neural networks to identify distributions of similar cells and expression of genes to remove the batch effect (Lopez

et al., 2018). *scANVI* as an extension of *scVI*, allows cell type annotations to estimate cell types in the datasets and allow integration more efficiently (C. Xu et al., 2021).

1.3.3.1.8 Dimensionality Reduction

Dimension reduction is needed to visualise and investigate the topological distribution of the high dimensional structure of the scRNA-seq data (Heumos et al., 2023). It can be achieved with principal component analysis (PCA) for reducing the dimension and further reduction is done with nonlinear dimensionality reduction techniques like t-distributed stochastic neighborhood (t-SNE) or uniform manifold approximation and projection (UMAP) (der Maaten & Hinton, 2008; Linderman, 2021; McInnes et al., 2020).

PCA is widely used to reduce smaller principal components by preserving maximum variance (Luecken & Theis, 2019). However, it fails for two-dimensional visualisation due to the non-linear structure and noisy nature of scRNA-seq (Moon et al., 2019). t-SNE aims to minimise divergence between probability distributions between high and low dimensional spaces to determine the locations of the data points (Kobak & Berens, 2019). UMAP first constructs high-dimensional data and then optimises a low-dimensional graph to preserve the data's original structure by plotting in two dimensions (McInnes et al., 2018). t-SNE was the most accurate benchmark comparing dimensionality reduction methods, but UMAP was the most stable (Xiang et al., 2021).

1.3.3.2 Downstream Analyses

When the data is processed, transcriptomic profiles within and between cells can be analysed by various streamlined methods, including identifying clusters, determining differential gene expression, annotating cell types, and exploring the dynamic processes of individual cells

(Heumos et al., 2023). These analyses allow investigators to investigate cellular heterogeneity, regulatory mechanisms, and functional states of the cells in different conditions.

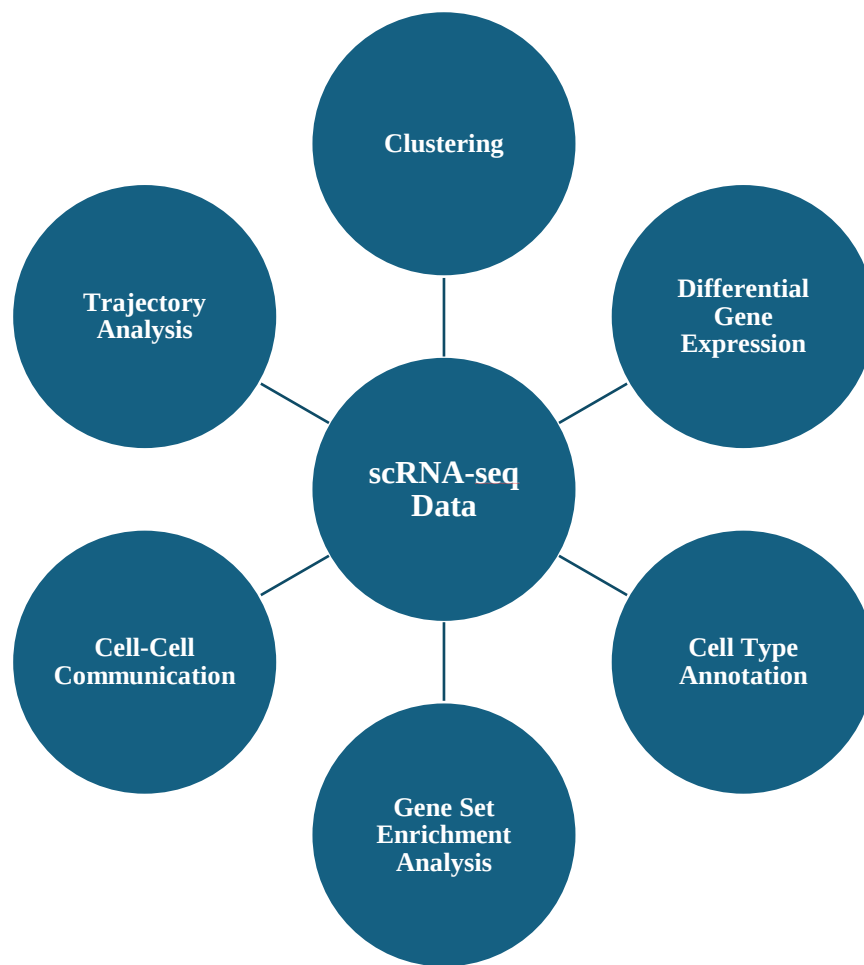


Figure 1.3 Approaches of scRNA-seq downstream analyses.

1.3.3.2.1 Clustering

Based on transcriptional profiles of individual cells, they can be clustered to identify and characterise different cell types and expression patterns, and methods can be categorised into four: k-means, hierarchical, density-based clustering, autoencoder and community detection (S. Zhang et al., 2023).

K-means clustering, a well-known methodology previously applied first to microarray-based expression studies, tries to find distinct subgroups based on a determined number of

clusters and adapted to high dimensional scRNA-seq data structure (Baker et al., 2021). Hierarchical clustering methods use bottom-up or top-down approaches to cluster cells (Feng et al., 2020). Density-based clustering generates clusters where graph data points are dense (Peyvandipour et al., 2020). Autoencoder methods, such as scVAE, use count data instead of principal components to estimate the clusters (Grønbech et al., 2020). Community detection-based algorithms, such as Louvain and Leiden, improved based on Louvain, using the k-nearest neighbors graph to identify subgroups (Traag et al., 2019). Although all the mentioned approaches have advantages and disadvantages, it is suggested that community-based detection methods, especially Leiden, to be used for clustering (Heumos et al., 2023; S. Zhang et al., 2023).

1.3.3.2.2 Differential Gene Expression Analysis

Differentially expressed genes can be determined by three main approaches: at the single-cell level, mixed models, which consider similarities and distinct features of the selection, and pseudobulk, where the counts are summed up to create a bulk RNA-seq-like profile (Crowell et al., 2020).

Single-cell level approaches for differential gene expression (DGE) analysis, such as the Wilcoxon rank sum test, compare each cell as statistically independent between two groups (Mou et al., 2019). However, this approach's false positive rates are higher than those from other methods (Junttila et al., 2022). Mixed models, such as *NEBULA*, take account of fixed effects for experimental conditions and random effects for variation across cells for DGE analysis (L. He, Davila-Velderrain, et al., 2021). Pseudobulk analysis methods use prior algorithms for bulk RNA-seq data, such as *edgeR*, *DESeq2*, and *limma*, for DGE analysis with aggregated counts, which offer high statistical power but lose single-cell resolution of the data

(Love et al., 2014; Ritchie et al., 2015; Robinson et al., 2010; Squair et al., 2021). While there is no consensus on the best-performing method, benchmarks are showing that pseudobulk-based methods provide the most accurate results, although some contradictory reports claim mixed models perform better (Gagnon et al., 2022; Junttila et al., 2022; Murphy & Skene, 2022; Squair et al., 2021; T. Wang et al., 2019; Zimmerman et al., 2021).

1.3.3.2.3 Cell Type Annotation

Cells that express specific genes or certain gene signatures can be used to classify the cells into distinct cell identities and subtypes by differentially expressed genes of clusters (Luecken & Theis, 2019). It can be done manually or using automated classifiers (Heumos et al., 2023).

Various automatic annotation tools use different approaches: marker-based classifications, label transferring, supervised classifications, and large language models (Hou & Ji, 2024; Pasquini et al., 2021). The marker-based methods, such as *CellAssign*, score the cells based on a given marker gene list (A. W. Zhang et al., 2019). Label transferring methods, such as *SingleR*, rely on statistical methods to compare annotated reference datasets and unannotated datasets for cell types for cell identity assignment (Aran et al., 2019). Supervised classifications, such as *CellTypist*, train data on previously annotated datasets to create a distribution model and use the model to assign identities to new datasets (Domínguez Conde et al., 2022; C. Xu et al., 2023). Large language models, such as *GPT-4*, also started to be used in annotation by providing differentially expressed genes (Hou & Ji, 2024). On the other hand, manual annotation can be done, and some databases provide marker lists for cell types (Franzén et al., 2019; X. Zhang et al., 2019). However, this could be time-consuming and not reproducible (Abdelaal et al., 2019). Hence, a three-step workflow is suggested: automatic

annotation, manual annotation and verification to ensure the automated cell annotations are missing cell identities or subtypes are assigned (Clarke et al., 2021).

1.3.3.2.4 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSE) is an overrepresentation method originally designed for microarray data to rank lists of genes to show statistical significance by using differential expressions of two different conditions or phenotypes (Subramanian et al., 2005). It is adapted to scRNA-seq analysis to compare single-cell level by comparing their absolute expression of genes in the list (Barbie et al., 2009). Various tools using different statistical approaches are in use for the enrichment of scRNA-seq data, including *GSVA*, which uses Kolmogorov-Smirnov like random walk statistics to calculate variation, and *AUCell*, which uses area under the curve to compare relative expressions (Aibar et al., 2017; Hänzelmann et al., 2013; Y. Zhang et al., 2020). Also, a framework *decoupler* allows access to both gene list databases and different enrichment tools, including *fast GSEA*, *GSVA*, and *AUCELL* (Badia-i-Mompel et al., 2022).

1.3.3.2.5 Cell-Cell Communication Analysis

Receptor ligand interactions and downstream effects of intercellular interactions can be analysed to understand tissue-level changes and impacts on regulatory mechanisms using scRNA-seq data based on curated databases or simulated interaction networks (Almet et al., 2021). There are tools that enable analysis at the single-cell level, such as *SoptSC*, or at the cluster level, such as *CellPhoneDB* (Efremova et al., 2020; S. Wang et al., 2019). Some frameworks like *LIANA* enable the use of different sources for receptor-ligand pairs and downstream networks with different approaches (Dimitrov et al., 2022).

1.3.3.2.6 Trajectory Analysis

The expression levels of individual cells or clusters represent a snapshot of the cellular mechanisms, but their dynamic states can be inferred by RNA velocity, pseudo-time, and lineage tracking methods (Heumos et al., 2023). Hence, regulatory networks that lead to changes in the transcriptomic profile can be inferred by calculating pseudo-time, lineage tracking, and RNA velocity (Nguyen et al., 2021; R. Singh et al., 2024).

Pseudo-time analysis allows the ordering of the cells based on the transition of the expression profiles in which the cells are in continuous processes such as disease or differentiation, either with or without contrast conditions (Hou et al., 2023). It can be calculated by tools that use dimension reduction first, such as *Monocle*, or gene-space methods which directly use expression data, such as Continuous-state Hidden Markov Models (CSHMM) (Ding et al., 2022; Van den Berge et al., 2020). RNA velocity methods compare unspliced and spliced readouts of transcripts to predict the direction and speed of the transitions of the cells (La Manno et al., 2018). There are various RNA velocity modelling methods, including static models, which assume constant rates for changes in transcripts, such as *velocity*, dynamic models, which assume variable rates over time, such as *scVelo*, and deep learning methods which use neural networks, such as *cellDancer* (Bergen et al., 2020; La Manno et al., 2018; S. Li et al., 2024). On the other hand, lineage-tracking methods are used to construct trajectories based on the progeny of cells over time and can be combined with different data acquisition techniques (C. Chen et al., 2022). There are tools to infer lineage tracking for specific experimental setups as well, for example, *CoSpar* is designed for scRNA-seq data with static barcoding and *LineageOT* for CRISPR/Cas9 based lineage tracing (Forrow & Schiebinger, 2021; Heumos et al., 2023; S.-W. Wang et al., 2022).

1.4 Use of Single-Cell RNA Sequencing in Liver Diseases

The human liver was sequenced to identify cell types and characterise their expression profiles in single-cell resolution, and novel subclasses of the liver cells were identified using scRNA-seq (MacParland et al., 2018). Furthermore, scRNA-seq has been widely used to investigate liver diseases, from fetal samples to hepatocellular carcinoma (Cao et al., 2020; Lu et al., 2022). These studies identified mechanisms of liver development, regeneration, immunity, subtypes of cellular identities, and responses against inflammation, fibrosis, and tumours (L. He, Lu, et al., 2021). Also, peripheral blood samples were used to characterise the immune cells of patients with liver diseases (Kim et al., 2021; Poch et al., 2021). On the other hand, some studies used previously published datasets by integration to characterize the transcriptomics profiles of liver cells that follow common profiles under similar conditions (Brancale & Vilarinho, 2021; J. Liu et al., 2023; Rocque et al., 2021).

1.4.1 Use of Single-Cell RNA Sequencing in MASLD

In recent years, samples from patients in the MASLD spectrum have been used to reveal underlying mechanisms and changes in the transcriptomics profile of non-parenchymal and parenchymal cells using scRNA-seq (Zhou et al., 2022).

For simple steatosis, Gulliams et al. used liver samples from obese and lean patients and with the combination of spatial mapping of proteomics and transcriptomics, they discovered differential zonation of immune cells, for example, lipid-associated macrophages were localized in steatotic regions (2022). Woestemeier et al. showed a change in transcriptional profiles in T cells and identified a helper 17 polarization during fibrosis development using scRNA-seq (2023). Using cirrhotic samples, Ramachandran et al. identified fibrotic niche-related cell subpopulations, including endothelial, mesenchymal, and

macrophage cells and potential interactions between cell identities such as platelet-derived growth factor receptors (PDGFR), tumor necrosis factor receptor superfamily member 12A (TNFRSF12A), and NOTCH signalling (2020). Another study combined scRNA-seq with three-dimensional imaging showed the cellular plasticity of cholangiocytes and hepatocytes during the advancement of the disease from the initial stages to the end stage of the disease (Gribben et al., 2024). Differential gene expressions of liver lymphatic endothelial cells in different chronic liver diseases and identified that the *IL13* signalling pathway was enriched in patients with (NASH) compared to those with hepatitis C virus infection using scRNA-seq (Tamburini et al., 2019). Another study revealed that liver stromal cells limit the maturation of monocytes into macrophages (Buonomo et al., 2022). Also, they showed macrophage differentiation away from cirrhosis-linked CD9+ scar-associated macrophages towards CD163+ Kupffer cell-like macrophages. The study by Buonomo et al., 2022, revealed that liver stromal cells restrict the maturation of monocytes into macrophages. Pfister et al. identified an exhausted CD8 and PD1 positive T cell accumulation in the liver, which was shown to cause impairment of immune surveillance in NASH patients(2021). Zhang et al. (2021) conducted a multiomics analysis, including scRNA-seq, on datasets from patients with cirrhosis. They uncovered maladaptation in vascular endothelial cells, where epigenetic reprogramming of these cells enriches fibrogenic TH17 cells, thereby contributing to liver fibrosis. Woestemeier et al. also identified polarized TH17 cell enrichment with advanced fibrosis (2023). Xiao et al. (2023) demonstrated distinct mechanisms between the stages of NAFLD and NASH. During the NAFL stage, hepatocytes exhibited metabolic adaptations. In contrast, gene expressions related to cell adhesion and migration were enriched at the NASH stage. Another study that explores hepatic stellate cells during advanced fibrosis in NASH detected autocrine signalling interactions that had emerged during fibrosis (S. Wang, Li, et al., 2023). Also, Filliol et al.

identified activated myofibroblastic hepatic stellate cells contributing to fibrosis and stiffness during NAFLD (2022).

On the other hand, there are integrative analyses where the available scRNA-seq datasets were investigated together to unveil cellular mechanisms between stages or cellular signalling between different cell identities. For example, in a recent study, datasets consisting of peripheral blood and liver samples were collected and integrated (Ma, Zhong, et al., 2023). Their analysis showed sulfhydryl oxidase 1 (QSOX1)-enriched fibroblast and interleukin-1 receptor accessory protein (IL1RAP)-enriched hepatocytes in MASLD spectrum patients. Also, their receptor-ligand interaction analysis showed differentially regulated pairs during disease progression.

1.4.2 Use of Single-Cell RNA Sequencing in MASLD with Mouse Models

Mouse models are frequently used in MASLD modelling, and various scRNA-seq datasets exist for different models using both dietary and chemically induced damage (Zhou et al., 2022). For example, Remmerie et al. used a Western diet on C68BL/6 mouse models for 12, 24, and 36 weeks to mimic different stages of MASLD, from simple steatosis to MASH (2020). Their model consistently showed similar attributes of MASLD, including increased liver weight, elevated liver enzymes, and glucose intolerance. Using scRNA-seq, they identified different subtypes of hepatic macrophages with heterogeneous profiles, especially in lipid metabolism. They discovered that bone marrow-derived macrophages replace resident Kupffer cells during disease progression. For example, in one study, C57BL6/J mice were used to treat carbon tetrachloride for two and four weeks to induce fibrosis in the liver (Terkelsen et al., 2020). They identified different subpopulations of hepatic stellate cells, liver sinusoidal endothelial cells, and Kupffer cells and their dynamic changes during fibrogenesis. They identified distinct states of hepatic stellate cells upon activation and their differential

transcriptomics profiles. Six subpopulations were identified for sinusoidal endothelial cells, and their gene expressions changed significantly during fibrosis. Notably, there was upregulation of vimentin, glycine-rich protein 2, and TSC22 domain family member 1 in the sinusoidal endothelial cells. Also, there are studies that explore conserved mechanisms between mice and humans in the MASLD spectrum. For example, Wang et al. identified receptor-ligand interactions in advanced fibrosis that are conserved between humans and mice (2023).

1.4.3 Use of Single-Cell RNA Sequencing in MASLD with Zebrafish Models

Although there are zebrafish models for MASLD in literature, there is only one available scRNA-seq data that can be related to MASLD (Morrison et al., 2022). scRNA-seq dataset Morrison et al., 2022, contains liver samples from eighteen-month-old mannose phosphate isomerase (mpi) knockout transgenic zebrafish, which were shown to have fibrotic conditions and scRNA-seq analysis NAFLD and captured both parenchymal and non-parenchymal cells. They identified distinct cell identities, including hepatic cells, hepatocytes, cholangiocytes, endothelial cells, and hepatic stellate cells and immune cells. Also, their integration with healthy human liver scRNA-seq showed that while zebrafish endothelial cells have distinct profiles, conservation of some subtypes and functions between species.

1.5 Web Tools for Single-Cell RNA Sequencing Analysis

There are web tools that enable scRNA-seq data analysis online and provide opportunities for hardware and coding-free use of frameworks. For example, *Shaoxia* allows the use of various tools on a web-based interactive platform to facilitate preprocessing, including mapping and dimension reduction, and downstream analyses, such as gene set enrichment and cell communication, and enable the process without programming expertise (Wei et al., 2024).

Also, some tools enable using specific tools over web servers without programming, such as *CellTypist* for cell type annotation (C. Xu et al., 2023). Moreover, tools such as *Single Cell Portal*, *UCSC Cell Browser*, *cellxgene*, which also allow annotation of the cells, and *ShinyCell* allow interactive visualisation of pre-processed datasets (Megill et al., 2021; Ouyang et al., 2021; Speir et al., 2021; Tarhan et al., 2023). They allow visualization of dimension reduction and gene expression, co-expression, and comparison of gene expression between labels of cells.

1.5.1 Web Tools for MASLD

There are web tools which enable scRNA-seq data analysis specific to MASLD-related datasets. *GepLiver*, which has integrated scRNA-seq data of liver samples in a large spectrum, including MASLD, allow users to explore changes in the gene expression patterns between conditions and samples, prediction of interacted cell identities by cell-cell communication analysis, abundance and enrichment of cell identities based for each sample (Z. Li et al., 2023). It was constructed by processing publicly available datasets in a standardised workflow. Although the integrated dataset is downloadable, the application does not allow custom downstream analyses.

ShinyCell was also used in two MASLD studies. In one of them, MASLD spectrum datasets were integrated and presented (Ma, Zhong, et al., 2023). *ShinyCell* was also deployed to present a recent dataset in a study that explores epithelial plasticity (Gribben et al., 2024). *Liver Cell Atlas* on the other hand allows the comparison of gene expression and visualisation of the authors' in-house mouse and human scRNA-seq datasets independently (Guilliams et al., 2022; Remmerie et al., 2020).

2. Aims and Rationale

Estimations from epidemiological studies show that at least 25% of the human population is affected by MASLD worldwide. The disease spectrum ranges from simple steatosis, which is reversible by lifestyle changes, to cirrhosis, which may cause liver failure and hepatocellular carcinoma and may require liver transplantation. Understanding the mechanisms behind the progression of the disease may allow for identifying targets for effective therapeutic strategies. Liver tissue is highly heterogeneous, and previous studies have shown that both parenchymal and non-parenchymal hepatic cells show different transcriptional profiles during MASLD when compared to healthy states.

Advancements in scRNA-seq techniques allow for a more comprehensive understanding of mechanistic changes specific to different cell types and their effects on diseases. An increasing number of scRNA-seq studies related to MASLD provide an opportunity to compare samples obtained from different stages. Integration of them may allow their direct comparison and downstream analyses. Also, data from animal models can be investigated to reveal conserved and distinct mechanisms of the disease. While some studies provide web tools for visualising and comparing gene expression levels between conditions and samples, there are no web tools for custom downstream analyses of MASLD datasets alone or integrated. Additionally, the preprocessing of the datasets significantly affects downstream analyses. A standardised preprocessing using recent approaches may reduce the noise and produce more accurate results than those previously reported in published studies.

This thesis work aims to develop a web tool that:

- Provides a database for MASLD spectrum human scRNA-seq datasets, which are pre-processed using a standardised workflow.

- Offers custom downstream analyses of individual and integrated scRNA-seq datasets without coding expertise.
- Offers downstream analyses of datasets from animal models and integrated with human datasets.
- Enable conserved marker detection with multi-species comparisons.



3. Materials and Methods

3.1 scRNA-seq Data Analyses

MASLD-related human scRNA-seq datasets and selected mouse and zebrafish model datasets based on the conditions with available raw counts were collected and pre-processed in a standardised workflow to construct the database and web tool. They were integrated with the respective MASLD stage. Then, using these datasets and downstream pipelines, a web tool was generated to allow further investigation of datasets, which is described in detail below.

3.1.1 Data Acquisition

scRNA-seq datasets, which have publicly available raw counts and unfiltered datasets, were gathered from *NCBI GEO* (Barrett et al., 2012). All available liver-sampled datasets were collected for human datasets, as listed in Table 3.1, while animal models representing different stages of the disease were collected in Table 3.2.

Table 3.1. Human scRNA-seq Datasets Used in This Study

Accession Code	Condition	# Condition	# Healthy	Reference
GSE136103	Cirrhosis	5	5	(Ramachandran et al., 2019)
GSE159977	Borderline NASH	3	4	(Pfister et al., 2021)
	NASH	3		
GSE181483	Cirrhosis	2	2	(H. Zhang et al., 2021)
GSE189600	NASH	3	3	(Xiao et al., 2023)
GSE190487	NAFLD	1	0	(Wallace S et al., 2022)
GSE192740	Obese	5	11	(Guilliams et al., 2022)
GSE212837	NASH	3	15	(S. Wang, Li, et al., 2023)
GSE174748	NAFLD	2	2	(Filliol et al., 2022)
GSE217235	NAFLD	3	0	(Woestemeier et al., 2023)
	NASH	3		
GSE129933	NASH	1	2	(Tamburini et al., 2019)
GSE169446	NASH	1	1	
GSE202379	NAFLD	7	6	(Gribben et al., 2024)
	NASH Without Cirrhosis	27		
	NASH With Cirrhosis	4		
	End Stage	15		
GSE168933	Cirrhosis	10	0	(Buonomo et al., 2022)
	Non-Cirrhotic	13		
Total		109	48	

Table 3.2. Animal Models scRNA-seq Datasets Used in This Study

Accession Code	Strain	Condition	# Condition	#Healthy	Reference
Mouse					
GSE156059	C57Bl/6J	Western Diet			
		12 Week	3	3	(Guilliams et al., 2022; Remmerie et al., 2020)
		24 Week	3		
36 Week	3				
Carbon Tetrachloride					
GSE145086	C57BL6/J	2 Week	3	3	(Terkelsen et al., 2020)
		4 Week	3		
Zebrafish					
GSE181987	AB	mpi ^{+/-}	3	3	(Morrison et al., 2022)
Total			30	9	

3.1.2 Preprocessing

Preprocessing of the datasets was done by using *Scanpy*, *Solo* from *scvi-tools* for doublet detection, *scVI* from *sci-tools* for integration, and *CellTypist* for annotation (Bernstein et al., 2020; Domínguez Conde et al., 2022; Gayoso et al., 2022; Hie et al., 2024; C. Xu et al., 2023). Preprocessing was performed using *Python 3.9.19* (Van Rossum & Drake, 2009). Orthologous gene name conversion was conducted using *Homologene* in *R 4.3.3* to make queries to *DIOPT* (Y. Hu et al., 2011; Mancarci & French, 2022; R Core Team, 2021). All datasets were pre-processed individually, and dataset integration was done with datasets *GSE136103*, *GSE181483*, *GSE145086*, and *GSE181987* in the context of fibrosis as a case study.

For basic filtering, genes present in less than 3 cells and cells with less than 200 total counts were removed (Satija et al., 2015). Mitochondrial genes were labelled based on their nomenclature. MADs were used to filter out outlier cells; exceeding 5 MADs for total counts, number of genes, and percentage of counts in the top 20 genes and 3 MADs for mitochondrial gene percentages were removed (Germain et al., 2020; Heumos et al., 2023). Then, doublets, which were predicted using *Solo* with default parameters, were removed (Bernstein et al., 2020). Ribosomal genes were labelled using genome annotations from *GRCz11* for zebrafish, *GRCh38* for humans, and *GRCm39* for mice (Schneider et al., 2017; *The Genome Reference Consortium*, 2024). Since some datasets used nuclei isolation protocols, ribosomal and mitochondrial genes were regressed from further analysis so as not to affect feature selection.

For integration, mouse and zebrafish orthologous genes were converted to human gene symbols using *DIOPT*, which uses multiple tools for orthologous gene prediction, by homologue with the selection of genes that the majority of the tools have predicted (Mancarci & French, 2022). The results table of *DIOPT* was first filtered based on the highest *DIOPT Score*. 14221 genes in the zebrafish dataset were converted to 9883 predicted human orthologous, and 13735 genes in the mouse dataset were converted to 12409 predicted human orthologous. The conversions were used for mouse and zebrafish datasets for integration and zebrafish dataset to be able to predict cell types with human healthy liver model from *CellTypist*.

Analytic Pearson residuals were used to select 2000 highly variable genes using sample identifier as the batch key parameter for individual datasets and dataset identifier for integrating different datasets (Lause et al., 2021; Wolf et al., 2018). The whole dimension of raw counts was kept in the dataset for further downstream analyses. Next, *scVI* was used for integration

(Gayoso et al., 2022). Batch key parameter was used to specify sample identifiers of individual datasets or dataset identifiers for dataset integration. Also, the categorical covariate keys parameter was used with the conditions of the samples. Latent representations from *scVI* results were used to compute neighbourhood graphs with default parameters. UMAP visualisation and clustering with Leiden were performed using these default parameters based on the calculated neighbour connectivities. Then, raw counts were log-normalized to 10000 counts per cell for *CellTypist*. To annotate clusters automatically, *CellTypist*, with default parameters, was used to use models trained on healthy human and mouse samples and results from majority voting deployed as cell identities (Domínguez Conde et al., 2022). The available healthy human liver model was used for human, zebrafish, and integrated datasets, and the available healthy mouse liver model was used with mouse datasets (Guilliams et al., 2022). Differentially expressed genes between cell identities were identified using the t-test by *Scanpy* (Wolf et al., 2018).

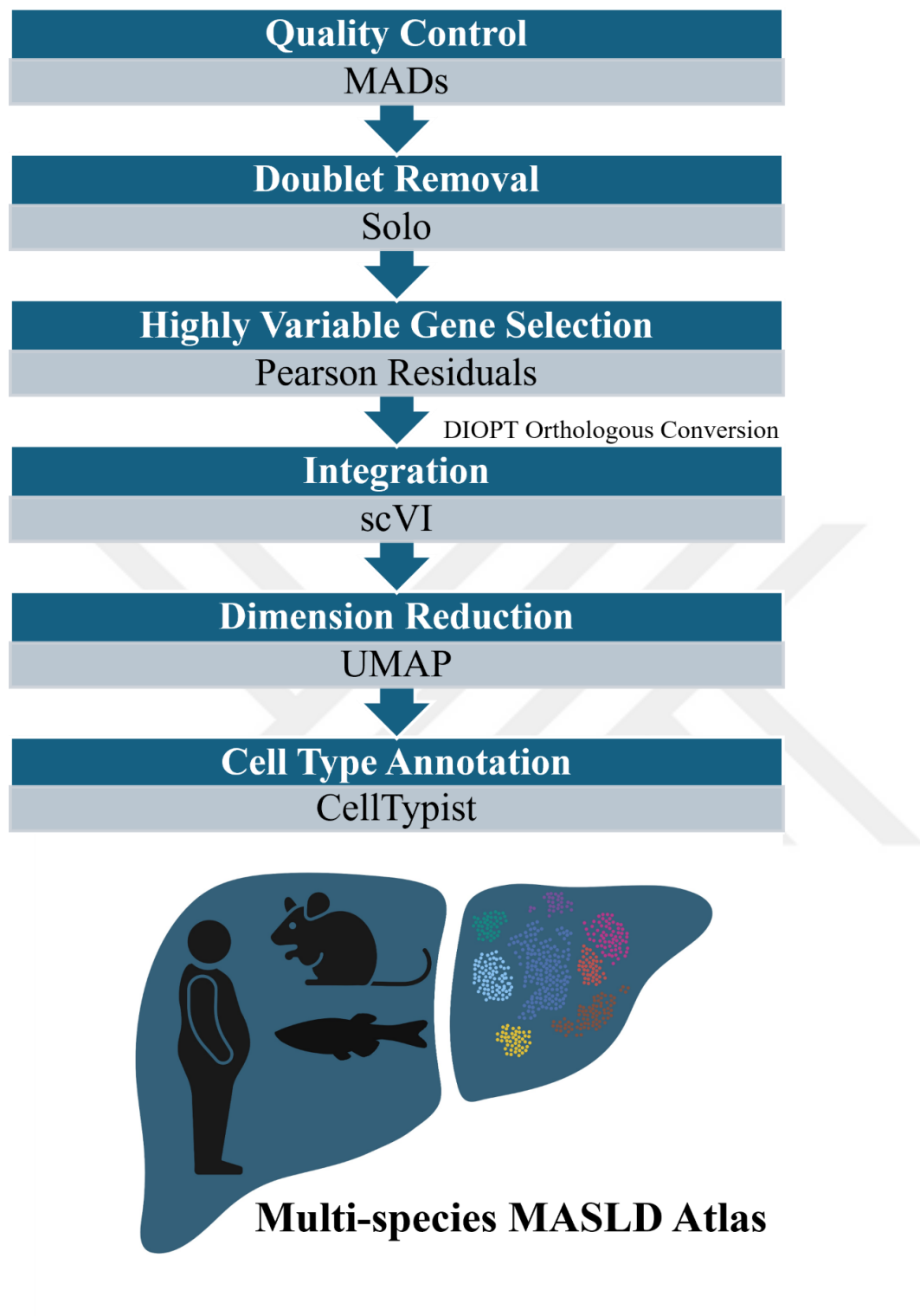


Figure 3.1 Conducted scRNA-seq data preprocessing workflow.

Table 3.3 Libraries Used in Preprocessing

Libraries	Version	Identifier	Reference
<i>Python</i>			
CellTypist	1.6.3	https://github.com/Teichlab/celltypist	(Domínguez Conde et al., 2022; C. Xu et al., 2023)
scvi-tools	1.1.2	https://github.com/scverse/scvi-tools	(Gayoso et al., 2022)
Scanpy	1.10.1	https://github.com/scverse/scanpy	(Wolf et al., 2018)
NumPy	1.26.4	https://github.com/numpy/numpy	(Harris et al., 2020)
pandas	2.2.2	https://github.com/pandas-dev/pandas	(McKinney, 2010)
PyTorch	2.2.2	https://github.com/pytorch/pytorch	(Paszke et al., 2017)
SciPy	1.13.0	https://github.com/scipy/scipy	(Virtanen et al., 2020)
<i>R</i>			
dplyr	1.1.4	https://github.com/tidyverse/dplyr	(Wickham, François, et al., 2023)
homologene	1.7.68.23.10.31	https://github.com/oganm/homologene	(Mancarci & French, 2022)
matplotlib	3.8.4	https://github.com/matplotlib/matplotlib	(Hunter, 2007)
seaborn	0.13.2	https://github.com/mwaskom/seaborn	(Waskom, 2021)

3.2 Multi-Species scRNA-seq Atlas of MASLD Shiny App Development

The web tool was constructed using *R Shiny* (W. Chang et al., 2024). The *reticulate* package was used to recruit Python in the *R Shiny* for using pre-processed objects and libraries in Python (Ushey et al., 2024). The web tool is designed to explore datasets based on standardised preprocessing and enable downstream analyses without the need for any coding; for this purpose, parameter selection was limited, and default parameters of tools were used in the web tool. Visualisation modalities include gene expression, cellular identities, and differentially expressed genes. Also, downstream analyses such as co-expression, gene enrichment, differential expression, and pseudo-bulk analyses can be done with the preprocessed datasets.

For UMAP and gene expression visualisations, *Scanpy* is used (Wolf et al., 2018). The co-expression of the selected genes is plotted using *ggpubr*, and Spearman and Pearson correlation is calculated using the base *cor* function of *R* (Kassambara, 2023; R Core Team, 2021). *Decoupler* was used for gene enrichment analyses on scRNA-seq datasets (Badia-i-Mompel et al., 2022). *fenr* was used for functional enrichment analysis with the results of differentially expressed genes for *GO*, *KEGG*, *WikiPathways*, and *Reactome* (Agrawal et al., 2024; Aleksander et al., 2023; Ashburner et al., 2000; Gierlinski, 2024; Kanehisa, 2000, 2019; Kanehisa et al., 2023; Milacic et al., 2024). The pseudo-bulk analysis is done with *decoupler*, and differential expression was conducted by *PyDESeq2* (Badia-i-Mompel et al., 2022; Muzellec et al., 2023). *decoupler* was also used for transcription factor analysis, using *CollectRI*, and pathway activity inference, using *PROGENy* and their visualisations (Müller-Dott et al., 2023; Schubert et al., 2018).

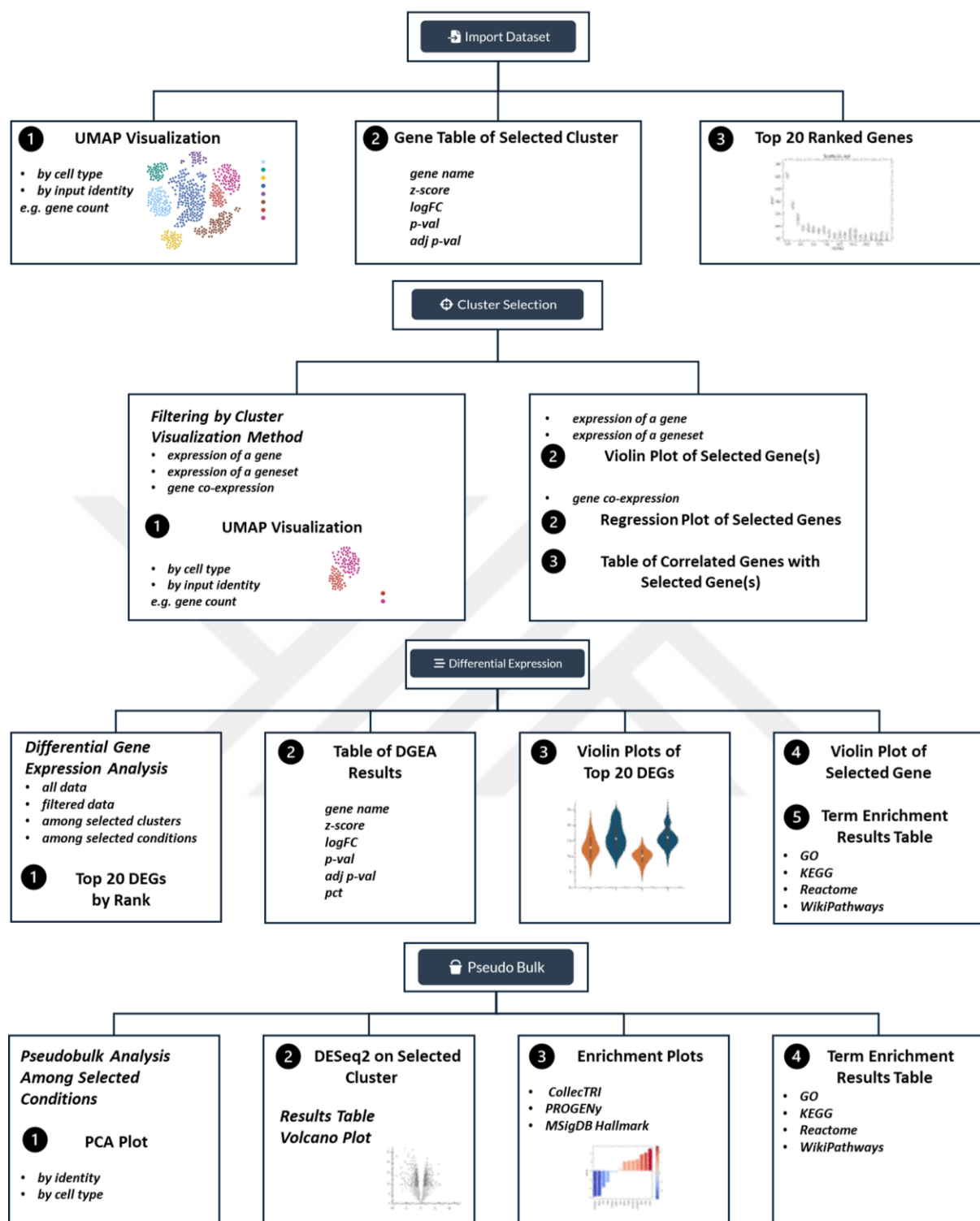


Figure 3.2 Workflow of *Multi-Species scRNA-seq Atlas of MASLD*.

Table 3.4. Libraries Used in Web Tool Development

Libraries	Version	Identifier	Reference
<i>R</i>			
shiny	1.8.1.1	https://github.com/rstudio/shiny	(W. Chang et al., 2024)
bslib	0.7	https://github.com/rstudio/bslib	(Sievert et al., 2024)
shinyBS	0.6.1.1	https://github.com/ebailey78/shinyBS	(Bailey & Detjens, 2021)
dplyr	1.1.4	https://github.com/tidyverse/dplyr	(Wickham, François, et al., 2023)
ggplot2	1.6.3	https://github.com/tidyverse/ggplot2	(Wickham, 2016)
shinydisconnect	0.1.1	https://github.com/daattali/shinydisconnect	(Attali, 2022)
shinycssloaders	1.0.0	https://github.com/daattali/shinycssloaders	(Andras Sali Luke Hass, 2023)
Shinyjs	2.1.0	https://github.com/daattali/shinyjs	(Attali, 2023)
reticulate	1.36.1	https://github.com/rstudio/reticulate	(Ushey et al., 2024)
DT	0.33	https://github.com/rstudio/DT	(Xie et al., 2023)
readr	2.1.5	https://github.com/tidyverse/readr	(Wickham, Hester, et al., 2023)
ggpubr	0.6.0	https://github.com/kassambara/ggpubr	(Kassambara, 2023)
shinyWidgets	0.8.6	https://github.com/dreamRs/shinyWidgets	(Perrier et al., 2023)
fenr	1.0.5	https://github.com/bartongroup/fenr	(Gierlinski, 2024)
stringr	1.5.1	https://github.com/tidyverse/stringr	(Wickham, 2023)
<i>Python</i>			
decoupler	1.6.0	https://github.com/saezlab/decoupler-py	(Badia-i-Mompel et al., 2022)
Scanpy	1.10.1	https://github.com/scverse/scanpy	(Wolf et al., 2018)
PyDESeq2	0.4.9	https://github.com/owkin/PyDESeq2	(Muzellec et al., 2023)

4. Results

4.1 Preprocessing of Individual and Integrated Datasets

Individual datasets were preprocessed in a standardized pipeline (**Figure 3.1-4.1, Appendix Figure 1-11**). In addition, human datasets containing only cirrhotic and healthy labels (GSE136103, GSE181483) and fibrotic mouse and zebrafish models (GSE145086, GSE181987) were integrated (**Figure 4.2**). UMAP was used to visualize the condition and annotations of the cells.



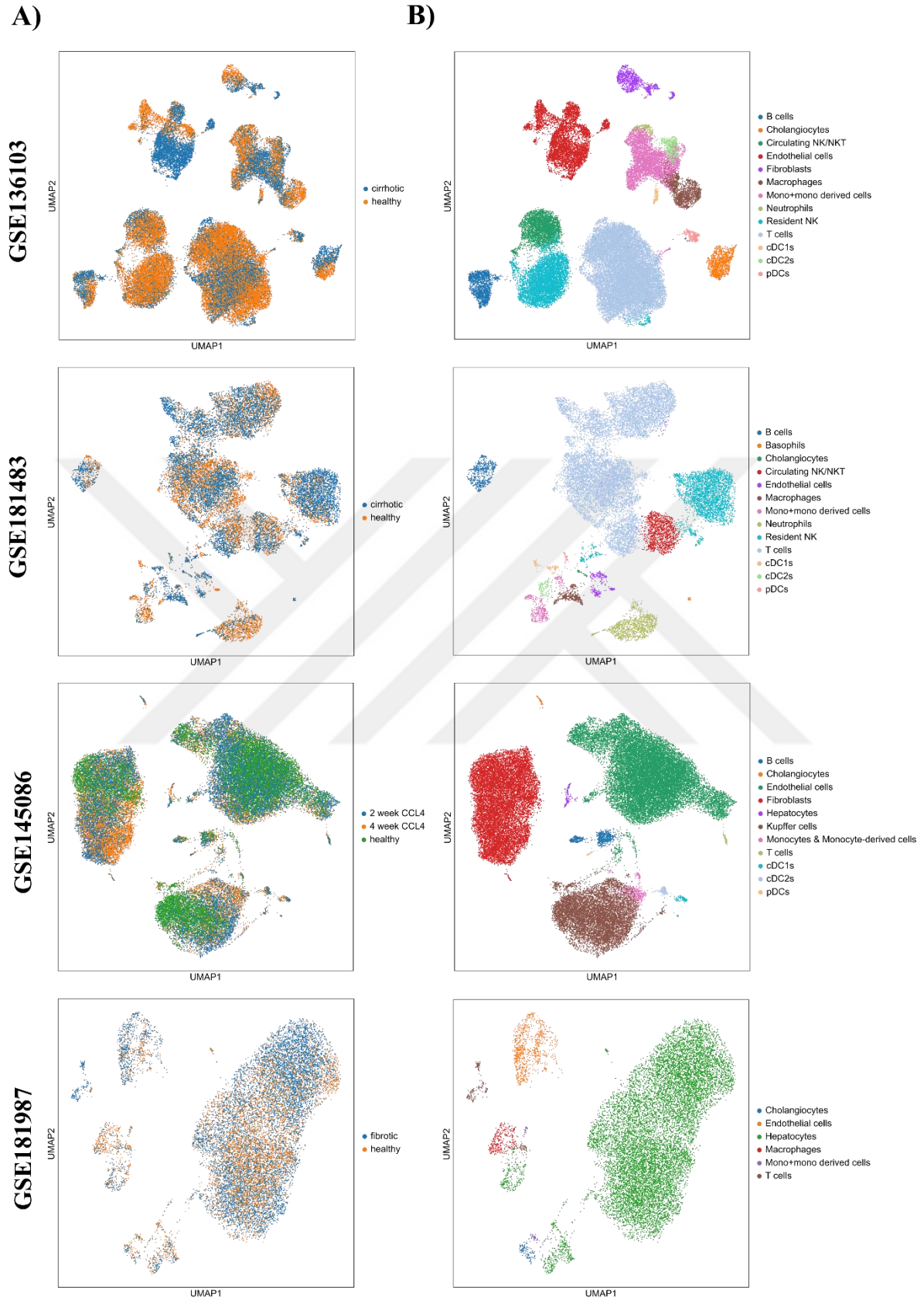


Figure 4.1 UMAP visualizations of preprocessed datasets used in dataset integration A) Categorized by conditions, B) Categorized by cell type annotations.

The cell type annotation of integrated datasets showed the presence of various cell types (**Figure 4.1**). GSE136103 human dataset, UMAP showed that several cell types were relatively more distinguished for cirrhotic and healthy cells, e.g., fibroblasts and cholangiocytes. Moreover, endothelial cells, one of this dataset's most populated cell types, also exhibited distinct cirrhotic and healthy subpopulations (**Figure 4.1A-B**). In the GSE181483 human dataset, endothelial cells exhibited dissimilarity between cirrhotic and healthy states. At the same time, multiple populations of nonparenchymal cells of the liver were differentiating between the cirrhotic and healthy cells, e.g., cDC1s and macrophages (**Figure 4.1A-B**). In the mouse CCL4 treated mice dataset, *GSE145086*, the time course of CCL4 allowed for visualization of increased fibrosis states in certain cell types. Indeed, fibroblasts effectively showed a gradient from healthy to 2 weeks and then to 4 weeks of treatment with CCL4. Hepatocytes also showed distinct clusters (**Figure 4.1A-B**). In the zebrafish scRNA-seq, *GSE181987*, cholangiocytes, although very limited in number and endothelial cells, were the cell types with a visible divergence between the fibrotic and healthy cells (**Figure 4.1A-B**).

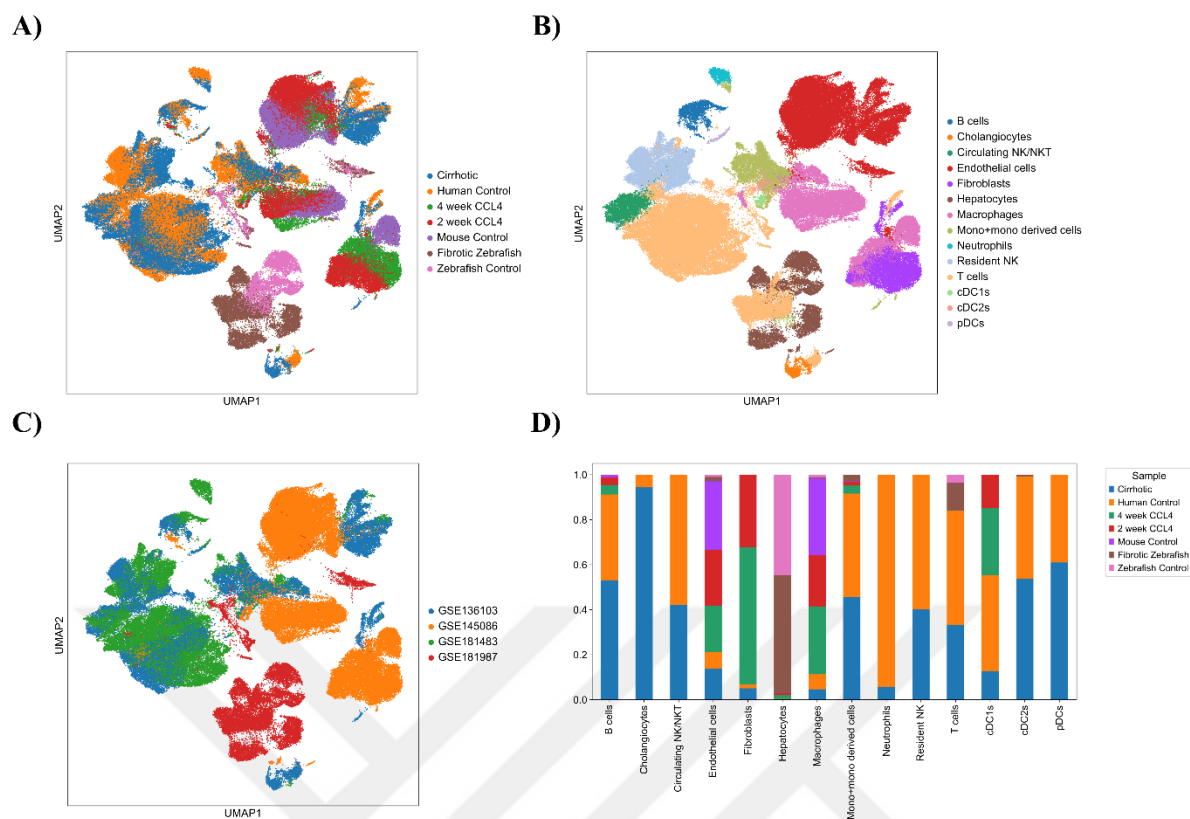


Figure 4.2 UMAP visualisation of cell type annotation, condition, and dataset identifiers of integrated datasets and distribution of samples per cell type A) Conditions of the cells on UMAP, B) Cell type annotations on UMAP, C) Datasets on UMAP, D) Proportions of sample conditions per cell type.

The integration of the fibrotic samples showed similar cell types common in datasets (**Figure 4.2**). Although some cell types were found in two species, endothelial cell types were present in all the datasets used in the integration.

Although individual datasets could be studied based on UMAP revealing the divergences between conditions and cell types, integration of these datasets could allow us to visualize all datasets simultaneously. Accordingly, in **Figure 4.2A**, where datasets from two humans, one mouse and one zebrafish healthy/fibrosis dataset were integrated, the species condition specificity could be observed more distinctly. When the conditions were mapped

against the cell types (**Figure 4.2B**), it became clear that endothelial cells comprised one of the largest clusters. Interestingly, all species were closely associated with this cluster (**Figure 4.3C**). However, zebrafish endothelial cells were relatively more distant from those of mouse and human datasets. The two human datasets were indistinguishable, and cirrhotic cells significantly diverged from healthy endothelial cells (**Figure 4.2A**). Mouse cells were placed next to the human cells within the endothelial cell cluster yet exhibited separation based on condition. Most interestingly, 4-week CCL4 treated endothelial cells were located at the edge of the 2-week CCL4 treated cells. Moreover, there was another endothelial subcluster in which treated cells were clustered apart from the healthy endothelial cells. The zebrafish endothelial cells were smaller when compared with those of the mouse and human datasets, and fibrosis and normal cells did not polarize but were relatively patchily clustered.

Fibroblasts were another cluster with large numbers of cells from the human and mouse datasets (**Figure 4.2B-C**). Endothelial cells from human and mouse datasets were relatively closer to each other yet all endothelial cells were still distinctly clustered. The cirrhotic cells were distinguishable among the human fibroblasts, while T cells were also closely placed next to them. On the other hand, there were more mouse fibroblasts, in close proximity to macrophages, which could be distinctly organized into control, 2 weeks of CCL4 and 4 weeks of CCL4 treatments (**Figure 4.2A-C**).

T cells and NK cells of human datasets were large in quantity in the integrated dataset and were clustered side by side (**Figure 4.2A-C**). The two datasets were relatively intermingled, suggesting that their integration was relatively successful. Identifying the cirrhotic T cells versus the healthy T cells in relatively well-distinguished multiple clusters of cells was possible. This suggested that there is heterogeneity among the cirrhotic cells of T-

cells. Circulating NK cells' condition specificity was better defined as in monocytes' case, suggesting they were relatively more homogeneous.

Integration has clarified the condition specificity for zebrafish hepatocytes, roughly divided into five clusters. In that regard, three of these five clusters belonged to fibrotic hepatocytes. **Figure 4.2D** showed that endothelial cells were relatively commonly presented in all four datasets and, hence, were further focused on in this thesis.

Altogether, these findings suggested that integrating datasets can help identify the condition-specific changes in the cell's transcriptomes when compared with individual dataset findings.

4.2 Multi-Species scRNA-seq Atlas of MASLD Shiny App

In this study, a web tool was developed to enable visualisation and downstream analyses of MASLD-related scRNA-seq datasets (**Figure 3.2**). To demonstrate the utilities of the tool, endothelial cells, which were common among four integrated datasets (**Figure 4.2**), were used. The application contains four main tabs with different functionalities: *Import Dataset*, *Cluster Selection*, *Differential Expression*, and *Pseudobulk*.

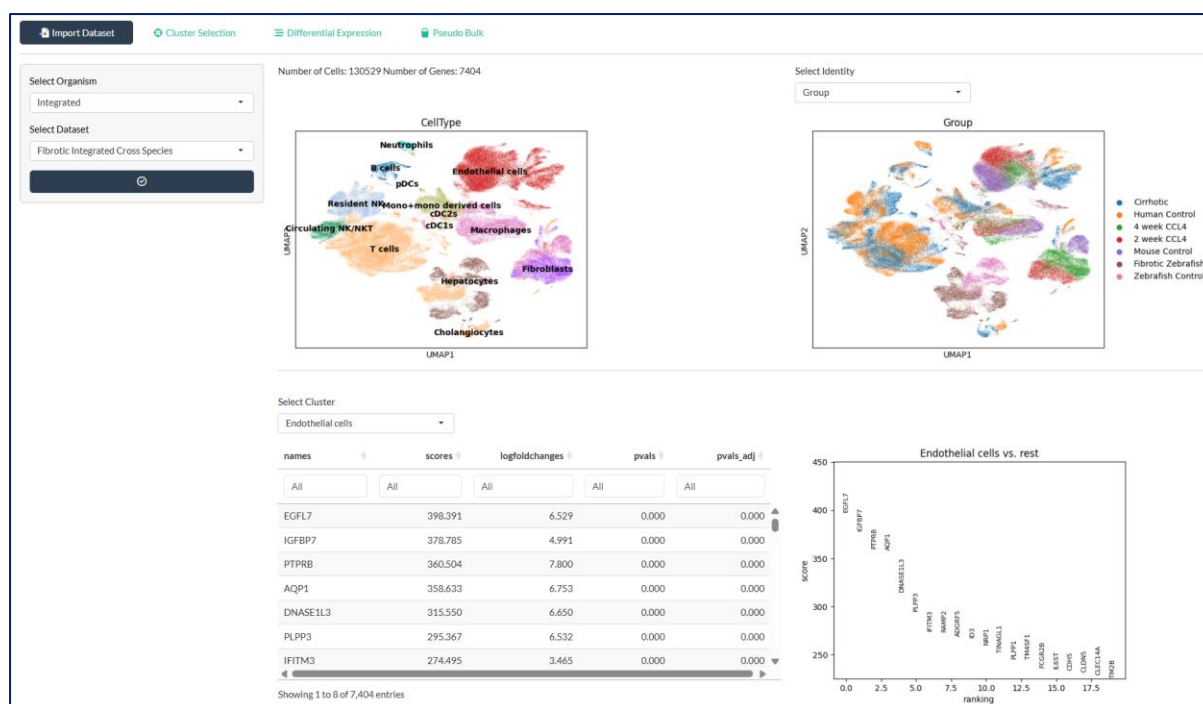


Figure 4.3 Import Dataset tab of *Multi-Species scRNA-seq Atlas of MASLD*.

The first tab of the app, *Import Dataset*, allows for the import of the selected dataset and the investigation of precalculated labels or annotations (**Figure 4.3**). The user must select the organism, i.e., Integrated, Human, Mouse, or Zebrafish, followed by the dataset. On the upper left side of the tab, cell type annotations on UMAP can be seen. On the upper right, labels of the cells can be visualized on UMAP. The lower half of the tab shows the differentially expressed genes of the selected cell type, where the left side provides the table, and the top 20 ranked genes based on the t-statistic can be seen on the right side (**Figure 4.3**). The table can be sorted according to the columns while a graphical representation of the ranked genes are provided with the ranking score.

When endothelial cells were selected, the top five ranked genes were EGF-like domain-containing protein 7 (EGFL7), insulin-like growth factor-binding protein 7 (IGFBP7),

receptor-type tyrosine-protein phosphatase beta (PTPRB), aquaporin 1 (AQP1), and deoxyribonuclease gamma (DNASE1L3) (**Figure 4.3**).

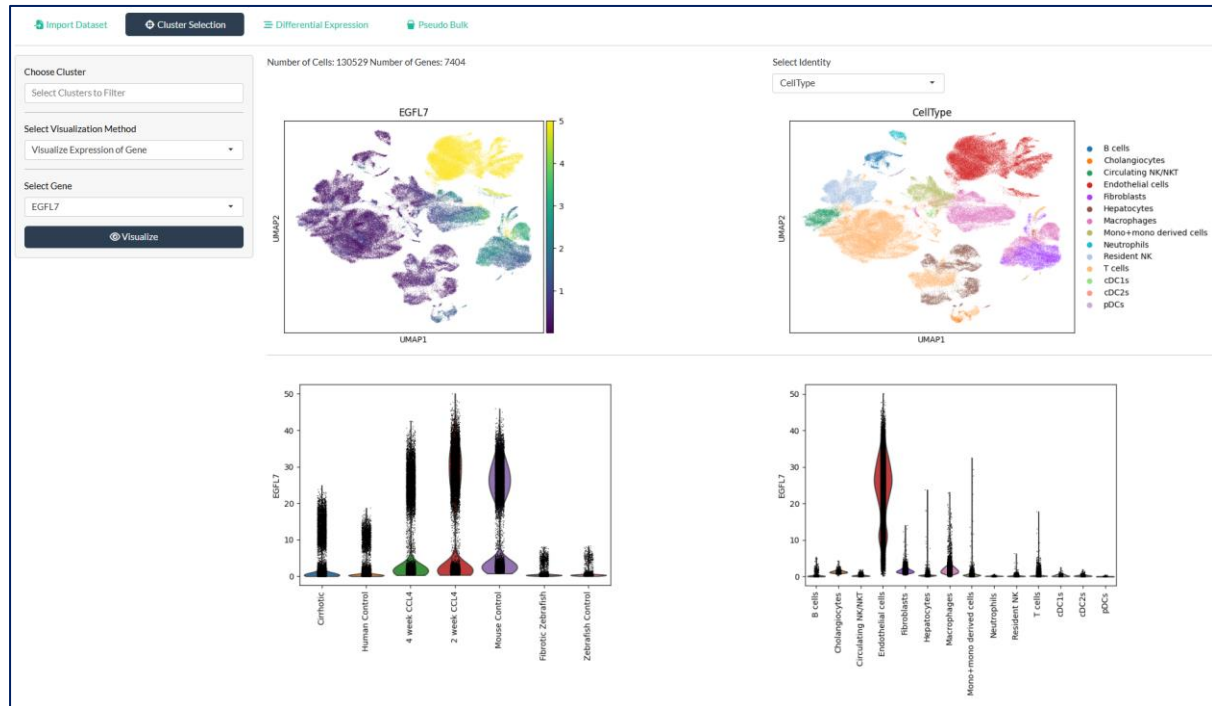


Figure 4.4 Cluster Selection Tab, Visualize Expression of Gene Section of *Multi-Species scRNA-seq Atlas of MASLD*.

The *Cluster Selection* tab was designed to allow users to investigate gene expression of selected gene with different methods to determine which cell type to investigate further (**Figure 4.4**). Single gene, gene set enrichment, and co-expression methods can be used to visualize the expression levels of the genes. The user can select a cell type on the upper left panel, although the default selection allows for visualization of all clusters. In addition, the user has to select a visualization method and the gene name.

Visualize Expression of Gene method allows the selection of a gene from lists of genes present in the dataset (**Figure 4.4**). UMAP visualization of the expression can be seen on the

upper left panel, indicated by colours ranging from blue to red based on the normalized count. The upper right panel also allows cell labels on UMAP to be investigated to compare expressions and labels. Violin plots based on condition and cell type can be visualized on the lower half of the tab.

EGFL7 was the top-ranked gene for endothelial cells, and its visualization showed significantly higher expression for endothelial cells than other cell types in the integrated dataset (**Figure 4.4**).





Figure 4.5 Cluster Selection Tab, Visualize Expression of Geneset Section of *Multi-Species scRNA-seq Atlas of MASLD*.

In the *Cluster Selection* tab, it is also possible to visualize the enrichment of genesets. To calculate and visualize the expression of genesets, *decoupler* was adopted for enrichment analysis (Badia-i-Mompel et al., 2022). Two enrichments can be done on the same page to compare the results simultaneously. **Figure 4.5** shows an example of geneset visualization where two separate gene lists can be entered to visualize. In this tab, on the left side, enrichment scores can be visualized on UMAP, and on the right-side the violin plots show the enrichment levels based on selected cell labels.

To validate that endothelial cells were labelled correctly, endothelial and all fibroblast cell markers were collected from *CellMarker 2.0* (C. Hu et al., 2023). Cancer-related markers

were filtered out to generate the marker lists (**Appendix Table 1**). The enrichment results showed endothelial and fibroblast labelled cells exhibit higher enrichment with their respective markers (**Figure 4.5**).



Figure 4.6 Cluster Selection Tab, Calculate Co-Expression Section of *Multi-Species scRNA-seq Atlas of MASLD*.

In the *Cluster Selection* tab, the third visualization method is *Calculate Co-Expression*. This allows users to select two genes in the dataset, calculate their correlations, and find the genes that correlate most closely with them in the dataset (**Figure 4.6**). Their expression levels can be seen on UMAP on the upper panel across all clusters or a selected cluster. On the lower left panel, plots show the correlations of the genes with the regression line plotted, and the correlation coefficient with the p-value is displayed. Spearman correlations between selected genes and all other genes in the dataset can be calculated on the lower right panel.

The correlation between *EGFL7* and smooth muscle α actin (*ACTA2*) was calculated to investigate further if fibroblast and endothelial cell types were dissociated. Their different distribution and statistically significant negative correlation show no associations between cells expressing them (**Figure 4.6**).

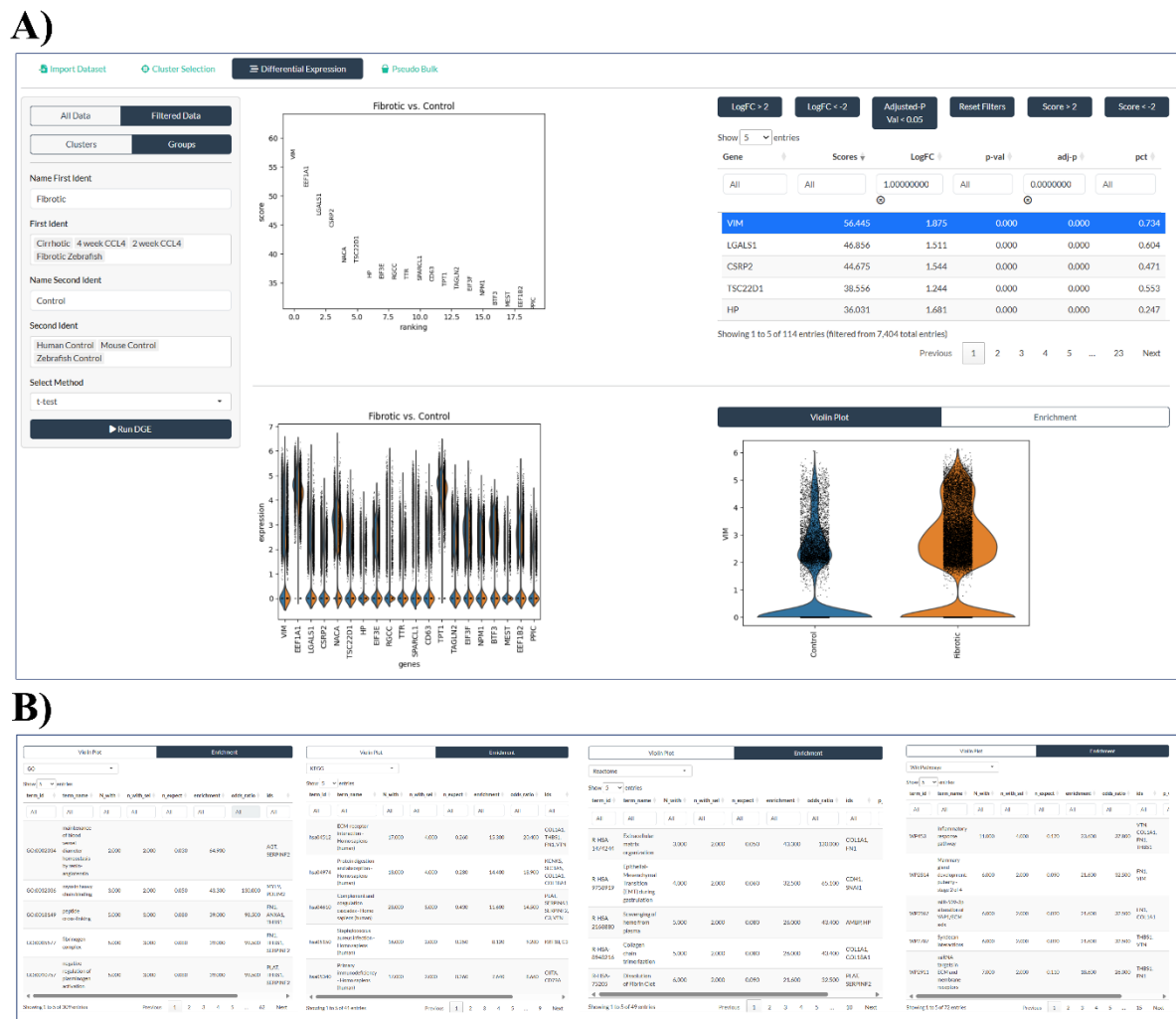


Figure 4.7 Differential Expression Tab of *Multi-Species scRNA-seq Atlas of MASLD*. A) Tab Overview, B) Term enrichment results.

The *Differential Expression* tab allows the calculation of differential gene expression statistics either by conditions or cell types (**Figure 4.7A**). It allows using *rank_genes_groups* of Scanpy with two methods: Wilcoxon and t-test. The user first need to identify the condition

and control groups by selecting on the list which shows the available conditions of the selected dataset. Once groups and the method are selected and *RUN DGE* is clicked on the analyses figures and tables are displayed on the tab. On the upper left, ranked genes based on ranking statistics can be seen, and below this, a violin plot is drawn to compare selection and reference. On the upper right, the results table can be investigated and filtered. On the lower right, the violin plot of the selected gene on the results table can be visualized via the violin plot. In addition, enrichment analysis can be done with filtered datasets using *fenr* on GO, KEGG, WikiPathways, and Reactome databases (Gierlinski, 2024).

To calculate differentially expressed genes between fibrotic samples from all three organisms, differential expression of the endothelial cells was selected in the Cluster Selection tab, and a t-test was used. Top-ranked five genes based on t-statistics after filtering adjusted p-value less than 0.5 and log fold change greater than one were: vimentin (VIM), galectin-1 (LGALS1), cysteine and glycine-rich protein 2 (CSRP2), TSC22 domain family protein 1 (TSC22D1), and haptoglobin (HP) (**Figure 4.7**).

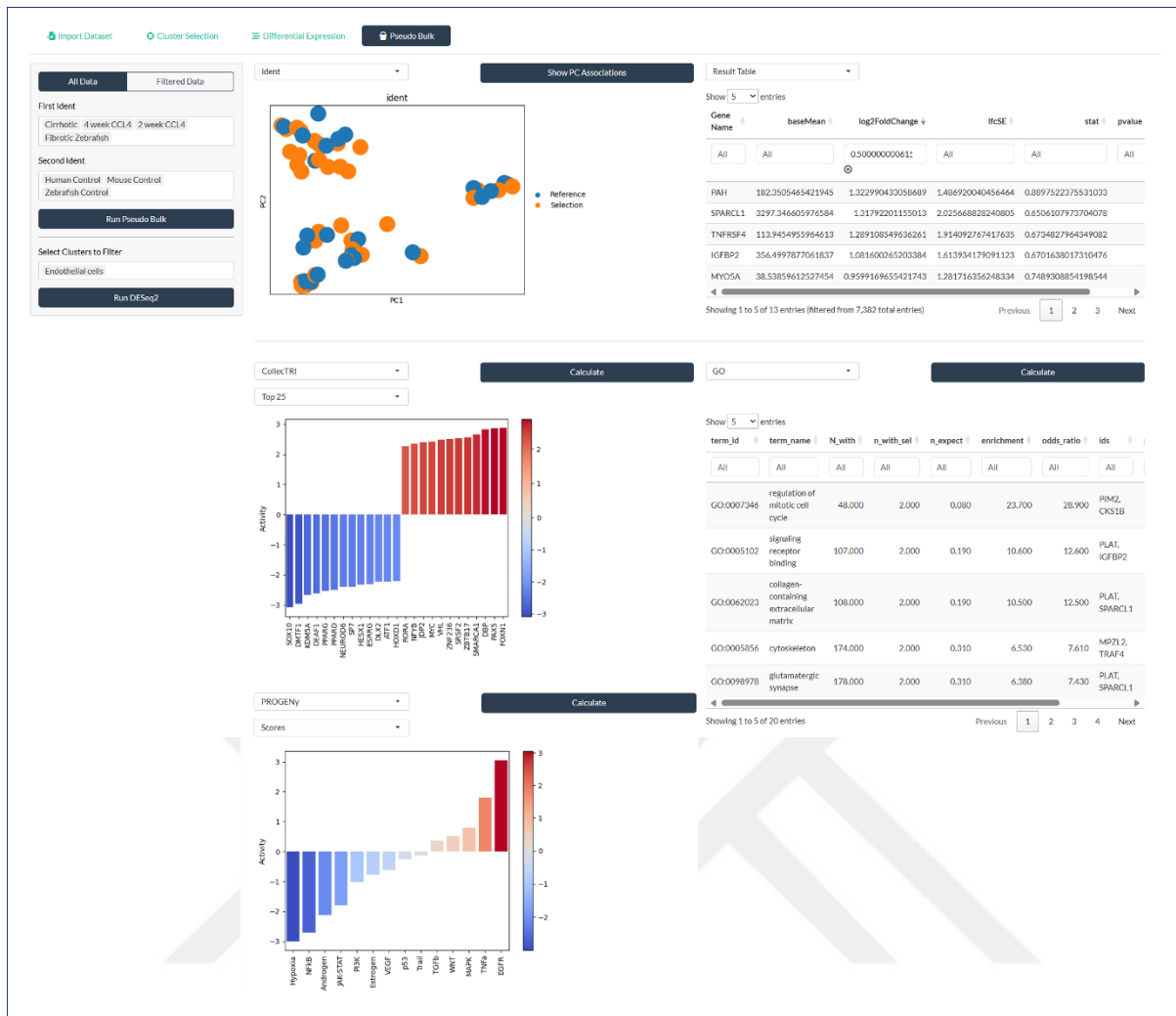


Figure 4.8 Pseudobulk Tab of *Multi-Species scRNA-seq Atlas of MASLD*.

The *Pseudobulk* tab allows pseudo-bulk analysis using *decoupler* (Badia-i-Mompel et al., 2022). In this tab, the user can use all the data or the filtered data, based on cell types and must select the groups to compare in addition to the cell cluster label. On the upper left, a *PCA* plot of the pseudo bulk samples can be seen where the reference and selection groups are colour coded. On the upper right, the results of the differential expression table can be investigated and filtered. As with the *Differential Expression Tab*, term enrichment can be conducted by filtering this table at the lower right panel. On the lower left panel, results of enrichment of pathways, *PROGENy* and transcriptional factors, *CollectRI* can be visualized (**Figure 4.8**).

The analysis was conducted with endothelial cells (**Figure 4.8**). After filtration of adjusted p-value less than 0.05 and log fold change greater than 0.5, top five upregulated genes were phenylalanine hydroxylase (PAH), SPARC-like protein 1 (SPARCL1), tumor necrosis factor receptor superfamily, member 4 (TNFRSF4), Insulin-like growth factor-binding protein 2 (IGFBP2), unconventional myosin-Va (MYO5A).



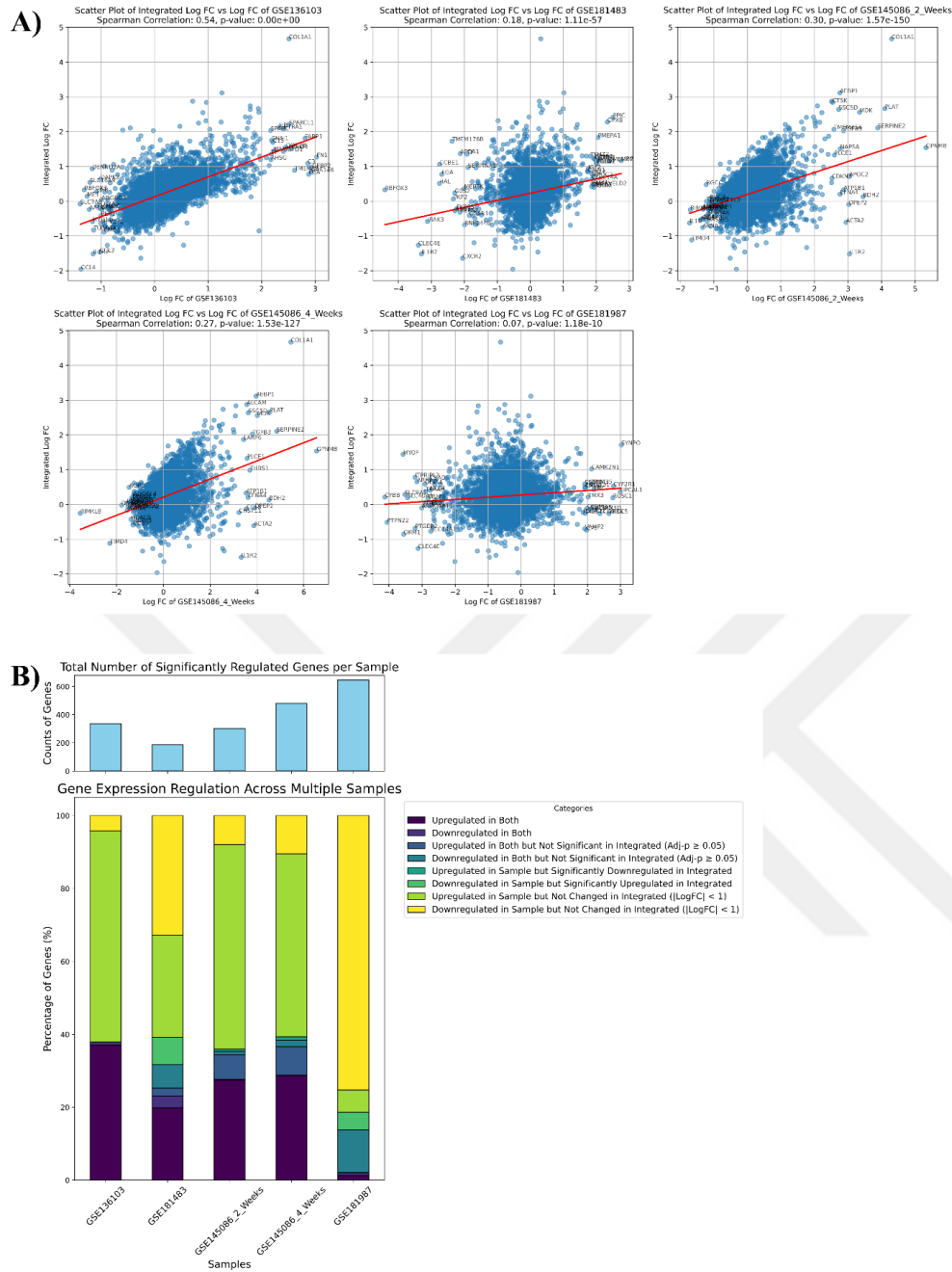


Figure 4.9 Comparison of Differentially Expressed Genes of Individually Preprocessed Datasets Against Integrated Dataset. **A)** Scatter plots of log fold changes of individual datasets on the x-axis and log fold changes of integrated datasets on the y-axis with Spearman correlation scores. **B)** Categorical comparison of significantly regulated genes ($|\log FC| > 1$ and adjusted p-val < 0.05) in the individual datasets and their regulation status in the integrated dataset where the first bar plot shows the number of significantly regulated genes and the second bar plot their categorical comparison with the integrated dataset.

In **Figure 4.9**, a comparison of differentially expressed genes conducted by endothelial cells and their categories can be seen for comparison of individually preprocessed datasets and integrated data. While *GSE136103* showed the highest correlation, the mouse dataset also exhibited a high correlation. On the other hand, the *GSE181483* and zebrafish dataset had low correlations. In **Figure 4.9B**, categories of the significant differentially expressed genes per dataset can be seen. In all the datasets, no change was seen when analysed on the integrated dataset among most of the genes that are differentially regulated during fibrosis. Also, the categorical comparison showed a similar trend for these two datasets (**Figure 4.9B**). A moderately low proportion of opposite trends was seen when gene expression profiles were analysed in the integrated dataset.

Table 4.1. Significantly Differentially Regulated Genes in Individual Datasets Used in Integration

Genes	Z-score	Log Fold Change	P-value	Adjusted P-value
SYNPO	Integrated: 41.76	Integrated: 1.71	Integrated: <0.01	Integrated: <0.01
	GSE136103: 28.41	GSE136103: 1.92	GSE136103: <0.01	GSE136103: <0.01
	GSE181483: 4.10	GSE181483: 1.33	GSE181483: <0.01	GSE181483: <0.01
	GSE145086(2-week-CCL4): 10.82	GSE145086(2-week-CCL4): 0.81	GSE145086(2-week-CCL4): <0.01	GSE145086(2-week-CCL4): <0.01
	GSE145086(4-week-CCL4): 24.82	GSE145086(4-week-CCL4): 1.62	GSE145086(4-week-CCL4): <0.01	GSE145086(4-week-CCL4): <0.01
	GSE181987: 4.38	GSE181987: 3.03	GSE181987: <0.01	GSE181987: <0.01
LRFN4	Integrated: 14.23	Integrated: 1.26	Integrated: <0.01	Integrated: <0.01
	GSE136103: 5.80	GSE136103: 1.12	GSE136103: <0.01	GSE136103: <0.01
	GSE181483: 3.47	GSE181483: 0.62	GSE181483: <0.01	GSE181483: <0.01
	GSE145086(2-week-CCL4): 4.38	GSE145086(2-week-CCL4): 0.82	GSE145086(2-week-CCL4): <0.01	GSE145086(2-week-CCL4): <0.01
	GSE145086(4-week-CCL4): 10.34	GSE145086(4-week-CCL4): 1.67	GSE145086(4-week-CCL4): <0.01	GSE145086(4-week-CCL4): <0.01
	GSE181987: 3.23	GSE181987: 1.38	GSE181987: <0.01	GSE181987: <0.01
MRC1	Integrated: -2.01	Integrated: -0.04	Integrated: <0.01	Integrated: <0.01
	GSE136103: -13.3	GSE136103: -0.67	GSE136103: <0.01	GSE136103: <0.01
	GSE181483: -5.83	GSE181483: -0.97	GSE181483: <0.01	GSE181483: <0.01
	GSE145086(2-week-CCL4): -42.63	GSE145086(2-week-CCL4): -0.87	GSE145086(2-week-CCL4): <0.01	GSE145086(2-week-CCL4): <0.01
	GSE145086(4-week-CCL4): -64.80	GSE145086(4-week-CCL4): -1.39	GSE145086(4-week-CCL4): <0.01	GSE145086(4-week-CCL4): <0.01
	GSE181987: -6.12	GSE181987: -2.77	GSE181987: <0.01	GSE181987: <0.01
TIMD4	Integrated: -25.31	Integrated: -1.12	Integrated: <0.01	Integrated: <0.01
	GSE136103: -4.52	GSE136103: -0.60	GSE136103: <0.01	GSE136103: <0.01
	GSE181483: -4.39	GSE181483: -1.06	GSE181483: <0.01	GSE181483: <0.01
	GSE145086(2-week-CCL4): -29.53	GSE145086(2-week-CCL4): -1.66	GSE145086(2-week-CCL4): <0.01	GSE145086(2-week-CCL4): <0.01
	GSE145086(4-week-CCL4): -36.64	GSE145086(4-week-CCL4): -2.29	GSE145086(4-week-CCL4): <0.01	GSE145086(4-week-CCL4): <0.01
	GSE181987: -8.40	GSE181987: -2.38	GSE181987: <0.01	GSE181987: <0.01

Table 4.1 shows commonly differentially expressed genes in endothelial cells in these four datasets after filtration of log fold changes greater than absolute 1 and adjusted p-values less than 0.05. Four genes were found to be significant among these datasets. The differential expression analysis of the integrated dataset also showed the same trends. While synaptopodin (*SYNPO*) and leucine rich repeat and fibronectin type III domain containing 4 (*LRFN4*) were upregulated in all of this sample, mannose receptor C-type 1 (*MRC1*) and T-cell immunoglobulin and mucin domain containing 4 (*TIMD4*) were downregulated.

5. Conclusion and Discussion and Future Perspectives

In this study, MASLD-related scRNA-seq datasets were preprocessed in a standardised workflow, and a web tool was developed to enable investigation and downstream analyses of these datasets. In addition, the integration of cirrhotic human samples and fibrotic animal models of zebrafish and mouse was performed to enable the investigation of fibrosis-related mechanisms among species. The application, as both a database and a tool for downstream analyses, is unique and includes various datasets from MASLD studies with different stages.

5.1 Preprocessing

The preprocessing workflow aimed to utilise best practices for scRNA-seq analysis in this study. The quality control matrices were standardised with MADs filtering (Germain et al., 2020; Heumos et al., 2023). *Solo* was used to predict doublets in the datasets, which showed efficiency in detecting doublets with high accuracy (Bernstein et al., 2020; Xi & Li, 2021). Pearson residual was used to detect the most variable 2000 genes before integration, which was shown in the literature to outperform other methods in accurately identifying variation (Heumos et al., 2023; Lause et al., 2021). *scVI*, which had shown high accuracy when integrating samples from multiple species, was used to integrate samples in the datasets and datasets from different species (Lopez et al., 2018; Y. Song et al., 2023). For annotation of the cells, *CellTypist*, which includes pre-trained datasets that were used in this study, also captured a variety of liver cell types (C. Xu et al., 2023).

Future studies should be conducted to compare the results using the different parameters and methods presented in this thesis. This is necessary to address the robustness

and reliability of the pipeline applied herein. For example, the preprocessing methodology used in this study has limitations and needs validation. Ambient RNA removal, which aims to discard reads from cell-free transcriptomes, was not conducted, but tools such as SoupX can be used (Young & Behjati, 2020). It may improve integration by removing noise, and cell-type classifications might improve. Moreover, the quality control matrices, which can vary among each sample, can be checked, and parameters can be set by considering plots of quality control metrics (Luecken & Theis, 2019). Although *Solo* was shown to be efficient, tools such as *DoubletFinder* and *scDblFinder* may be more accurate and stable than *Solo* (Germain et al., 2021; Xi & Li, 2021). They can be tested by comparing downstream results and adopted instead of *Solo*. In the literature, three steep workflows have been recommended to annotate the cells: automatic annotation, manual annotation, and verification (Clarke et al., 2021). However, in this study, only automatic annotation was used. Manual annotation, by referencing curated marker lists such as *CellMarker 2.0*, can be used for Leiden clusters and their differentially expressed genes, and verification can be further used to label the cells accurately (C. Hu et al., 2023). One important limitation of this study was that the *CellTypist* did not have a pre-trained model for zebrafish liver, and hence, zebrafish genes were labelled using a human liver model upon conversion of orthologous genes by *DIOPT* (Y. Hu et al., 2011; Mancarci & French, 2022; C. Xu et al., 2023). An annotated zebrafish liver scRNA-seq dataset can be used to train the model for better accuracy. Additionally, alternative gene name conversions can be obtained from *HGNC Symbol Reports* (Seal et al., 2023). One-to-many and many-to-one conversions can be conducted by comparing different databases. Possible changes can be tested one at a time and based on the accuracy of the preprocessing results, can be adopted. For example, differential gene expression analysis or resulting cell type annotations on UMAP can be used to compare the effects of parameters or changes in approach.

5.2 Multi-Species scRNA-seq Atlas of MASLD Shiny App

In **Figure 4.3-Figure 4.8**, the capabilities of the web tool are shown. The current version of the application allows visualization of individual genes, genesets, calculation of co-expression, differential gene expression, gene set enrichment, and pseudobulk analysis. The utilities of the application are the same for both individual and integrated datasets (**Figure 4.3, Appendix Figure 12**).

Datasets with cirrhotic and healthy labels were integrated in this study mainly to compare fibrotic animal models. All human datasets that are processed (**Table 3.1**) can be integrated to examine the mechanistic changes during the initiation and progression of the MASLD. This could be important since the findings may reveal potential targets for new therapeutic approaches and diagnostic markers that predict early or advanced liver disease prognosis. In addition, multiple Western diet-treated mouse model datasets can be integrated with datasets that include samples from patients that contain not only cirrhotic and healthy samples but also steatotic and MASH samples. This approach may reveal conserved features of the MASLD effects in non-cirrhotic stages. Also, in this study, available mouse scRNA-seq datasets were sampled from different mouse models, and not all models were included in the application (S. Cheng et al., 2023). Hence, they can be added to the web tool for further analysis of the datasets and integration with respective human samples based on the stage of the disease. During the development of the web tool, all MASLD spectrum scRNA-seq datasets were collected from NCBI GEO (Barrett et al., 2012). A literature survey can be done to follow and add new human MASLD scRNA-seq datasets to the web tool. Also, it has been shown that pseudobulk analysis outperforms single-cell differential expression analysis by providing

fewer false discoveries (Squair et al., 2021). However, in this study, their comparison was not conducted. The results from differential gene expression from *Differential Expression* and *Pseudobulk* tabs can be compared according to their log fold changes and p-values to validate the performances of the methods.

Also, additional downstream analyses can be enabled by the web tool. For example, trajectory analyses, such as pseudotime, can be conducted to understand differential mechanisms among the same cell types in different profiles (Van den Berge et al., 2020). For example, another analysis to be added could be cell-cell communication analysis, which could help investigate changes in the interactions in different stages of the disease using existing tools, e.g. *LIANA* (Dimitrov et al., 2023). On the other hand, although scRNA-seq datasets provide valuable information at the single-cell level, the location of cell types and proximity cannot be determined, and spatial transcriptomics may be used to map gene expression and cellular locations to understand MASLD (Wallace et al., 2022). Spatial transcriptomics datasets are available in both single-cell and spot levels in the MASLD spectrum (Guilliams et al., 2022; Moreno et al., 2024; Watson et al., 2024). Along with cell-cell communication analysis, a section exploring spatial datasets may provide a further understanding of cellular interactions during MASLD. Furthermore, Bulk RNA-seq datasets are also available in the MASLD spectrum, and studies investigate common changes by using published studies (Ma, Huang, et al., 2023; Ye et al., 2023). Results published in these studies or a novel signature database for differentially expressed genes in bulk RNA-seq datasets can be added to enable enrichment analysis on scRNA-seq datasets.

In addition, since the app does not specifically ask for parameter setup or complex prerequisites, an R library that constructs a web tool using *Shiny R* can also be developed instead. *ShinyCell*, already used to present scRNA-seq datasets for the MASLD spectrum, does not conduct downstream analyses such as differential gene expression, pseudo-bulk, and gene set enrichment (Ouyang et al., 2021). A new library may enable researchers to create a web tool with downstream analysis capabilities to present their preprocessed scRNA-seq datasets easily.

5.3 Multi-Species Integration of Fibrotic Samples

Integrated datasets included endothelial cells in common (**Figure 4.1B, Figure 4.2D**). The functionalities of the tabs were demonstrated by focusing on endothelial-labelled cells. The endothelial cells have already been shown to impact both the progression and resolution of liver fibrosis (Lafoz et al., 2020).

Top five ranked genes of endothelial cells were *EGFL7*, *IGFBP7*, *PTPRB*, *AQP1*, and *DNASE1L3* (**Figure 4.3**) and they were all shown to be expressed in liver endothelial cells (Da Silva et al., 2018; W. Li et al., 2023; Lit et al., 2024; Nichol & Stuhlmann, 2012; Q. Zhang et al., 2024). Also, the top-ranked gene for endothelial cells was *EGFL7*, which is highly specific to endothelial cells (Nichol & Stuhlmann, 2012). Also, as can be seen in **Figure 4.4**, the expression of it on the endothelial cells seems higher when compared to other cell types. When datasets were preprocessed independently, its expression was again specific to endothelial cells (**Figure 4.5, Appendix Figure 2**). The knockdown of *egfl7* in zebrafish has been shown to suppress endothelial cell proliferation, adhesion, migration, and vascular tube formation (J. Chen et al., 2024). Top-upregulated genes from differential expression analysis also show a

correlation with the literature. As such, the relationship between *VIM*, *LGALS1* and liver fibrosis was established (Angelini et al., 2023; Hermenean et al., 2022; P.-W. Wang et al., 2019). Also, pseudobulk analysis, **Figure 4.8**, showed *SPARCL1*, *TNFRSF4*, and *IGFBP2* in the top-upregulated genes, which were already shown to be related to liver fibrosis (B. Liu et al., 2021; Sun et al., 2018; Sveinbjornsson et al., 2022).

A comparison of results from individual and integrated datasets was conducted to investigate further the consistency of differentially expressed genes (**Figure 4.9**). While *GSE136103* and the mouse dataset showed a high correlation for log fold changes, the zebrafish and *GSE181483* datasets showed a low correlation (**Figure 4.9A**). Also, when categorical comparison of differentially expressed genes, again datasets showed similar trends (**Figure 4.9B**). However, when the composition of the endothelial cells is investigated in **Figure 4.2c-D**, the contribution of these two datasets is low compared to other datasets. The sample sizes can cause false discoveries during the statistical test (Mascha & Vetter, 2018). Also, their biological differences might cause this divergence. So, since the web tool enables the same analyses for individual datasets, validation for specific datasets can be conducted. On the other hand, subsampling of datasets with respect to cell types can be conducted to reveal more consistent changes in common with adjusting sample size contribution of datasets (D. Song et al., 2022). However, to identify common genes, the app can be used to investigate datasets or compare two different organisms in the integrated dataset, including their conditions.

When comparing the integrated dataset results with the individual datasets, it became evident that cross-species or dataset-specific differences significantly influenced the formation of cell clusters. Distinct clusters specific to each dataset rather than consisting of mixed cells,

such as mouse versus human versus zebrafish endothelial cells, were seen (**Figure 4.2B**). This outcome was expected because analysing an individual dataset independently provides more detailed information about genes, resulting in more homogeneous clusters. While species-specific separation was observed within the endothelial cell populations, the two human datasets were successfully intermingled, indicating effective integration when multiple datasets from the same species were combined (**Figure 4.2B**). Furthermore, when analysed separately, the integrated human datasets exhibited greater dissociation of condition labels compared to the individual human studies (**Figure 4.1A**). This suggests that integration enhances the differentiation of conditions within the same species, providing a more nuanced understanding of the data.

Individual investigation of four integrated datasets revealed 4 differentially regulated genes in endothelial cells: upregulation of *SYNPO*, *LRFN4* and downregulation of *MRC1*, *TIMD4* (**Table 4.1**). Their regulation was also the same in the integrated dataset's differential gene expression analysis results. Also, the log fold changes of the upregulated genes in the mouse dataset were higher in 4-week CCL4 treated samples than in 2-week CCL4 treated samples. The 4-week CCL4 treatment for downregulated genes also showed lower log fold changes. This result might indicate that the advancement of fibrosis might have a relation with them. *SYNPO* was shown to be upregulated in cirrhosis, chronic hepatitis B, and hepatitis B virus-related hepatocellular carcinoma (Zhao et al., 2022). In another study, it is also upregulated in hepatocellular carcinoma patients (Moriyama et al., 2022). In an MDR2 knock-out mouse model, it was also shown to be upregulated at the precancerous stage compared to the control liver (Stoyanov et al., 2015). On the other hand, *TIMD4* was shown to be downregulated in both fibrotic patients and mouse models, thioacetamide (TAA) and CCL4-treated models (Petrenko et al., 2024). While direct relations with fibrosis and *LRFN4* and

MRC1 were not found, *MRC1* expression was downregulated in liver endothelial cells in ageing mice (Almanzar et al., 2020). It is also known that tissue fibrosis is highly associated with ageing (Selman & Pardo, 2021). So, these common changes might help reveal conserved mechanisms during liver fibrosis among humans, mice, and zebrafish. They can be tested on other datasets to see if the changes are consistent between stages, and further validation might be accompanied by in vitro experiments.

These results suggested that the app could reveal significant mechanisms by specific cell types and helped combine three species to reveal that endothelial cells were distinct from the other cell types. They also showed that there are species-specific differences in endothelial cells. Furthermore, since the integration pipeline uses datasets from different animal models, conserved mechanisms among species can be investigated. New insights might be gained to understand the occurrence and progression of the MASLD. Also, the selected markers can be validated on animal models. On the other hand, since there is a lack of zebrafish scRNA-seq datasets that can be interpreted in the MASLD spectrum, new experiments may significantly contribute to understanding the conserved mechanisms of the disease among species.

To conclude, this study provides a novel web tool for exploring and conducting downstream analyses of the human, mouse, and zebrafish scRNA-seq datasets in the MASLD spectrum. Also, the integration of human samples with respective animal models allows the investigation of conserved and differential mechanisms of the disease. The common mechanisms might enable the development of more advanced animal models that represent the pathogenesis and impacts of the disease.

REFERENCES

- Abdelaal, T., Michielsens, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, 20(1), 194. <https://doi.org/10.1186/s13059-019-1795-z>
- Agrawal, A., Balci, H., Hanspers, K., Coort, S. L., Martens, M., Slenter, D. N., Ehrhart, F., Digles, D., Waagmeester, A., Wassink, I., Abbassi-Daloui, T., Lopes, E. N., Iyer, A., Acosta, J. M., Willighagen, L. G., Nishida, K., Riutta, A., Basaric, H., Evelo, C. T., ... Pico, A. R. (2024). WikiPathways 2024: next generation pathway database. *Nucleic Acids Research*, 52(D1), D679–D689. <https://doi.org/10.1093/nar/gkad960>
- Ahlmann-Eltze, C., & Huber, W. (2023). Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, 20(5), 665–672. <https://doi.org/10.1038/s41592-023-01814-1>
- Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11), 1083–1086. <https://doi.org/10.1038/nmeth.4463>
- Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., Van Auken, K., ... Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *GENETICS*, 224(1). <https://doi.org/10.1093/genetics/iyad031>
- Alexander, M., Loomis, A. K., Fairburn-Beech, J., van der Lei, J., Duarte-Salles, T., Prieto-Alhambra, D., Ansell, D., Pasqua, A., Lapi, F., Rijnbeek, P., Mosseveld, M., Avillach, P., Egger, P., Kendrick, S., Waterworth, D. M., Sattar, N., & Alazawi, W. (2018). Real-world data reveal a diagnostic gap in non-alcoholic fatty liver disease. *BMC Medicine*, 16(1), 130. <https://doi.org/10.1186/s12916-018-1103-x>
- AlJanahi, A. A., Danielsen, M., & Dunbar, C. E. (2018). An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Molecular Therapy - Methods & Clinical Development*, 10, 189–196. <https://doi.org/10.1016/j.omtm.2018.07.003>
- Almanzar, N., Antony, J., Baghel, A. S., Bakerman, I., Bansal, I., Barres, B. A., Beachy, P. A., Berdnik, D., Bilen, B., Brownfield, D., Cain, C., Chan, C. K. F., Chen, M. B., Clarke, M. F., Conley, S. D., Darmanis, S., Demers, A., Demir, K., de Morree, A., ... Zou, J. (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817), 590–595. <https://doi.org/10.1038/s41586-020-2496-1>
- Almet, A. A., Cang, Z., Jin, S., & Nie, Q. (2021). The landscape of cell–cell communication through single-cell transcriptomics. *Current Opinion in Systems Biology*, 26, 12–23. <https://doi.org/10.1016/j.coisb.2021.03.007>
- Ampuero, J., Aller, R., Gallego-Durán, R., Banales, J. M., Crespo, J., García-Monzón, C., Pareja, M. J., Vilar-Gómez, E., Caballería, J., Escudero-García, D., Gomez-Camarero, J., Calleja, J. L., Latorre, M., Albillos, A., Salmeron, J., Aspichueta, P., Lo Iacono, O.,

- Francés, R., Benlloch, S., ... Romero-Gómez, M. (2018). The effects of metabolic status on non-alcoholic fatty liver disease-related outcomes, beyond the presence of obesity. *Alimentary Pharmacology & Therapeutics*, 48(11–12), 1260–1270. <https://doi.org/10.1111/apt.15015>
- Andras Sali Luke Hass, D. A. (2023). *shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating*. <https://cran.r-project.org/web/packages/shinycssloaders/index.html>
- Angelini, G., Panunzi, S., Castagneto-Gissey, L., Pellicanò, F., De Gaetano, A., Pompili, M., Riccardi, L., Garcovich, M., Raffaelli, M., Ciccoritti, L., Verrastro, O., Russo, M. F., Vecchio, F. M., Casella, G., Casella-Mariolo, J., Papa, L., Marini, P. L., Rubino, F., le Roux, C. W., ... Mingrone, G. (2023). Accurate liquid biopsy for the diagnosis of non-alcoholic steatohepatitis and liver fibrosis. *Gut*, 72(2), 392–403. <https://doi.org/10.1136/gutjnl-2022-327498>
- Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2), 163–172. <https://doi.org/10.1038/s41590-018-0276-y>
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29. <https://doi.org/10.1038/75556>
- Attali, D. (2022). *shinydisconnect: Show a Nice Message When a 'Shiny' App Disconnects or Errors*. <https://cran.r-project.org/web/packages/shinydisconnect/index.html>
- Attali, D. (2023). *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*. <https://cran.r-project.org/web/packages/shinyjs/index.html>
- Badia-i-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., & Saez-Rodriguez, J. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, 2(1). <https://doi.org/10.1093/bioadv/vbac016>
- Bailey, E., & Detjens, E. (2021). *shinyBS: Twitter Bootstrap Components for Shiny*. <https://cran.r-project.org/web/packages/shinyBS/index.html>
- Bais, A. S., & Kostka, D. (2020). scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics*, 36(4), 1150–1158. <https://doi.org/10.1093/bioinformatics/btz698>
- Baker, D. N., Dyjack, N., Braverman, V., Hicks, S. C., & Langmead, B. (2021). Fast and memory-efficient scRNA-seq k-means clustering with various distances. *ACM-BCB ... : The ... ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, Computational Biology and Biomedicine, 2021*. <https://doi.org/10.1145/3459930.3469523>
- Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., ... Hahn, W. C. (2009).

- Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269), 108–112. <https://doi.org/10.1038/nature08460>
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1), D991–D995. <https://doi.org/10.1093/nar/gks1193>
- Bayoumi, A., Grønbaek, H., George, J., & Eslam, M. (2020). The Epigenetic Drug Discovery Landscape for Metabolic-associated Fatty Liver Disease. *Trends in Genetics*, 36(6), 429–441. <https://doi.org/10.1016/j.tig.2020.03.003>
- Benedé-Ubieto, R., Cubero, F. J., & Nevzorova, Y. A. (2024). Breaking the barriers: the role of gut homeostasis in Metabolic-Associated Steatotic Liver Disease (MASLD). *Gut Microbes*, 16(1). <https://doi.org/10.1080/19490976.2024.2331460>
- Berg, M., Petoukhov, I., van den Ende, I., Meyer, K. B., Guryev, V., Vonk, J. M., Carpaij, O., Banchemo, M., Hendriks, R. W., van den Berge, M., & Nawijn, M. C. (2023). FastCAR: fast correction for ambient RNA to facilitate differential gene expression analysis in single-cell RNA-sequencing datasets. *BMC Genomics*, 24(1), 722. <https://doi.org/10.1186/s12864-023-09822-3>
- Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12), 1408–1414. <https://doi.org/10.1038/s41587-020-0591-3>
- Bernstein, N. J., Fong, N. L., Lam, I., Roy, M. A., Hendrickson, D. G., & Kelley, D. R. (2020). Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. *Cell Systems*, 11(1), 95–101.e5. <https://doi.org/10.1016/j.cels.2020.05.010>
- Berzigotti, A., Tsochatzis, E., Boursier, J., Castera, L., Cazzagon, N., Friedrich-Rust, M., Petta, S., & Thiele, M. (2021). EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis – 2021 update. *Journal of Hepatology*, 75(3), 659–689. <https://doi.org/10.1016/j.jhep.2021.05.025>
- Bessone, F., Razori, M. V., & Roma, M. G. (2019). Molecular pathways of nonalcoholic fatty liver disease development and progression. *Cellular and Molecular Life Sciences*, 76(1), 99–128. <https://doi.org/10.1007/s00018-018-2947-0>
- Brancale, J., & Vilarinho, S. (2021). A single cell gene expression atlas of 28 human livers. *Journal of Hepatology*, 75(1), 219–220. <https://doi.org/10.1016/j.jhep.2021.03.005>
- Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S., & John, D. (2022). Comparative analysis of common alignment tools for single-cell RNA sequencing. *GigaScience*, 11. <https://doi.org/10.1093/gigascience/giac001>
- Buonomo, E. L., Mei, S., Guinn, S. R., Leo, I. R., Peluso, M. J., Nolan, M.-A., Schildberg, F. A., Zhao, L., Lian, C., Xu, S., Misdraji, J., Kharchenko, P. V., & Sharpe, A. H. (2022). Liver stromal cells restrict macrophage maturation and stromal IL-6 limits the differentiation of cirrhosis-linked macrophages. *Journal of Hepatology*, 76(5), 1127–1137. <https://doi.org/10.1016/j.jhep.2021.12.036>
- Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., Zager, M. A., Aldinger, K. A., Blecher-Gonen, R., Zhang, F., Spielmann, M., Palis, J., Doherty, D.,

- Steemers, F. J., Glass, I. A., Trapnell, C., & Shendure, J. (2020). A human cell atlas of fetal gene expression. *Science*, 370(6518). <https://doi.org/10.1126/science.aba7721>
- Cervantes-Alvarez, E., Vilatoba, M., Limon-de la Rosa, N., Mendez-Guerrero, O., Kershenobich, D., Torre, A., & Navarro-Alvarez, N. (2022). Liver transplantation is beneficial regardless of cirrhosis stage or acute-on-chronic liver failure grade: A single-center experience. *World Journal of Gastroenterology*, 28(40), 5881–5892. <https://doi.org/10.3748/wjg.v28.i40.5881>
- Chambers, J. C., Zhang, W., Sehmi, J., Li, X., Wass, M. N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S. E., Coin, L. J., Deng, G., Gieger, C., Heard-Costa, N. L., Hottenga, J.-J., Kühnel, B., Kumar, V., Lagou, V., Liang, L., ... Kooner, J. S. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature Genetics*, 43(11), 1131–1138. <https://doi.org/10.1038/ng.970>
- Chang, C., Li, H., & Zhang, R. (2023). Zebrafish facilitate non-alcoholic fatty liver disease research: Tools, models and applications. *Liver International*, 43(7), 1385–1398. <https://doi.org/10.1111/liv.15601>
- Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web Application Framework for R*. <https://shiny.posit.co/>
- Chen, B., Zheng, Y.-M., & Zhang, J.-P. (2018). Comparative Study of Different Diets-Induced NAFLD Models of Zebrafish. *Frontiers in Endocrinology*, 9. <https://doi.org/10.3389/fendo.2018.00366>
- Chen, C., Liao, Y., & Peng, G. (2022). Connecting past and present: single-cell lineage tracing. *Protein & Cell*, 13(11), 790–807. <https://doi.org/10.1007/s13238-022-00913-7>
- Chen, J., Ding, J., Li, Y., Feng, F., Xu, Y., Wang, T., He, J., Cang, J., & Luo, L. (2024). Epidermal growth factor-like domain 7 drives brain lymphatic endothelial cell development through integrin $\alpha\beta 3$. *Nature Communications*, 15(1), 5986. <https://doi.org/10.1038/s41467-024-50389-8>
- Cheng, C., Chen, W., Jin, H., & Chen, X. (2023). A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell–Cell Communication. *Cells*, 12(15), 1970. <https://doi.org/10.3390/cells12151970>
- Cheng, S., Zou, Y., Zhang, M., Bai, S., Tao, K., Wu, J., Shi, Y., Wu, Y., Lu, Y., He, K., Sun, P., Su, X., Hou, S., & Han, B. (2023). Single-cell RNA sequencing reveals the heterogeneity and intercellular communication of hepatic stellate cells and macrophages during liver fibrosis. *MedComm*, 4(5). <https://doi.org/10.1002/mco2.378>
- Cheng, X., Chen, J., Liu, S., & Bu, S. (2024). Assessing Causal Relationships Between Periodontitis and Non-alcoholic Fatty Liver Disease: A Two-Sample Bidirectional Mendelian Randomisation Study. *Oral Health & Preventive Dentistry*, 22, 189–202. <https://doi.org/10.3290/j.ohpd.b5395053>
- Clarke, Z. A., Andrews, T. S., Atif, J., Pouyababar, D., Innes, B. T., MacParland, S. A., & Bader, G. D. (2021). Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature Protocols*, 16(6), 2749–2764. <https://doi.org/10.1038/s41596-021-00534-0>

- Crowell, H. L., Sonesson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., & Robinson, M. D. (2020). muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1), 6077. <https://doi.org/10.1038/s41467-020-19894-4>
- Cusi, K., Isaacs, S., Barb, D., Basu, R., Caprio, S., Garvey, W. T., Kashyap, S., Mechanick, J. I., Mouzaki, M., Nadolsky, K., Rinella, M. E., Vos, M. B., & Younossi, Z. (2022). American Association of Clinical Endocrinology Clinical Practice Guideline for the Diagnosis and Management of Nonalcoholic Fatty Liver Disease in Primary Care and Endocrinology Clinical Settings. *Endocrine Practice*, 28(5), 528–562. <https://doi.org/10.1016/j.eprac.2022.03.010>
- Da Silva, I., Barroso, M., Moura, T., Castro, R., & Soveral, G. (2018). Endothelial Aquaporins and Hypomethylation: Potential Implications for Atherosclerosis and Cardiovascular Disease. *International Journal of Molecular Sciences*, 19(1), 130. <https://doi.org/10.3390/ijms19010130>
- De, A., & Duseja, A. (2020). Natural History of Simple Steatosis or Nonalcoholic Fatty Liver. *Journal of Clinical and Experimental Hepatology*, 10(3), 255–262. <https://doi.org/10.1016/j.jceh.2019.09.005>
- der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).
- Dimitrov, D., Schäfer, P. S. L., Farr, E., Rodriguez Mier, P., Lobentanzer, S., Dugourd, A., Tanevski, J., Ramirez Flores, R. O., & Saez-Rodriguez, J. (2023). LIANA+: an all-in-one cell-cell communication framework. *BioRxiv*. <https://doi.org/10.1101/2023.08.19.553863>
- Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P. L., Nagai, J. S., Boys, C., Ramirez Flores, R. O., Kim, H., Szalai, B., Costa, I. G., Valdeolivas, A., Dugourd, A., & Saez-Rodriguez, J. (2022). Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature Communications*, 13(1), 3224. <https://doi.org/10.1038/s41467-022-30755-0>
- Ding, J., Sharon, N., & Bar-Joseph, Z. (2022). Temporal modelling using single-cell transcriptomics. In *Nature Reviews Genetics* (Vol. 23, Issue 6, pp. 355–368). Nature Research. <https://doi.org/10.1038/s41576-021-00444-7>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., King, H. W., Mamanova, L., Huang, N., Szabo, P. A., Richardson, L., Bolt, L., Fasouli, E. S., Mahbubani, K. T., Prete, M., Tuck, L., ... Teichmann, S. A. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594). <https://doi.org/10.1126/science.abl5197>
- Driessen, S., Francque, S. M., Anker, S. D., Castro Cabezas, M., Grobbee, D. E., Tushuizen, M. E., & Holleboom, A. G. (2023). Metabolic dysfunction-associated steatotic liver disease and the heart. *Hepatology*. <https://doi.org/10.1097/HEP.0000000000000735>

- Du, Y., Huang, Q., Arisdakessian, C., & Garmire, L. X. (2020). Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment. *G3 (Bethesda, Md.)*, 10(5), 1775–1783. <https://doi.org/10.1534/g3.120.401160>
- Eeda, V., Patil, N. Y., Joshi, A. D., & Awasthi, V. (2024). Advancements in metabolic-associated steatotic liver disease research: Diagnostics, small molecule developments, and future directions. *Hepatology Research*. <https://doi.org/10.1111/hepr.14008>
- Efremova, M., Vento-Tormo, M., Teichmann, S. A., & Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols*, 15(4), 1484–1506. <https://doi.org/10.1038/s41596-020-0292-x>
- Eng, J. M., & Estall, J. L. (2021). Diet-Induced Models of Non-Alcoholic Fatty Liver Disease: Food for Thought on Sugar, Fat, and Cholesterol. *Cells*, 10(7), 1805. <https://doi.org/10.3390/cells10071805>
- Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1), 390. <https://doi.org/10.1038/s41467-018-07931-2>
- Eslam, M., Sanyal, A. J., George, J., Sanyal, A., Neuschwander-Tetri, B., Tiribelli, C., Kleiner, D. E., Brunt, E., Bugianesi, E., Yki-Järvinen, H., Grønbaek, H., Cortez-Pinto, H., George, J., Fan, J., Valenti, L., Abdelmalek, M., Romero-Gomez, M., Rinella, M., Arrese, M., ... Younossi, Z. (2020). MAFLD: A Consensus-Driven Proposed Nomenclature for Metabolic Associated Fatty Liver Disease. *Gastroenterology*, 158(7), 1999-2014.e1. <https://doi.org/10.1053/j.gastro.2019.11.312>
- Fang, J., Yu, C.-H., Li, X.-J., Yao, J.-M., Fang, Z.-Y., Yoon, S.-H., & Yu, W.-Y. (2022). Gut dysbiosis in nonalcoholic fatty liver disease: pathogenesis, diagnosis, and therapeutic implications. *Frontiers in Cellular and Infection Microbiology*, 12, 997018. <https://doi.org/10.3389/fcimb.2022.997018>
- Feng, C., Liu, S., Zhang, H., Guan, R., Li, D., Zhou, F., Liang, Y., & Feng, X. (2020). Dimension Reduction and Clustering Models for Single-Cell RNA Sequencing Data: A Comparative Study. *International Journal of Molecular Sciences*, 21(6). <https://doi.org/10.3390/ijms21062181>
- Filliol, A., Saito, Y., Nair, A., Dapito, D. H., Yu, L.-X., Ravichandra, A., Bhattacharjee, S., Affo, S., Fujiwara, N., Su, H., Sun, Q., Savage, T. M., Wilson-Kanamori, J. R., Caviglia, J. M., Chin, L., Chen, D., Wang, X., Caruso, S., Kang, J. K., ... Schwabe, R. F. (2022). Opposing roles of hepatic stellate cell subpopulations in hepatocarcinogenesis. *Nature*, 610(7931), 356–365. <https://doi.org/10.1038/s41586-022-05289-6>
- Fleming, S. J., Chaffin, M. D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J. C., Philippakis, A. A., Ellinor, P. T., & Babadi, M. (2023). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature Methods*, 20(9), 1323–1335. <https://doi.org/10.1038/s41592-023-01943-7>
- Flessa, C.-M., Nasiri-Ansari, N., Kyrou, I., Leca, B. M., Lianou, M., Chatzigeorgiou, A., Kaltsas, G., Kassi, E., & Randeva, H. S. (2022). Genetic and Diet-Induced Animal Models for Non-Alcoholic Fatty Liver Disease (NAFLD) Research. *International Journal of Molecular Sciences*, 23(24), 15791. <https://doi.org/10.3390/ijms232415791>

- Forrow, A., & Schiebinger, G. (2021). LineageOT is a unified framework for lineage tracing and trajectory inference. *Nature Communications*, 12(1), 4940. <https://doi.org/10.1038/s41467-021-25133-1>
- Franzén, O., Gan, L.-M., & Björkegren, J. L. M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019. <https://doi.org/10.1093/database/baz046>
- Fresneda, S., Abbate, M., Busquets-Cortés, C., López-González, A., Fuster-Parra, P., Bennasar-Veny, M., & Yáñez, A. M. (2022). Sex and age differences in the association of fatty liver index-defined non-alcoholic fatty liver disease with cardiometabolic risk factors: a cross-sectional study. *Biology of Sex Differences*, 13(1), 64. <https://doi.org/10.1186/s13293-022-00475-7>
- Gagnon, J., Pi, L., Ryals, M., Wan, Q., Hu, W., Ouyang, Z., Zhang, B., & Li, K. (2022). Recommendations of scRNA-seq Differential Gene Expression Analysis Based on Comprehensive Benchmarking. *Life (Basel, Switzerland)*, 12(6). <https://doi.org/10.3390/life12060850>
- Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., Xu, C., Ashuach, T., Gabitto, M., Lotfollahi, M., ... Yosef, N. (2022). A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2), 163–166. <https://doi.org/10.1038/s41587-021-01206-w>
- Geng, Y., Faber, K. N., de Meijer, V. E., Blokzijl, H., & Moshage, H. (2021). How does hepatic lipid accumulation lead to lipotoxicity in non-alcoholic fatty liver disease? *Hepatology International*, 15(1), 21–35. <https://doi.org/10.1007/s12072-020-10121-2>
- Germain, P.-L., Lun, A., Macnair, W., & Robinson, M. D. (2021). Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research*, 10, 979. <https://doi.org/10.12688/f1000research.73600.1>
- Germain, P.-L., Sonrel, A., & Robinson, M. D. (2020). pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. *Genome Biology*, 21(1), 227. <https://doi.org/10.1186/s13059-020-02136-7>
- Gierlinski, M. (2024). *fenr: Fast functional enrichment for interactive applications*. <https://github.com/bartongroup/fenr>
- Ginès, P., Krag, A., Abalde, J. G., Solà, E., Fabrellas, N., & Kamath, P. S. (2021). Liver cirrhosis. *The Lancet*, 398(10308), 1359–1376. [https://doi.org/10.1016/S0140-6736\(21\)01374-X](https://doi.org/10.1016/S0140-6736(21)01374-X)
- Goessling, W., & Sadler, K. C. (2015). Zebrafish: An Important Tool for Liver Disease Research. *Gastroenterology*, 149(6), 1361–1377. <https://doi.org/10.1053/j.gastro.2015.08.034>
- Gribben, C., Galanakis, V., Calderwood, A., Williams, E. C., Chazarra-Gil, R., Larraz, M., Frau, C., Puengel, T., Guillot, A., Rouhani, F. J., Mahbubani, K., Godfrey, E., Davies, S. E., Athanasiadis, E., Saeb-Parsy, K., Tacke, F., Allison, M., Mohorianu, I., & Vallier, L. (2024). Acquisition of epithelial plasticity in human chronic liver disease. *Nature*, 630(8015), 166–173. <https://doi.org/10.1038/s41586-024-07465-2>

- Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., & Winther, O. (2020). scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16), 4415–4422. <https://doi.org/10.1093/bioinformatics/btaa293>
- Guilliams, M., Bonnardel, J., Haest, B., Vanderborght, B., Wagner, C., Remmerie, A., Bujko, A., Martens, L., Thoné, T., Browaeys, R., De Ponti, F. F., Vanneste, B., Zwicker, C., Svedberg, F. R., Vanhalewyn, T., Gonçalves, A., Lippens, S., Devriendt, B., Cox, E., ... Scott, C. L. (2022). Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell*, 185(2), 379–396.e38. <https://doi.org/10.1016/j.cell.2021.12.018>
- Gunes, A., & Estall, J. L. (2024). Is MASLD lost in translation in mice? *Trends in Endocrinology & Metabolism*. <https://doi.org/10.1016/j.tem.2024.03.005>
- Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSEA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1), 7. <https://doi.org/10.1186/1471-2105-14-7>
- Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1), 75. <https://doi.org/10.1186/s13073-017-0467-4>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Harrison, S. A., Bedossa, P., Guy, C. D., Schattenberg, J. M., Loomba, R., Taub, R., Labriola, D., Moussa, S. E., Neff, G. W., Rinella, M. E., Anstee, Q. M., Abdelmalek, M. F., Younossi, Z., Baum, S. J., Francque, S., Charlton, M. R., Newsome, P. N., Lanthier, N., Schiefke, I., ... Ratziu, V. (2024). A Phase 3, Randomized, Controlled Trial of Resmetirom in NASH with Liver Fibrosis. *New England Journal of Medicine*, 390(6), 497–509. <https://doi.org/10.1056/NEJMoa2309000>
- He, D., Zakeri, M., Sarkar, H., Soneson, C., Srivastava, A., & Patro, R. (2022). Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. *Nature Methods*, 19(3), 316–322. <https://doi.org/10.1038/s41592-022-01408-3>
- He, L., Davila-Velderrain, J., Sumida, T. S., Hafler, D. A., Kellis, M., & Kulminski, A. M. (2021). NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Communications Biology*, 4(1), 629. <https://doi.org/10.1038/s42003-021-02146-6>
- He, L., Lu, A., Qin, L., Zhang, Q., Ling, H., Tan, D., & He, Y. (2021). Application of single-cell RNA sequencing technology in liver diseases: a narrative review. *Annals of Translational Medicine*, 9(20), 1598–1598. <https://doi.org/10.21037/atm-21-4824>
- Hermenean, A., Oatis, D., Herman, H., Ciceu, A., D’Amico, G., & Trotta, M. C. (2022). Galectin 1—A Key Player between Tissue Repair and Fibrosis. *International Journal of Molecular Sciences*, 23(10), 5548. <https://doi.org/10.3390/ijms23105548>
- Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Aliee, H., Ansari, M., Badia-i-Mompel, P., Büttner, M., Dann, E., Dimitrov, D., Dony, L., Frishberg, A., He, D., ... Theis, F. J. (2023). Best

- practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8), 550–572. <https://doi.org/10.1038/s41576-023-00586-w>
- Hie, B. L., Kim, S., Rando, T. A., Bryson, B., & Berger, B. (2024). Scanorama: integrating large and diverse single-cell transcriptomic datasets. *Nature Protocols*. <https://doi.org/10.1038/s41596-024-00991-3>
- Hou, W., & Ji, Z. (2024). Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods*. <https://doi.org/10.1038/s41592-024-02235-4>
- Hou, W., Ji, Z., Chen, Z., Wherry, E. J., Hicks, S. C., & Ji, H. (2023). A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples. *Nature Communications*, 14(1), 7286. <https://doi.org/10.1038/s41467-023-42841-y>
- Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P., & Zhang, Y. (2023). CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, 51(D1), D870–D876. <https://doi.org/10.1093/nar/gkac947>
- Hu, P., Zhang, W., Xin, H., & Deng, G. (2016). Single Cell Isolation and Analysis. *Frontiers in Cell and Developmental Biology*, 4. <https://doi.org/10.3389/fcell.2016.00116>
- Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., & Mohr, S. E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*, 12(1), 357. <https://doi.org/10.1186/1471-2105-12-357>
- Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Ipsen, D. H., Lykkesfeldt, J., & Tveden-Nyborg, P. (2018). Molecular mechanisms of hepatic lipid accumulation in non-alcoholic fatty liver disease. *Cellular and Molecular Life Sciences*, 75(18), 3313–3327. <https://doi.org/10.1007/s00018-018-2860-6>
- Ipsen, D. H., Lykkesfeldt, J., & Tveden-Nyborg, P. (2020). Animal Models of Fibrosis in Nonalcoholic Steatohepatitis: Do They Reflect Human Disease? *Advances in Nutrition*, 11(6), 1696–1711. <https://doi.org/10.1093/advances/nmaa081>
- Jahromi, M. K., Daftari, G., Farhadnejad, H., Tehrani, A. N., Teymoori, F., Salehi-Sahlabadi, A., & Mirmiran, P. (2023). The association of healthy lifestyle score and risk of non-alcoholic fatty liver disease. *BMC Public Health*, 23(1), 973. <https://doi.org/10.1186/s12889-023-15816-3>
- Janssen, P., Kliesmete, Z., Vieth, B., Adiconis, X., Simmons, S., Marshall, J., McCabe, C., Heyn, H., Levin, J. Z., Enard, W., & Hellmann, I. (2023). The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biology*, 24(1), 140. <https://doi.org/10.1186/s13059-023-02978-x>
- Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), 118–127. <https://doi.org/10.1093/biostatistics/kxj037>
- Jonas, W., & Schürmann, A. (2021). Genetic and epigenetic factors determining NAFLD risk. *Molecular Metabolism*, 50, 101111. <https://doi.org/10.1016/j.molmet.2020.101111>
- Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3), e694. <https://doi.org/10.1002/ctm2.694>

- Juanola, O., Martínez-López, S., Francés, R., & Gómez-Hurtado, I. (2021). Non-Alcoholic Fatty Liver Disease: Metabolic, Genetic, Epigenetic and Environmental Risk Factors. *International Journal of Environmental Research and Public Health*, 18(10), 5227. <https://doi.org/10.3390/ijerph18105227>
- Junttila, S., Smolander, J., & Elo, L. L. (2022). Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. *Briefings in Bioinformatics*, 23(5). <https://doi.org/10.1093/bib/bbac286>
- Kaminow, B., Yunusov, D., & Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *BioRxiv*, 2021.05.05.442755. <https://doi.org/10.1101/2021.05.05.442755>
- Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11), 1947–1951. <https://doi.org/10.1002/pro.3715>
- Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1), D587–D592. <https://doi.org/10.1093/nar/gkac963>
- Kassambara, A. (2023). *ggpubr: 'ggplot2' Based Publication Ready Plots*. <https://cran.r-project.org/web/packages/ggpubr/index.html>
- 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*, 74(2), 987–1002. <https://doi.org/10.1002/hep.31759>
- Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1), 5416. <https://doi.org/10.1038/s41467-019-13056-x>
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4), 610–620. <https://doi.org/10.1016/j.molcel.2015.04.005>
- Konyn, P., Ahmed, A., & Kim, D. (2023). Causes and risk profiles of mortality among individuals with nonalcoholic fatty liver disease. *Clinical and Molecular Hepatology*, 29(Suppl), S43–S57. <https://doi.org/10.3350/cmh.2022.0351>
- Koutoukidis, D. A., Astbury, N. M., Tudor, K. E., Morris, E., Henry, J. A., Noreik, M., Jebb, S. A., & Aveyard, P. (2019). Association of Weight Loss Interventions With Changes in Biomarkers of Nonalcoholic Fatty Liver Disease. *JAMA Internal Medicine*, 179(9), 1262. <https://doi.org/10.1001/jamainternmed.2019.2248>
- Kouvari, M., Valenzuela-Vallejo, L., Guatibonza-Garcia, V., Polyzos, S. A., Deng, Y., Kokkorakis, M., Agraz, M., Mylonakis, S. C., Katsarou, A., Verrastro, O., Markakis, G., Eslam, M., Papatheodoridis, G., George, J., Mingrone, G., & Mantzoros, C. S. (2023). Liver biopsy-based validation, confirmation and comparison of the diagnostic performance of established and novel non-invasive steatotic liver disease indexes: Results from a large multi-center study. *Metabolism*, 147, 155666. <https://doi.org/10.1016/j.metabol.2023.155666>
- Kulkarni, A., Ibrahim, S., Haider, I., Basha, A., Montgomery, E., Ermis, E., Mirmira, R. G., & Anderson, R. M. (2022). A Novel 2-Hit Zebrafish Model to Study Early Pathogenesis

- of Non-Alcoholic Fatty Liver Disease. *Biomedicines*, 10(2), 479.
<https://doi.org/10.3390/biomedicines10020479>
- Kumar, S., Duan, Q., Wu, R., Harris, E. N., & Su, Q. (2021). Pathophysiological communication between hepatocytes and non-parenchymal cells in liver injury from NAFLD to liver fibrosis. *Advanced Drug Delivery Reviews*, 176, 113869.
<https://doi.org/10.1016/j.addr.2021.113869>
- La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastrioti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., ... Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719), 494–498.
<https://doi.org/10.1038/s41586-018-0414-6>
- Lafoz, E., Ruart, M., Anton, A., Oncins, A., & Hernández-Gea, V. (2020). The Endothelium as a Driver of Liver Fibrosis and Regeneration. *Cells*, 9(4), 929.
<https://doi.org/10.3390/cells9040929>
- Lause, J., Berens, P., & Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, 22(1), 258.
<https://doi.org/10.1186/s13059-021-02451-7>
- Le, M. H., Yeo, Y. H., Li, X., Li, J., Zou, B., Wu, Y., Ye, Q., Huang, D. Q., Zhao, C., Zhang, J., Liu, C., Chang, N., Xing, F., Yan, S., Wan, Z. H., Tang, N. S. Y., Mayumi, M., Liu, X., Liu, C., ... Nguyen, M. H. (2022). 2019 Global NAFLD Prevalence: A Systematic Review and Meta-analysis. *Clinical Gastroenterology and Hepatology*, 20(12), 2809–2817.e28. <https://doi.org/10.1016/j.cgh.2021.12.002>
- Lee, Y., Cho, E. J., Choe, E. K., Kwak, M.-S., Yang, J. I., Oh, S.-W., Yim, J. Y., & Chung, G. E. (2024). Genome-wide association study of metabolic dysfunction-associated fatty liver disease in a Korean population. *Scientific Reports*, 14(1), 9753.
<https://doi.org/10.1038/s41598-024-60152-0>
- Lekakis, V., & Papatheodoridis, G. V. (2024). Natural history of metabolic dysfunction-associated steatotic liver disease. *European Journal of Internal Medicine*, 122, 3–10.
<https://doi.org/10.1016/j.ejim.2023.11.005>
- Li, S., Zhang, P., Chen, W., Ye, L., Brannan, K. W., Le, N.-T., Abe, J., Cooke, J. P., & Wang, G. (2024). A relay velocity model infers cell-dependent RNA velocity. *Nature Biotechnology*, 42(1), 99–108. <https://doi.org/10.1038/s41587-023-01728-5>
- Li, W., Nakano, H., Fan, W., Li, Y., Sil, P., Nakano, K., Zhao, F., Karmaus, P. W., Grimm, S. A., Shi, M., Xu, X., Mizuta, R., Kitamura, D., Wan, Y., Fessler, M. B., Cook, D. N., Shats, I., Li, X., & Li, L. (2023). DNASE1L3 enhances antitumor immunity and suppresses tumor progression in colon cancer. *JCI Insight*, 8(17).
<https://doi.org/10.1172/jci.insight.168161>
- Li, X., Zhou, L., Zheng, Y., He, T., Guo, H., Li, J., & Zhang, J. (2023). Establishment of a non-alcoholic fatty liver disease model by high fat diet in adult zebrafish. *Animal Models and Experimental Medicine*. <https://doi.org/10.1002/ame2.12309>
- Li, Y., Yang, P., Ye, J., Xu, Q., Wu, J., & Wang, Y. (2024). Updated mechanisms of MASLD pathogenesis. *Lipids in Health and Disease*, 23(1), 117.
<https://doi.org/10.1186/s12944-024-02108-x>

- Li, Z., Zhang, H., Li, Q., Feng, W., Jia, X., Zhou, R., Huang, Y., Li, Y., Hu, Z., Hu, X., Zhu, X., & Huang, S. (2023). GepLiver: an integrative liver expression atlas spanning developmental stages and liver disease phases. *Scientific Data*, 10(1), 376. <https://doi.org/10.1038/s41597-023-02257-1>
- Linderman, G. C. (2021). Dimensionality Reduction of Single-Cell RNA-Seq Data. In E. Picardi (Ed.), *RNA Bioinformatics* (pp. 331–342). Springer US. https://doi.org/10.1007/978-1-0716-1307-8_18
- Lit, K. K., Zhirenova, Z., & Blocki, A. (2024). Insulin-like growth factor-binding protein 7 (IGFBP7): A microenvironment-dependent regulator of angiogenesis and vascular remodeling. *Frontiers in Cell and Developmental Biology*, 12. <https://doi.org/10.3389/fcell.2024.1421438>
- Liu, B., Xiang, L., Ji, J., Liu, W., Chen, Y., Xia, M., Liu, Y., Liu, W., Zhu, P., Jin, Y., Han, Y., Lu, J., Li, X., Zheng, M., & Lu, Y. (2021). Sparcl1 promotes nonalcoholic steatohepatitis progression in mice through upregulation of CCL2. *Journal of Clinical Investigation*, 131(20). <https://doi.org/10.1172/JCI144801>
- Liu, J., Hu, S., Chen, L., Daly, C., Prada Medina, C. A., Richardson, T. G., Traylor, M., Dempster, N. J., Mbasu, R., Monfeuga, T., Vujkovic, M., Tsao, P. S., Lynch, J. A., Voight, B. F., Chang, K.-M., Million, V. A., Cobbald, J. F., Tomlinson, J. W., van Duijn, C. M., & Howson, J. M. M. (2023). Profiling the genome and proteome of metabolic dysfunction-associated steatotic liver disease identifies potential therapeutic targets. *MedRxiv : The Preprint Server for Health Sciences*. <https://doi.org/10.1101/2023.11.30.23299247>
- Liu, X., Powell, C. A., & Wang, X. (2022). Forward single-cell sequencing into clinical application: Understanding of cancer microenvironment at single-cell solution. *Clinical and Translational Medicine*, 12(4). <https://doi.org/10.1002/ctm2.782>
- Liu, Y., Wang, T., Zhou, B., & Zheng, D. (2021). Robust integration of multiple single-cell RNA sequencing datasets using a single reference space. *Nature Biotechnology*, 39(7), 877–884. <https://doi.org/10.1038/s41587-021-00859-x>
- Loomba, R., Friedman, S. L., & Shulman, G. I. (2021). Mechanisms and disease consequences of nonalcoholic fatty liver disease. *Cell*, 184(10), 2537–2564. <https://doi.org/10.1016/j.cell.2021.04.015>
- Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 1053–1058. <https://doi.org/10.1038/s41592-018-0229-2>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Lu, Y., Yang, A., Quan, C., Pan, Y., Zhang, H., Li, Y., Gao, C., Lu, H., Wang, X., Cao, P., Chen, H., Lu, S., & Zhou, G. (2022). A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nature Communications*, 13(1), 4594. <https://doi.org/10.1038/s41467-022-32283-3>
- Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F. J. (2022).

- Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1), 41–50. <https://doi.org/10.1038/s41592-021-01336-8>
- Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6). <https://doi.org/10.15252/msb.20188746>
- Ma, W., Huang, J., Cai, B., Shao, M., Yu, X., Kjær, M. B., Lv, M., Zhong, X., Xu, S., Zhan, B., Li, Q., Huang, Q., Ma, M., Cheng, L., Luo, Y., Grønbaek, H., Zhou, X., & Lin, L. (2023). A novel gene-screening approach reveals QSOX1/IL1RAP as promising biomarkers for the severity of non-alcoholic fatty liver disease. *MedRxiv*. <https://doi.org/10.1101/2023.07.26.23293038>
- Ma, W., Wu, W., Wen, W., Xu, F., Han, D., Lyu, J., & Huang, Y. (2022). Association of NAFLD with cardiovascular disease and all-cause mortality: a large-scale prospective cohort study based on UK Biobank. *Therapeutic Advances in Chronic Disease*, 13, 204062232211224. <https://doi.org/10.1177/20406223221122478>
- Ma, W., Zhong, X., Cai, B., Shao, M., Yu, X., Lv, M., Xu, S., Zhan, B., Li, Q., Ma, M., Kjær, M. B., Huang, J., Luo, Y., Grønbaek, H., & Lin, L. (2023). Unveiling the pathogenesis of non-alcoholic fatty liver disease by decoding biomarkers through integrated single-cell and single-nucleus profiles. *MedRxiv*, 2023.10.05.23296635. <https://doi.org/10.1101/2023.10.05.23296635>
- MacParland, S. A., Liu, J. C., Ma, X.-Z., Innes, B. T., Bartczak, A. M., Gage, B. K., Manuel, J., Khuu, N., Echeverri, J., Linares, I., Gupta, R., Cheng, M. L., Liu, L. Y., Camat, D., Chung, S. W., Seliga, R. K., Shao, Z., Lee, E., Ogawa, S., ... McGilvray, I. D. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature Communications*, 9(1), 4383. <https://doi.org/10.1038/s41467-018-06318-7>
- Mancarci, O., & French, L. (2022). *homologene: Quick Access to Homologene and Gene Annotation Updates*. <https://github.com/oganm/homologene>
- Mantovani, A., Petracca, G., Beatrice, G., Csermely, A., Tilg, H., Byrne, C. D., & Targher, G. (2022). Non-alcoholic fatty liver disease and increased risk of incident extrahepatic cancers: a meta-analysis of observational cohort studies. *Gut*, 71(4), 778–788. <https://doi.org/10.1136/gutjnl-2021-324191>
- Mantovani, A., Scorletti, E., Mosca, A., Alisi, A., Byrne, C. D., & Targher, G. (2020). Complications, morbidity and mortality of nonalcoholic fatty liver disease. *Metabolism*, 111, 154170. <https://doi.org/10.1016/j.metabol.2020.154170>
- Mascha, E. J., & Vetter, T. R. (2018). Significance, Errors, Power, and Sample Size: The Blocking and Tackling of Statistics. *Anesthesia & Analgesia*, 126(2), 691–698. <https://doi.org/10.1213/ANE.0000000000002741>
- McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, 8(4), 329–337.e4. <https://doi.org/10.1016/j.cels.2019.03.003>
- McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.
- McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. <https://arxiv.org/abs/1802.03426>

- McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61.
<https://doi.org/10.25080/Majora-92bf1922-00a>
- McTeer, M., Applegate, D., Mesenbrink, P., Ratziu, V., Schattenberg, J. M., Bugianesi, E., Geier, A., Romero Gomez, M., Dufour, J.-F., Ekstedt, M., Francque, S., Yki-Jarvinen, H., Allison, M., Valenti, L., Miele, L., Pavlides, M., Cobbald, J., Papatheodoridis, G., Holleboom, A. G., ... Missier, P. (2024). Machine learning approaches to enhance diagnosis and staging of patients with MASLD using routinely available clinical information. *PLOS ONE*, 19(2), e0299487.
<https://doi.org/10.1371/journal.pone.0299487>
- Megill, C., Martin, B., Weaver, C., Bell, S., Prins, L., Badajoz, S., McCandless, B., Pisco, A. O., Kinsella, M., Griffin, F., Kiggins, J., Haliburton, G., Mani, A., Weiden, M., Dunitz, M., Lombardo, M., Huang, T., Smith, T., Chambers, S., ... Carr, A. (2021). cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *BioRxiv*.
<https://doi.org/10.1101/2021.04.05.438318>
- Melsted, P., Booesbaghi, A. S., Liu, L., Gao, F., Lu, L., Min, K. H., da Veiga Beltrame, E., Hjörleifsson, K. E., Gehring, J., & Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, 39(7), 813–818.
<https://doi.org/10.1038/s41587-021-00870-2>
- Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., Petryszak, R., Ragueneau, E., Rothfels, K., Sevilla, C., Shamovsky, V., Stephan, R., Tiwari, K., Varusai, T., Weiser, J., ... D'Eustachio, P. (2024). The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, 52(D1), D672–D678. <https://doi.org/10.1093/nar/gkad1025>
- Montero-Vallejo, R., Maya-Miles, D., Ampuero, J., Martín, F., Romero-Gómez, M., & Gallego-Durán, R. (2024). Novel insights into metabolic-associated steatotic liver disease preclinical models. *Liver International*, 44(3), 644–662.
<https://doi.org/10.1111/liv.15830>
- Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. van den, Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12), 1482–1492.
<https://doi.org/10.1038/s41587-019-0336-3>
- Moreno, J., Gluud, L. L., Galsgaard, E. D., Hvid, H., Mazzoni, G., & Das, V. (2024). Identification of ligand and receptor interactions in CKD and MASH through the integration of single cell and spatial transcriptomics. *PLOS ONE*, 19(5), e0302853.
<https://doi.org/10.1371/journal.pone.0302853>
- Moriyama, M., Kanda, T., Midorikawa, Y., Matsumura, H., Masuzaki, R., Nakamura, H., Ogawa, M., Matsuoka, S., Shibata, T., Yamazaki, M., Kuroda, K., Nakayama, H., Higaki, T., Kanemaru, K., Miki, T., Sugitani, M., & Takayama, T. (2022). The proliferation of atypical hepatocytes and CDT1 expression in noncancerous tissue are associated with the postoperative recurrence of hepatocellular carcinoma. *Scientific Reports*, 12(1), 20508. <https://doi.org/10.1038/s41598-022-25201-6>
- Morrison, J. K., DeRossi, C., Alter, I. L., Nayar, S., Giri, M., Zhang, C., Cho, J. H., & Chu, J. (2022). Single-cell transcriptomics reveals conserved cell identities and fibrogenic

- phenotypes in zebrafish and human liver. *Hepatology Communications*, 6(7), 1711–1724. <https://doi.org/10.1002/hep4.1930>
- Mou, T., Deng, W., Gu, F., Pawitan, Y., & Vu, T. N. (2019). Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. *Frontiers in Genetics*, 10, 1331. <https://doi.org/10.3389/fgene.2019.01331>
- Müller-Dott, S., Tsirvouli, E., Vazquez, M., Ramirez Flores, R. O., Badia-i-Mompel, P., Fallegger, R., Türei, D., Lægreid, A., & Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20), 10934–10949. <https://doi.org/10.1093/nar/gkad841>
- Murphy, A. E., & Skene, N. G. (2022). A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nature Communications*, 13(1), 7851. <https://doi.org/10.1038/s41467-022-35519-4>
- Muzellec, B., Teleńczuk, M., Cabeli, V., & Andreux, M. (2023). PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics*, 39(9). <https://doi.org/10.1093/bioinformatics/btad547>
- Neavin, D., Senabouth, A., Arora, H., Lee, J. T. H., Ripoll-Cladellas, A., Franke, L., Prabhakar, S., Ye, C. J., McCarthy, D. J., Melé, M., Hemberg, M., & Powell, J. E. (2024). Demuxafy: improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *Genome Biology*, 25(1), 94. <https://doi.org/10.1186/s13059-024-03224-8>
- Nevzorova, Y. A., Boyer-Diaz, Z., Cubero, F. J., & Gracia-Sancho, J. (2020). Animal models for liver disease – A practical approach for translational research. *Journal of Hepatology*, 73(2), 423–440. <https://doi.org/10.1016/j.jhep.2020.04.011>
- Nguyen, H., Tran, D., Tran, B., Pehlivan, B., & Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. In *Briefings in Bioinformatics* (Vol. 22, Issue 3). Oxford University Press. <https://doi.org/10.1093/bib/bbaa190>
- Nichol, D., & Stuhlmann, H. (2012). EGFL7: a unique angiogenic signaling factor in vascular development and disease. *Blood*, 119(6), 1345–1352. <https://doi.org/10.1182/blood-2011-10-322446>
- Noureddin, M., Vipani, A., Bresee, C., Todo, T., Kim, I. K., Alkhouri, N., Setiawan, V. W., Tran, T., Ayoub, W. S., Lu, S. C., Klein, A. S., Sundaram, V., & Nissen, N. N. (2018). NASH Leading Cause of Liver Transplant in Women: Updated Analysis of Indications For Liver Transplant and Ethnic and Gender Variances. *American Journal of Gastroenterology*, 113(11), 1649–1659. <https://doi.org/10.1038/s41395-018-0088-6>
- Ouyang, J. F., Kamaraj, U. S., Cao, E. Y., & Rackham, O. J. L. (2021). ShinyCell: simple and sharable visualization of single-cell gene expression data. *Bioinformatics*, 37(19), 3374–3376. <https://doi.org/10.1093/bioinformatics/btab209>
- Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, 19, 961–969. <https://doi.org/10.1016/j.csbj.2021.01.015>
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). *Automatic differentiation in pytorch*.

- Perrier, V., Meyer, F., & Granjon, D. (2023). *shinyWidgets: Custom Inputs Widgets for Shiny*. <https://cran.r-project.org/web/packages/shinyWidgets/index.html>
- Petrenko, O., Königshofer, P., Brusilovskaya, K., Hofer, B. S., Bareiner, K., Simbrunner, B., Jühling, F., Baumert, T. F., Lupberger, J., Trauner, M., Kauschke, S. G., Pfisterer, L., Simon, E., Rendeiro, A. F., de Rooij, L. P. M. H., Schwabl, P., & Reiberger, T. (2024). Transcriptomic signatures of progressive and regressive liver fibrosis and portal hypertension. *IScience*, 27(3), 109301. <https://doi.org/10.1016/j.isci.2024.109301>
- Peyvandipour, A., Shafi, A., Saberian, N., & Draghici, S. (2020). Identification of cell types from single cell data using stable clustering. *Scientific Reports*, 10(1), 12349. <https://doi.org/10.1038/s41598-020-66848-3>
- Pfister, D., Núñez, N. G., Pinyol, R., Govaere, O., Pinter, M., Szydlowska, M., Gupta, R., Qiu, M., Deczkowska, A., Weiner, A., Müller, F., Sinha, A., Friebe, E., Engleitner, T., Lenggenhager, D., Moncsek, A., Heide, D., Stirm, K., Kosla, J., ... Heikenwalder, M. (2021). NASH limits anti-tumour surveillance in immunotherapy-treated HCC. *Nature*, 592(7854), 450–456. <https://doi.org/10.1038/s41586-021-03362-0>
- Pischke, S., Shiprov, A., Peters, U., Schulze zur Wiesch, J., Kluwe, J., Westphal, T., Fischer, F., Mader, M., Fründt, T., Horvatits, K., Horvatits, T., Aarabi, G., & Beikler, T. (2023). High prevalence of periodontal disease in patients with NASH- possible association of poor dental health with NASH severity. *Annals of Hepatology*, 28(2), 100887. <https://doi.org/10.1016/j.aohep.2022.100887>
- Poch, T., Krause, J., Casar, C., Liwinski, T., Glau, L., Kaufmann, M., Ahrenstorf, A. E., Hess, L. U., Ziegler, A. E., Martrus, G., Lunemann, S., Sebode, M., Li, J., Schwinge, D., Krebs, C. F., Franke, A., Friese, M. A., Oldhafer, K. J., Fischer, L., ... Schramm, C. (2021). Single-cell atlas of hepatic T cells reveals expansion of liver-resident naive-like CD4⁺ T cells in primary sclerosing cholangitis. *Journal of Hepatology*, 75(2), 414–423. <https://doi.org/10.1016/j.jhep.2021.03.016>
- R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/>
- Ramachandran, P., Dobie, R., Wilson-Kanamori, J. R., Dora, E. F., Henderson, B. E. P., Luu, N. T., Portman, J. R., Matchett, K. P., Brice, M., Marwick, J. A., Taylor, R. S., Efremova, M., Vento-Tormo, R., Carragher, N. O., Kendall, T. J., Fallowfield, J. A., Harrison, E. M., Mole, D. J., Wigmore, S. J., ... Henderson, N. C. (2019). Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature*, 575(7783), 512–518. <https://doi.org/10.1038/s41586-019-1631-3>
- Rao, G., Peng, X., Li, X., An, K., He, H., Fu, X., Li, S., & An, Z. (2023). Unmasking the enigma of lipid metabolism in metabolic dysfunction-associated steatotic liver disease: from mechanism to the clinic. *Frontiers in Medicine*, 10. <https://doi.org/10.3389/fmed.2023.1294267>
- Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., ... Yosef, N. (2017). The Human Cell Atlas. *ELife*, 6. <https://doi.org/10.7554/eLife.27041>
- Remmerie, A., Martens, L., Thoné, T., Castoldi, A., Seurinck, R., Pavie, B., Roels, J., Vanneste, B., De Prijck, S., Vanhockerhout, M., Binte Abdul Latib, M., Devisscher, L.,

- Hoorens, A., Bonnardel, J., Vandamme, N., Kremer, A., Borghgraef, P., Van Vlierberghe, H., Lippens, S., ... Scott, C. L. (2020). Osteopontin Expression Identifies a Subset of Recruited Macrophages Distinct from Kupffer Cells in the Fatty Liver. *Immunity*, 53(3), 641–657.e14. <https://doi.org/10.1016/j.immuni.2020.08.004>
- Rinella, M. E., Lazarus, J. V., Ratzliff, V., Francque, S. M., Sanyal, A. J., Kanwal, F., Romero, D., Abdelmalek, M. F., Anstee, Q. M., Arab, J. P., Arrese, M., Bataller, R., Beuers, U., Boursier, J., Bugianesi, E., Byrne, C. D., Castro Narro, G. E., Chowdhury, A., Cortez-Pinto, H., ... Newsome, P. N. (2023). A multisociety Delphi consensus statement on new fatty liver disease nomenclature. *Journal of Hepatology*, 79(6), 1542–1556. <https://doi.org/10.1016/j.jhep.2023.06.003>
- Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., & Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1), 284. <https://doi.org/10.1038/s41467-017-02554-5>
- 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–e47. <https://doi.org/10.1093/nar/gkv007>
- 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. <https://doi.org/10.1093/bioinformatics/btp616>
- Rockey, D. C., Caldwell, S. H., Goodman, Z. D., Nelson, R. C., & Smith, A. D. (2009). Liver biopsy. *Hepatology*, 49(3), 1017–1044. <https://doi.org/10.1002/hep.22742>
- Rocque, B., Barbetta, A., Singh, P., Goldbeck, C., Helou, D. G., Loh, Y.-H. E., Ung, N., Lee, J., Akbari, O., & Emamaullee, J. (2021). Creation of a Single Cell RNASeq Meta-Atlas to Define Human Liver Immune Homeostasis. *Frontiers in Immunology*, 12. <https://doi.org/10.3389/fimmu.2021.679521>
- Ryu, Y., Han, G. H., Jung, E., & Hwang, D. (2023). Integration of Single-Cell RNA-Seq Datasets: A Review of Computational Methods. *Molecules and Cells*, 46(2), 106–119. <https://doi.org/10.14348/molcells.2023.0009>
- Sanyal, A. J., Shankar, S. S., Yates, K. P., Bolognese, J., Daly, E., Dehn, C. A., Neuschwander-Tetri, B., Kowdley, K., Vuppalanchi, R., Behling, C., Tonascia, J., Samir, A., Sirlin, C., Sherlock, S. P., Fowler, K., Heymann, H., Kamphaus, T. N., Loomba, R., & Calle, R. A. (2023). Diagnostic performance of circulating biomarkers for non-alcoholic steatohepatitis. *Nature Medicine*, 29(10), 2656–2664. <https://doi.org/10.1038/s41591-023-02539-6>
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5), 495–502. <https://doi.org/10.1038/nbt.3192>
- Schiebout, C., Lust, H., Huang, Y., & Frost, H. R. (2023). Cell type-specific interaction analysis using doublets in scRNA-seq. *Bioinformatics Advances*, 3(1). <https://doi.org/10.1093/bioadv/vbad120>
- Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., ... Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid

- genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5), 849–864. <https://doi.org/10.1101/gr.213611.116>
- Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M. J., Blüthgen, N., & Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications*, 9(1), 20. <https://doi.org/10.1038/s41467-017-02391-6>
- Seal, R. L., Braschi, B., Gray, K., Jones, T. E. M., Tweedie, S., Haim-Vilmovsky, L., & Bruford, E. A. (2023). Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research*, 51(D1), D1003–D1009. <https://doi.org/10.1093/nar/gkac888>
- Selman, M., & Pardo, A. (2021). Fibroageing: An ageing pathological feature driven by dysregulated extracellular matrix-cell mechanobiology. *Ageing Research Reviews*, 70, 101393. <https://doi.org/10.1016/j.arr.2021.101393>
- Sharma, B., & John, S. (2018). *Hepatic cirrhosis*.
- Sheka, A. C., Adeyi, O., Thompson, J., Hameed, B., Crawford, P. A., & Ikramuddin, S. (2020). Nonalcoholic Steatohepatitis. *JAMA*, 323(12), 1175. <https://doi.org/10.1001/jama.2020.2298>
- Sheng, C., Lopes, R., Li, G., Schuierer, S., Waldt, A., Cuttat, R., Dimitrieva, S., Kauffmann, A., Durand, E., Galli, G. G., Roma, G., & de Weck, A. (2022). Probabilistic machine learning ensures accurate ambient denoising in droplet-based single-cell omics. *BioRxiv*, 2022.01.14.476312. <https://doi.org/10.1101/2022.01.14.476312>
- Sievert, C., Cheng, J., & Aden-Buie, G. (2024). *bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'*. <https://rstudio.github.io/bslib/>
- Singh, M. K., Yadav, R., Bhaskar, A. K., Sengupta, S., & Sachidanandan, C. (2023). A diet-independent zebrafish model for NAFLD recapitulates patient lipid profiles and offers a system for small molecule screening. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1868(1), 159246. <https://doi.org/10.1016/j.bbalip.2022.159246>
- Singh, R., Wu, A. P., Mudide, A., & Berger, B. (2024). Causal gene regulatory analysis with RNA velocity reveals an interplay between slow and fast transcription factors. *Cell Systems*, 15(5), 462-474.e5. <https://doi.org/10.1016/j.cels.2024.04.005>
- Soneson, C., Srivastava, A., Patro, R., & Stadler, M. B. (2021). Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLOS Computational Biology*, 17(1), e1008585. <https://doi.org/10.1371/journal.pcbi.1008585>
- Song, D., Xi, N. M., Li, J. J., & Wang, L. (2022). scSampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. *Bioinformatics*, 38(11), 3126–3127. <https://doi.org/10.1093/bioinformatics/btac271>
- Song, Y., Miao, Z., Brazma, A., & Papatheodorou, I. (2023). Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nature Communications*, 14(1), 6495. <https://doi.org/10.1038/s41467-023-41855-w>
- Speir, M. L., Bhaduri, A., Markov, N. S., Moreno, P., Nowakowski, T. J., Papatheodorou, I., Pollen, A. A., Raney, B. J., Seninge, L., Kent, W. J., & Haeussler, M. (2021). UCSC Cell Browser: visualize your single-cell data. *Bioinformatics*, 37(23), 4578–4580. <https://doi.org/10.1093/bioinformatics/btab503>

- Speliotes, E. K., Yerges-Armstrong, L. M., Wu, J., Hernaez, R., Kim, L. J., Palmer, C. D., Gudnason, V., Eiriksdottir, G., Garcia, M. E., Launer, L. J., Nalls, M. A., Clark, J. M., Mitchell, B. D., Shuldiner, A. R., Butler, J. L., Tomas, M., Hoffmann, U., Hwang, S.-J., Massaro, J. M., ... Borecki, I. B. (2011). Genome-Wide Association Analysis Identifies Variants Associated with Nonalcoholic Fatty Liver Disease That Have Distinct Effects on Metabolic Traits. *PLoS Genetics*, 7(3), e1001324. <https://doi.org/10.1371/journal.pgen.1001324>
- Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1), 5692. <https://doi.org/10.1038/s41467-021-25960-2>
- Srivastava, A., Malik, L., Smith, T., Sudbery, I., & Patro, R. (2019). Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biology*, 20(1), 65. <https://doi.org/10.1186/s13059-019-1670-y>
- Stähli, P., Becchetti, C., Korta Martiartu, N., Berzigotti, A., Frenz, M., & Jaeger, M. (2023). First-in-human diagnostic study of hepatic steatosis with computed ultrasound tomography in echo mode. *Communications Medicine*, 3(1), 176. <https://doi.org/10.1038/s43856-023-00409-3>
- Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11), 631–656. <https://doi.org/10.1038/s41576-019-0150-2>
- Stoyanov, E., Ludwig, G., Mizrahi, L., Olam, D., Schnitzer-Perlman, T., Tasika, E., Sass, G., Tiegs, G., Jiang, Y., Nie, T., Kohler, J., Schinazi, R. F., Vertino, P. M., Cedar, H., Galun, E., & Goldenberg, D. (2015). Chronic liver inflammation modifies DNA methylation at the precancerous stage of murine hepatocarcinogenesis. *Oncotarget*, 6(13), 11047–11060. <https://doi.org/10.18632/oncotarget.3567>
- Su, K., Yu, T., & Wu, H. (2021). Accurate feature selection improves single-cell RNA-seq cell clustering. *Briefings in Bioinformatics*, 22(5). <https://doi.org/10.1093/bib/bbab034>
- Subramanian, A., Alperovich, M., Yang, Y., & Li, B. (2022). Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biology*, 23(1), 267. <https://doi.org/10.1186/s13059-022-02820-w>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & 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*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Sun, G., Jin, H., Zhang, C., Meng, H., Zhao, X., Wei, D., Ou, X., Wang, Q., Li, S., Wang, T., Sun, X., Shi, W., Tian, D., Liu, K., Xu, H., Tian, Y., Li, X., Guo, W., Jia, J., ... Zhang, D. (2018). OX40 Regulates Both Innate and Adaptive Immunity and Promotes Nonalcoholic Steatohepatitis. *Cell Reports*, 25(13), 3786-3799.e4. <https://doi.org/10.1016/j.celrep.2018.12.006>
- Suvà, M. L., & Tirosh, I. (2019). Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Molecular Cell*, 75(1), 7–12. <https://doi.org/10.1016/j.molcel.2019.05.003>

- Sveinbjornsson, G., Ulfarsson, M. O., Thorolfsdottir, R. B., Jonsson, B. A., Einarsson, E., Gunnlaugsson, G., Rognvaldsson, S., Arnar, D. O., Baldvinsson, M., Bjarnason, R. G., Eirisdottir, T., Erikstrup, C., Ferkingstad, E., Halldorsson, G. H., Helgason, H., Helgadottir, A., Hindhede, L., Hjorleifsson, G., Jones, D., ... Stefansson, K. (2022). Multiomics study of nonalcoholic fatty liver disease. *Nature Genetics*, 54(11), 1652–1663. <https://doi.org/10.1038/s41588-022-01199-5>
- Tamburini, B. A. J., Finlon, J. M., Gillen, A. E., Kriss, M. S., Riemondy, K. A., Fu, R., Schuyler, R. P., Hesselberth, J. R., Rosen, H. R., & Burchill, M. A. (2019). Chronic Liver Disease in Humans Causes Expansion and Differentiation of Liver Lymphatic Endothelial Cells. *Frontiers in Immunology*, 10, 1036. <https://doi.org/10.3389/fimmu.2019.01036>
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5), 377–382. <https://doi.org/10.1038/nmeth.1315>
- Tapper, E. B., & Lok, A. S.-F. (2017). Use of Liver Imaging and Biopsy in Clinical Practice. *New England Journal of Medicine*, 377(8), 756–768. <https://doi.org/10.1056/NEJMra1610570>
- Tapper, E. B., & Parikh, N. D. (2023). Diagnosis and Management of Cirrhosis and Its Complications. *JAMA*, 329(18), 1589. <https://doi.org/10.1001/jama.2023.5997>
- Tarhan, L., Bistline, J., Chang, J., Galloway, B., Hanna, E., & Weitz, E. (2023). Single Cell Portal: an interactive home for single-cell genomics data. *BioRxiv*. <https://doi.org/10.1101/2023.07.13.548886>
- Terkelsen, M. K., Bendixen, S. M., Hansen, D., Scott, E. A. H., Moeller, A. F., Nielsen, R., Mandrup, S., Schlosser, A., Andersen, T. L., Sorensen, G. L., Krag, A., Natarajan, K. N., Detlefsen, S., Dimke, H., & Ravnkjaer, K. (2020). Transcriptional Dynamics of Hepatic Sinusoid-Associated Cells After Liver Injury. *Hepatology*, 72(6), 2119–2133. <https://doi.org/10.1002/hep.31215>
- The Genome Reference Consortium. (2024, July 8). <https://www.ncbi.nlm.nih.gov/grc>
- Townes, F. W., & Irizarry, R. A. (2020). Quantile normalization of single-cell RNA-seq read counts without unique molecular identifiers. *Genome Biology*, 21(1), 160. <https://doi.org/10.1186/s13059-020-02078-0>
- Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), 5233. <https://doi.org/10.1038/s41598-019-41695-z>
- Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21(1), 12. <https://doi.org/10.1186/s13059-019-1850-9>
- Trépo, E., & Valenti, L. (2020). Update on NAFLD genetics: From new variants to the clinic. *Journal of Hepatology*, 72(6), 1196–1209. <https://doi.org/10.1016/j.jhep.2020.02.020>
- U.S. Food and Drug Administration. (2024, March 14). *FDA Approves First Treatment for Patients with Liver Scarring Due to Fatty Liver Disease*. <https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-patients-liver-scarring-due-fatty-liver-disease>

- Ushey, K., Allaire, J. J., & Tang, Y. (2024). *reticulate: Interface to 'Python'*.
<https://rstudio.github.io/reticulate/>
- Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., & Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6), 565–571. <https://doi.org/10.1038/nmeth.4292>
- Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., & Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, 11(1).
<https://doi.org/10.1038/s41467-020-14766-3>
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Wallace S, Dobie R, Henderson NC, & Wilson-Kanamori JR. (2022, February 1). *Gastrointestinal B-cells license metabolic T-cell activation in NASH microbiota anigen-independently and contribute to fibrosis by IgA-FcRy signalling*.
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE190487>.
- Wallace, S. J., Tacke, F., Schwabe, R. F., & Henderson, N. C. (2022). Understanding the cellular interactome of non-alcoholic fatty liver disease. *JHEP Reports*, 4(8), 100524.
<https://doi.org/10.1016/j.jhepr.2022.100524>
- Wang, H., Shen, H., Seo, W., & Hwang, S. (2023). Experimental models of fatty liver diseases: Status and appraisal. *Hepatology Communications*, 7(7).
<https://doi.org/10.1097/HC9.0000000000000200>
- Wang, P.-W., Wu, T.-H., Lin, T.-Y., Chen, M.-H., Yeh, C.-T., & Pan, T.-L. (2019). Characterization of the Roles of Vimentin in Regulating the Proliferation and Migration of HSCs during Hepatic Fibrogenesis. *Cells*, 8(10), 1184.
<https://doi.org/10.3390/cells8101184>
- Wang, S., & Friedman, S. L. (2023). Found in translation—Fibrosis in metabolic dysfunction–associated steatohepatitis (MASH). *Science Translational Medicine*, 15(716). <https://doi.org/10.1126/scitranslmed.adi0759>
- Wang, S., Karikomi, M., MacLean, A. L., & Nie, Q. (2019). Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Research*, 47(11), e66–e66. <https://doi.org/10.1093/nar/gkz204>
- Wang, S., Li, K., Pickholz, E., Dobie, R., Matchett, K. P., Henderson, N. C., Carrico, C., Driver, I., Borch Jensen, M., Chen, L., Petitjean, M., Bhattacharya, D., Fiel, M. I., Liu, X., Kisseleva, T., Alon, U., Adler, M., Medzhitov, R., & Friedman, S. L. (2023). An autocrine signaling circuit in hepatic stellate cells underlies advanced fibrosis in nonalcoholic steatohepatitis. *Science Translational Medicine*, 15(677), eadd3949.
<https://doi.org/10.1126/scitranslmed.add3949>
- Wang, S., Sun, S.-T., Zhang, X.-Y., Ding, H.-R., Yuan, Y., He, J.-J., Wang, M.-S., Yang, B., & Li, Y.-B. (2023). The Evolution of Single-Cell RNA Sequencing Technology and

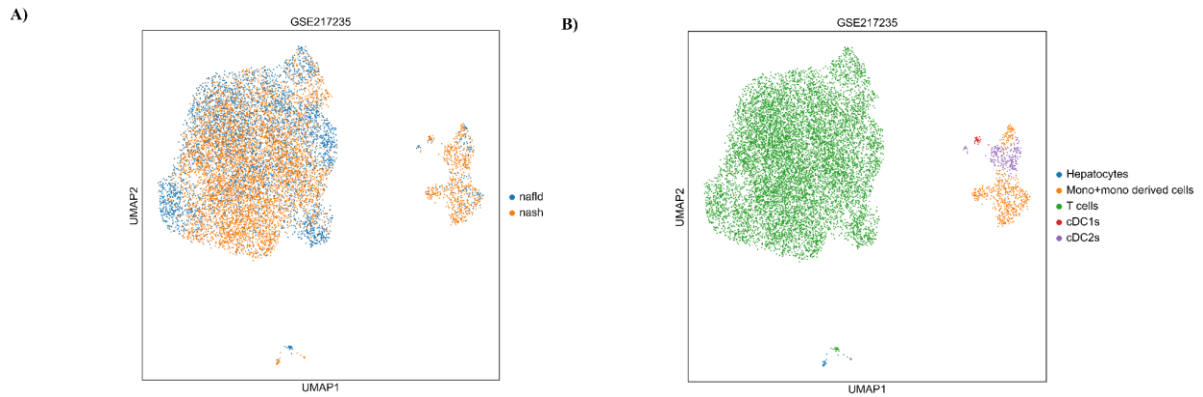
- Application: Progress and Perspectives. *International Journal of Molecular Sciences*, 24(3), 2943. <https://doi.org/10.3390/ijms24032943>
- Wang, S.-W., Herriges, M. J., Hurley, K., Kotton, D. N., & Klein, A. M. (2022). CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nature Biotechnology*, 40(7), 1066–1074. <https://doi.org/10.1038/s41587-022-01209-1>
- Wang, T., Li, B., Nelson, C. E., & Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, 20(1), 40. <https://doi.org/10.1186/s12859-019-2599-6>
- Wang, W., Cen, Y., Lu, Z., Xu, Y., Sun, T., Xiao, Y., Liu, W., Li, J. J., & Wang, C. (2024). scCDC: a computational method for gene-specific contamination detection and correction in single-cell and single-nucleus RNA-seq data. *Genome Biology*, 25(1), 136. <https://doi.org/10.1186/s13059-024-03284-w>
- Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>
- Watson, B., Paul, B., Amir-Zilberstein, L., Segerstolpe, A., Rahman, R. U., Shih, A., Deguine, J., Xavier, R. J., Moffitt, J. R., & Mullen, A. C. (2024). Spatial transcriptomics of healthy and fibrotic human liver at single-cell resolution. *BioRxiv*. <https://doi.org/10.1101/2024.02.02.578633>
- Wei, W., Xia, X., Li, T., Chen, Q., & Feng, X. (2024). Shaoxia: a web-based interactive analysis platform for single cell RNA sequencing data. *BMC Genomics*, 25(1), 402. <https://doi.org/10.1186/s12864-024-10322-1>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. <https://cran.r-project.org/web/packages/stringr/index.html>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. <https://dplyr.tidyverse.org>
- Wickham, H., Hester, J., & François, R. (2023). *readr: Read Rectangular Text Data*. <https://cran.r-project.org/web/packages/readr/index.html>
- Woestemeier, A., Scognamiglio, P., Zhao, Y., Wagner, J., Muscate, F., Casar, C., Siracusa, F., Cortesi, F., Agaloti, T., Müller, S., Sagebiel, A., Konczalla, L., Wahib, R., Karstens, K.-F., Giannou, A. D., Duprée, A., Wolter, S., Wong, M. N., Mühlig, A. K., ... Gagliani, N. (2023). Multicytokine-producing CD4⁺ T cells characterize the livers of patients with NASH. *JCI Insight*, 8(1). <https://doi.org/10.1172/jci.insight.153831>
- Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 15. <https://doi.org/10.1186/s13059-017-1382-0>
- Wolock, S. L., Lopez, R., & Klein, A. M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4), 281-291.e9. <https://doi.org/10.1016/j.cels.2018.11.005>
- Wong, W.-K., & Chan, W.-K. (2021). Nonalcoholic Fatty Liver Disease: A Global Perspective. *Clinical Therapeutics*, 43(3), 473–499. <https://doi.org/10.1016/j.clinthera.2021.01.007>

- Wu, S., Wang, X., Xing, W., Li, F., Liang, M., Li, K., He, Y., & Wang, J. (2023). An update on animal models of liver fibrosis. *Frontiers in Medicine*, 10. <https://doi.org/10.3389/fmed.2023.1160053>
- Wu, Y., & Zhang, K. (2020). Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nature Reviews Nephrology*, 16(7), 408–421. <https://doi.org/10.1038/s41581-020-0262-0>
- Xi, N. M., & Li, J. J. (2021). Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Systems*, 12(2), 176-194.e6. <https://doi.org/10.1016/j.cels.2020.11.008>
- Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Frontiers in Genetics*, 12. <https://doi.org/10.3389/fgene.2021.646936>
- Xiao, Y., Batmanov, K., Hu, W., Zhu, K., Tom, A. Y., Guan, D., Jiang, C., Cheng, L., McCright, S. J., Yang, E. C., Lanza, M. R., Liu, Y., Hill, D. A., & Lazar, M. A. (2023). Hepatocytes demarcated by EphB2 contribute to the progression of nonalcoholic steatohepatitis. *Science Translational Medicine*, 15(682), eadc9653. <https://doi.org/10.1126/scitranslmed.adc9653>
- Xie, Y., Cheng, J., & Tan, X. (2023). *DT: A Wrapper of the JavaScript Library 'DataTables'*. <https://cran.r-project.org/web/packages/DT/index.html>
- Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., & Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1). <https://doi.org/10.15252/msb.20209620>
- Xu, C., Prete, M., Webb, S., Jardine, L., Stewart, B. J., Hoo, R., He, P., Meyer, K. B., & Teichmann, S. A. (2023). Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell*, 186(26), 5876-5891.e20. <https://doi.org/10.1016/j.cell.2023.11.026>
- Xu, X., Poulsen, K. L., Wu, L., Liu, S., Miyata, T., Song, Q., Wei, Q., Zhao, C., Lin, C., & Yang, J. (2022). Targeted therapeutics and novel signaling pathways in non-alcohol-associated fatty liver/steatohepatitis (NAFL/NASH). *Signal Transduction and Targeted Therapy*, 7(1), 287. <https://doi.org/10.1038/s41392-022-01119-3>
- Yang, C.-R., Lin, W.-J., Shen, P.-C., Liao, P.-Y., Dai, Y.-C., Hung, Y.-C., Lai, H., Mehmood, S., Cheng, W.-C., & Ma, W.-L. (2024). Phenotypic and metabolomic characteristics of mouse models of metabolic associated steatohepatitis. *Biomarker Research*, 12(1), 6. <https://doi.org/10.1186/s40364-023-00555-9>
- Yang, P., Huang, H., & Liu, C. (2021). Feature selection revisited in the single-cell era. *Genome Biology*, 22(1), 321. <https://doi.org/10.1186/s13059-021-02544-3>
- Yang, S., Corbett, S. E., Koga, Y., Wang, Z., Johnson, W. E., Yajima, M., & Campbell, J. D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology*, 21(1), 57. <https://doi.org/10.1186/s13059-020-1950-6>
- Ye, H., Sun, M., Su, M., Chen, D., Liu, H., Ma, Y., Luo, W., Li, H., & Xu, F. (2023). Identification of disease-related genes and construction of a gene co-expression database in non-alcoholic fatty liver disease. *Frontiers in Genetics*, 14. <https://doi.org/10.3389/fgene.2023.1070605>

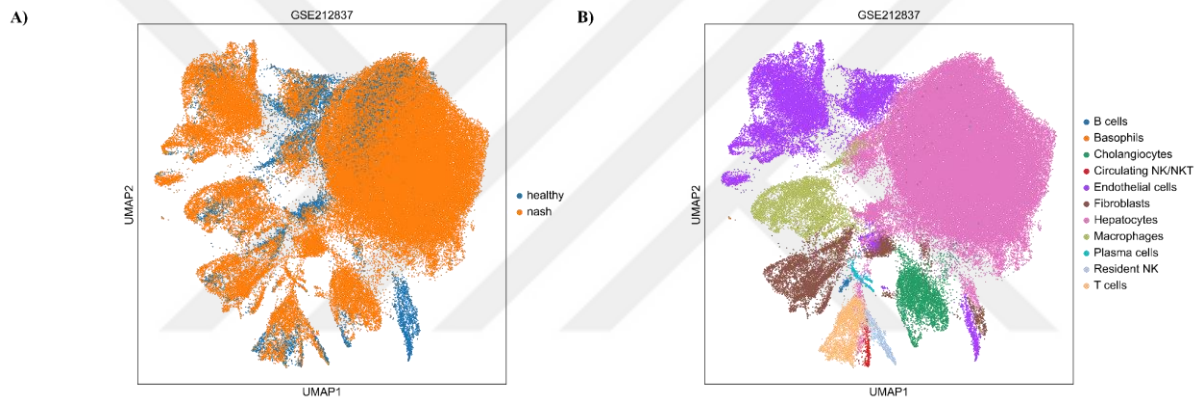
- You, Y., Tian, L., Su, S., Dong, X., Jabbari, J. S., Hickey, P. F., & Ritchie, M. E. (2021). Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. *Genome Biology*, 22(1), 339. <https://doi.org/10.1186/s13059-021-02552-3>
- Young, M. D., & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, 9(12). <https://doi.org/10.1093/gigascience/giaa151>
- Younossi, Z. M., Golabi, P., Paik, J. M., Henry, A., Van Dongen, C., & Henry, L. (2023). The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. *Hepatology*, 77(4), 1335–1347. <https://doi.org/10.1097/HEP.0000000000000004>
- Younossi, Z. M., Koenig, A. B., Abdelatif, D., Fazel, Y., Henry, L., & Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*, 64(1), 73–84. <https://doi.org/10.1002/hep.28431>
- Younossi, Z., Tacke, F., Arrese, M., Chander Sharma, B., Mostafa, I., Bugianesi, E., Wai-Sun Wong, V., Yilmaz, Y., George, J., Fan, J., & Vos, M. B. (2019). Global Perspectives on Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis. *Hepatology*, 69(6), 2672–2682. <https://doi.org/10.1002/hep.30251>
- Yuan, X., Waterworth, D., Perry, J. R. B., Lim, N., Song, K., Chambers, J. C., Zhang, W., Vollenweider, P., Stirnadel, H., Johnson, T., Bergmann, S., Beckmann, N. D., Li, Y., Ferrucci, L., Melzer, D., Hernandez, D., Singleton, A., Scott, J., Elliott, P., ... Mooser, V. (2008). Population-Based Genome-wide Association Studies Reveal Six Loci Influencing Plasma Levels of Liver Enzymes. *The American Journal of Human Genetics*, 83(4), 520–528. <https://doi.org/10.1016/j.ajhg.2008.09.012>
- Zhai, M., Liu, Z., Long, J., Zhou, Q., Yang, L., Zhou, Q., Liu, S., & Dai, Y. (2021). The incidence trends of liver cirrhosis caused by nonalcoholic steatohepatitis via the GBD study 2017. *Scientific Reports*, 11(1), 5195. <https://doi.org/10.1038/s41598-021-84577-z>
- Zhang, A. W., O’Flanagan, C., Chavez, E. A., Lim, J. L. P., Ceglia, N., McPherson, A., Wiens, M., Walters, P., Chan, T., Hewitson, B., Lai, D., Mottok, A., Sarkozy, C., Chong, L., Aoki, T., Wang, X., Weng, A. P., McAlpine, J. N., Aparicio, S., ... Shah, S. P. (2019). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods*, 16(10), 1007–1015. <https://doi.org/10.1038/s41592-019-0529-1>
- Zhang, H., Ma, Y., Cheng, X., Wu, D., Huang, X., Chen, B., Ren, Y., Jiang, W., Tang, X., Bai, T., Chen, Y., Zhao, Y., Zhang, C., Xiao, X., Liu, J., Deng, Y., Ye, T., Chen, L., Liu, H.-M., ... Cao, Z. (2021). Targeting epigenetically maladapted vascular niche alleviates liver fibrosis in nonalcoholic steatohepatitis. *Science Translational Medicine*, 13(614), eabd1206. <https://doi.org/10.1126/scitranslmed.abd1206>
- Zhang, Q., Zhang, B., Yang, F., Hu, Y., Fan, R., Wang, M., & Chen, S. (2024). Forsythoside A regulates pulmonary fibrosis by inhibiting endothelial-to-mesenchymal transition and lung fibroblast proliferation via the PTPRB signaling. *Phytomedicine*, 130, 155715. <https://doi.org/10.1016/j.phymed.2024.155715>

- Zhang, S., Li, X., Lin, J., Lin, Q., & Wong, K.-C. (2023). Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *RNA*, 29(5), 517–530. <https://doi.org/10.1261/rna.078965.121>
- Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., Ping, Y., Li, F., Shi, A., Bai, J., Zhao, T., Li, X., & Xiao, Y. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*, 47(D1), D721–D728. <https://doi.org/10.1093/nar/gky900>
- Zhang, Y., Ma, Y., Huang, Y., Zhang, Y., Jiang, Q., Zhou, M., & Su, J. (2020). Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Computational and Structural Biotechnology Journal*, 18, 2953–2961. <https://doi.org/10.1016/j.csbj.2020.10.007>
- Zhao, D., Zhang, X., Tang, Y., Guo, P., Ai, R., Hou, M., Wang, Y., Yuan, X., Cui, L., Zhang, Y., Zhao, S., Li, W., Wang, Y., Sun, X., Liu, L., Dong, S., Li, L., Zhao, W., & Nan, Y. (2022). Identification and Validation of Novel Biomarkers for Hepatocellular Carcinoma, Liver Fibrosis/Cirrhosis and Chronic Hepatitis B via Transcriptome Sequencing Technology. *Journal of Hepatocellular Carcinoma, Volume 9*, 389–403. <https://doi.org/10.2147/JHC.S357380>
- Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., ... Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), 14049. <https://doi.org/10.1038/ncomms14049>
- Zhou, T., Kiran, M., Lui, K. O., & Ding, Q. (2022). Decoding liver fibrogenesis with single-cell technologies. *Life Medicine*, 1(3), 333–344. <https://doi.org/10.1093/lifemedi/lnac040>
- Zimmerman, K. D., Espeland, M. A., & Langefeld, C. D. (2021). A practical solution to pseudoreplication bias in single-cell studies. *Nature Communications*, 12(1), 738. <https://doi.org/10.1038/s41467-021-21038-1>

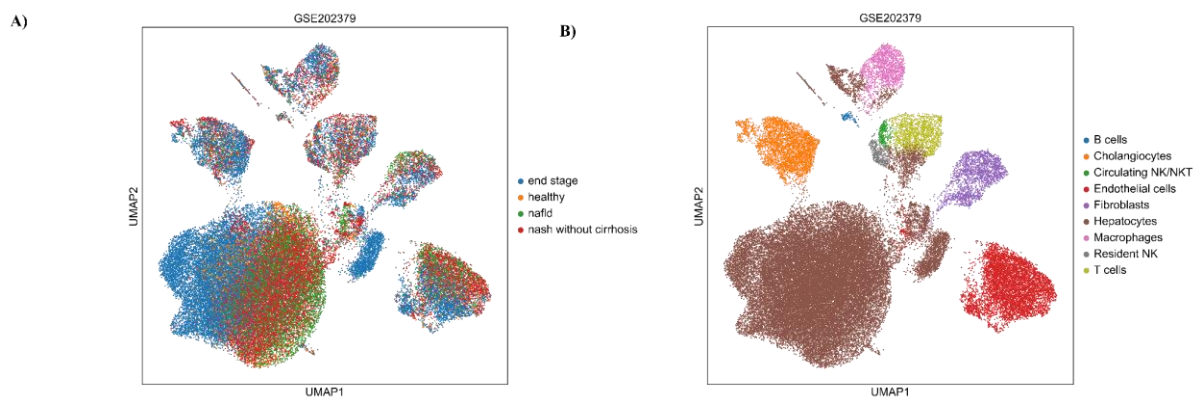
Appendix



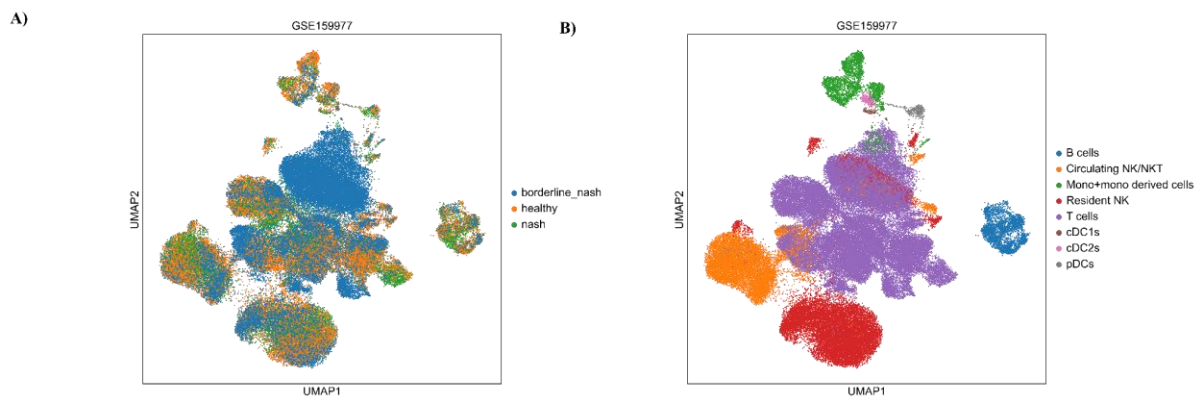
Appendix Figure 1 UMAP visualisation of cell type annotation of GSE212735 A) Conditions of the cells, B) Cell type annotations.



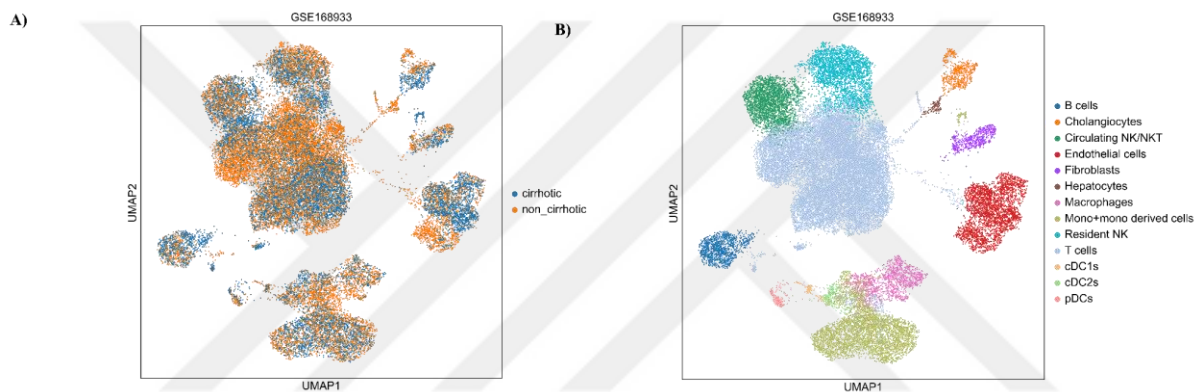
Appendix Figure 2 UMAP visualisation of cell type annotation of GSE212837 A) Conditions of the cells, B) Cell type annotations.



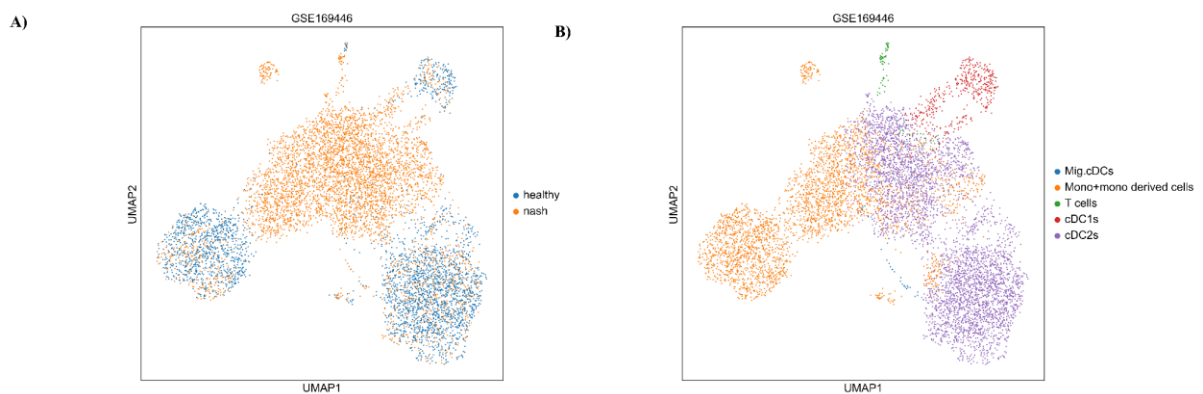
Appendix Figure 3 UMAP visualisation of cell type annotation of GSE202379 A) Conditions of the cells, B) Cell type annotations.



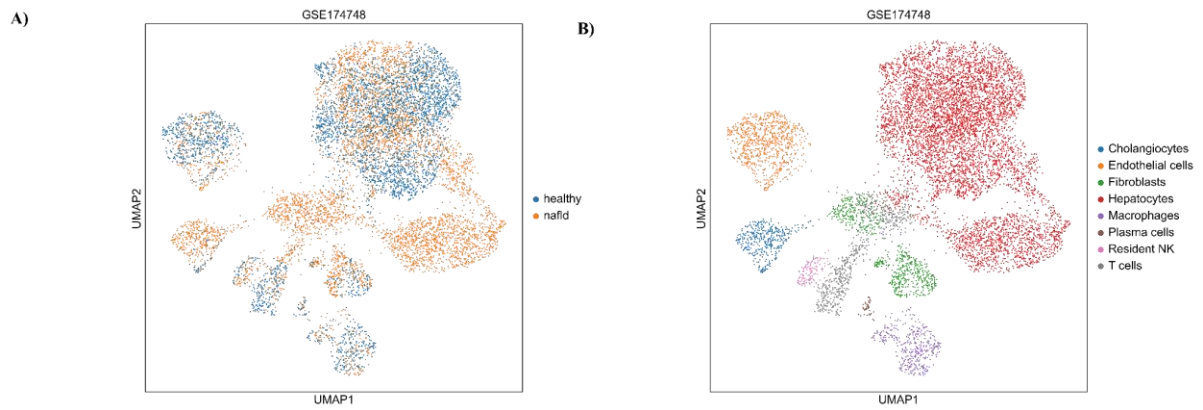
Appendix Figure 4 UMAP visualisation of cell type annotation of GSE159977 A) Conditions of the cells, B) Cell type annotations.



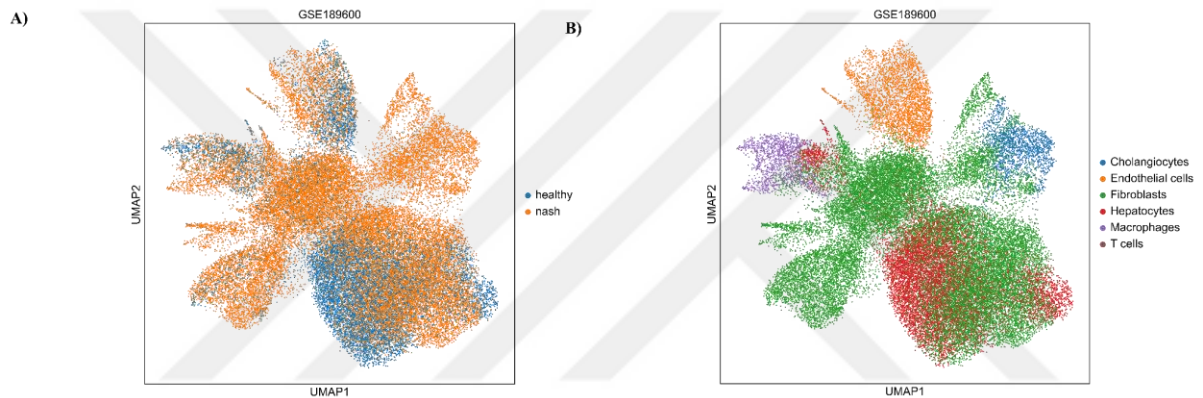
Appendix Figure 5 UMAP visualisation of cell type annotation of GSE168933 A) Conditions of the cells, B) Cell type annotations.



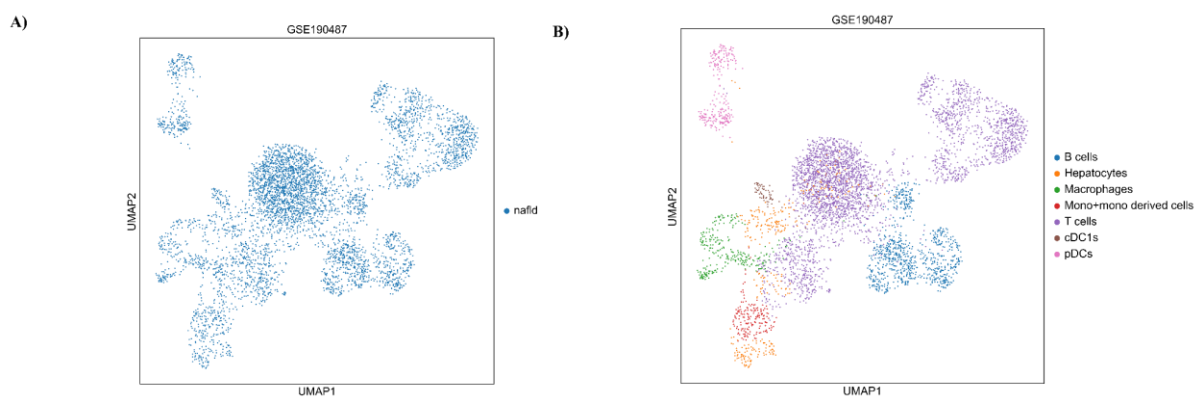
Appendix Figure 6 UMAP visualisation of cell type annotation of GSE169446 A) Conditions of the cells, B) Cell type annotations.



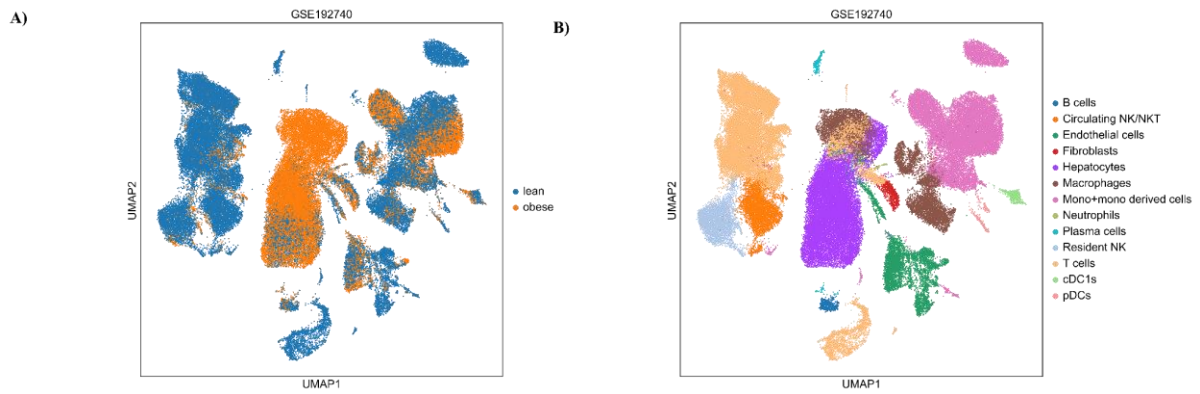
Appendix Figure 7 UMAP visualisation of cell type annotation of GSE174748 A) Conditions of the cells, B) Cell type annotations.



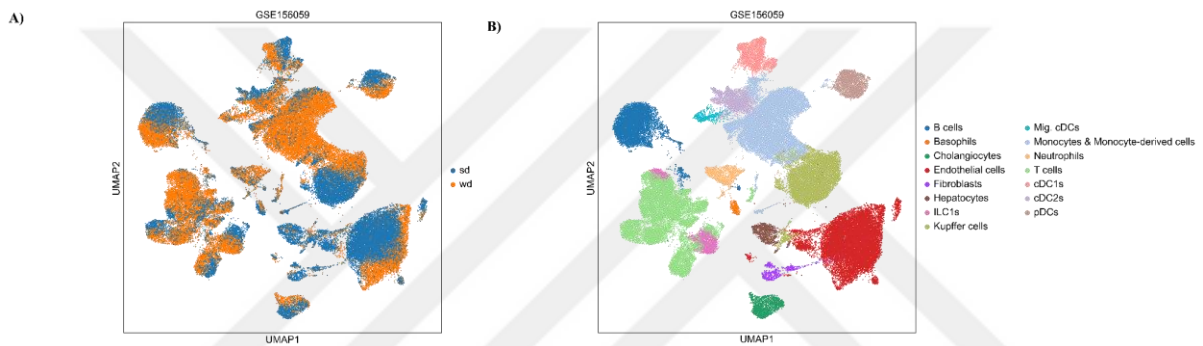
Appendix Figure 8 UMAP visualisation of cell type annotation of GSE189600 A) Conditions of the cells, B) Cell type annotations.



Appendix Figure 9 UMAP visualisation of cell type annotation of GSE190487 A) Conditions of the cells, B) Cell type annotations.

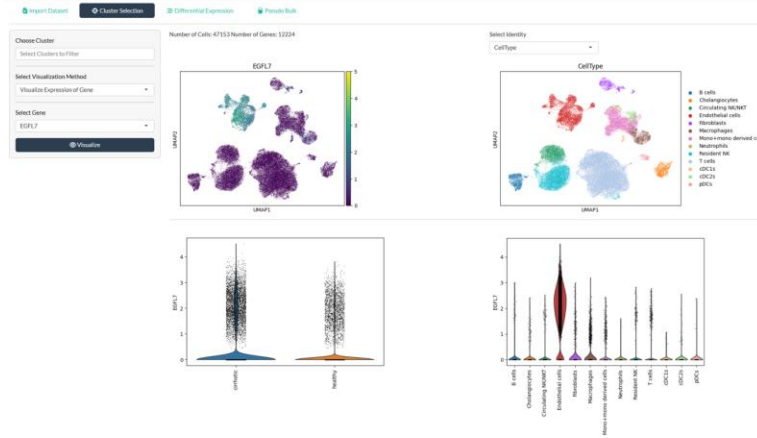


Appendix Figure 10 UMAP visualisation of cell type annotation of GSE192740 A) Conditions of the cells, B) Cell type annotations.

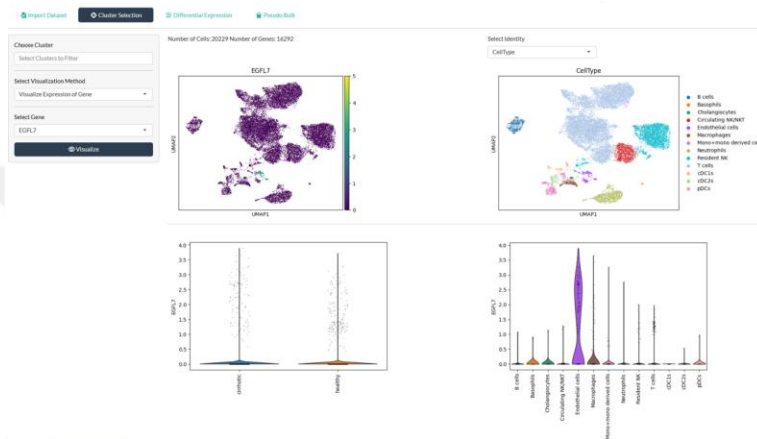


Appendix Figure 11 UMAP visualisation of cell type annotation of GSE156059 A) Conditions of the cells, B) Cell type annotations.

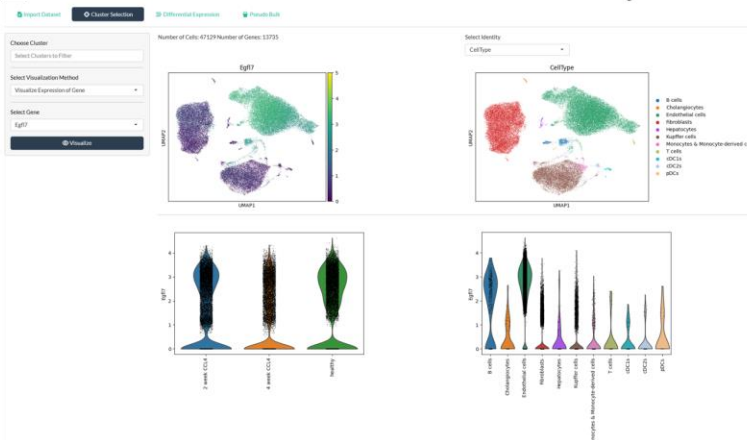
A)



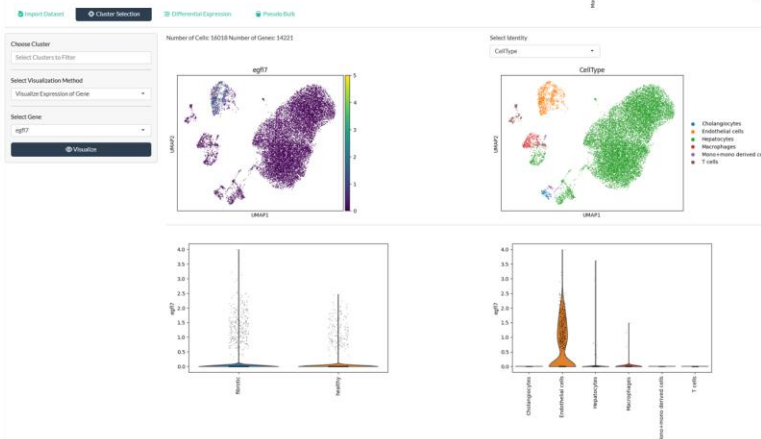
B)



C)



D)



Appendix Figure 12 Cluster Selection Tab: *EGFL7* Expression in Integrated Datasets

Appendix Table 1. Gene List Used as Cell Type Markers Used for Enrichment

Endothelial		Fibroblast
APPBP2	GGA2	NDUFA4L2
ARGLU1	GNG11	RGS5
ATP10D	GPR116	ACTA2
BNIP1	HES1	ACTA2
BST2	HYI	COL1A1
BTNL9	IL1R1	RBP1
C11orf96	IL33	TIMP1
C3orf58	KIF1B	VIM
CCDC85B	KLHL28	
CCL14	LDB2	
CD4	MCM3AP	
CDC73	MEF2C	
CDKL1	MFN1	
CHD4	MGAT5	
CLEC1B	MPZL3	
CLEC2B	MRC1	
CLEC4G	OIT3	
CLEC4M	PIK3C2A	
CRBN	PPAP2B	
CRHBP	PPWD1	
CTSL	RASGEF1B	
DCLRE1C	RASGRP3	
DLC1	RELN	
DLG1	RIN2	
DNASE1L3	SECISBP2L	
DOCK1	STAB1	
DUSP5	STAB2	
EFNB2	STRN3	
F8	TCF12	
FCN2	TEX264	
FCN3	TRAPPC11	
FILIP1	TSPAN7	
FOSB	USP48	
FXYD6	ZNF286B	
GBP4		