Technische Universität München
TUM School of Medicine and Health

Single cell transcriptomics unravels cellular heterogeneity in hepatocytes

Maria Lucia Johanna Richter

Vollständiger Abdruck der von der TUM School of Medicine and Health der
Technischen Universität München zur Erlangung einer
Doktorin der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation

Vorsitz: Prof. Dr. Eleftheria Zeggini

Prüfer*innen der Dissertation:

1. TUM Junior Fellow Dr. Celia Martinez Jimenez

2. Prof. Dr. Maria Colomé-Tatché

Die Dissertation wurde am 11.07.2023 bei der
Technischen Universität München eingereicht
und durch die TUM School of Medicine and Health am 08.11.2023
angenommen.

Abstract

The liver is a histologically homogeneous tissue responsible for the metabolism of endogenous and exogenous compounds. These metabolic functions are dominantly performed by its most abundant cell type: the hepatocytes. Apart from the hepatocytes, that make up between 50 and 85 % of the cells, the liver is host to other cell types, such as cholangiocytes, endothelial cells, hepatic stellate cells, Kupffer cells, and immune cells [1].

Despite their homogeneity under the microscope, evidence has accumulated suggesting several layers of hepatocyte heterogeneity. For instance, hepatocytes can naturally undergo whole genome duplication, resulting in polyploid hepatocytes. Whilst failure of nuclear division leads to cells containing a single nucleus with duplicated genome content, i.e. tetraploidy, complete nuclear division before failure in cytokinesis results in cells containing two nuclei each containing one set of chromosomes, i.e. bi-nucleated diploid cells [2]. Initially, polyploidy has been suggested to act protective against hepatocellular carcinoma (HCC) as more genome copies could prevent loss of heterozygosity events and buffer against mutational damage [3]. However, higher amounts of polyploid hepatocytes have been observed in HCC as well as in non-alcoholic and alcoholic fatty liver disease (NAFLD/ALFD) [4][5]. The first project of this doctoral thesis therefore aims to investigate the role of tetraploid hepatocytes in young healthy adult mice at single cell resolution to define a reference baseline for future studies.

Apart from polyploidyzation, the structural organization of the liver serves as another source of heterogeneity for hepatocytes. The liver is organized into lobules, with each lobule featuring a central vein in the middle and the portal triad at the vertices. Nutrients and oxygen are supplied to the tissue through the portal triad, whereas the central vein collects oxygen-deprived blood. Zonation describes the resulting difference in expression profiles along the gradient of nutrients and oxygen between portal and central vein [6]. Analyzing data obtained through single nucleus RNA sequencing in this thesis revealed that zonation and polyploidyzation are intertwined.

Moreover, hepatic steatosis has been shown to impact zonation in individual hepatocytes [7]. Hepatic steatosis is the first indication for the onset and progression of NAFLD, which affects 25 % of the worldwide population [8]. NAFLD has furthermore been associated to changes in the drug-related metabolic pathways [9]. During clinical trials, when testing new drugs for efficacy and safety, the gold standard model is the 2D culture of primary human hepatocytes. However, in the absence of the 3D architecture in this culture system, liver zonation is lost. In the second project of this doctoral thesis, it was hence investigated to which extent hepatocyte heterogeneity is preserved in the absence of zonation for a seemingly homogeneous population of cells. To address the impact of intracellular fat accumulation on the metabolism, this thesis presents a thorough analysis of how it changes the transcriptional profile and the drug-metabolic capacity of single hepatocytes.

Zusammenfassung

Die Leber ist ein histologisch homogenes Organ, das für den Metabolismus von endogenen und exogenen Stoffen verantwortlich ist. Diese Stoffwechselfunktionen werden hauptsächlich vom häufigsten Zelltyp der Leber ausgeführt: Hepatozyten. Neben Hepatozyten, die zwischen 50 und 85 % der Zellen ausmachen, gibt es in der Leber weitere Zelltypen wie Cholangiozyten, Endothelzellen, hepatische Sternzellen, Kupffer Zellen und andere Immunzellarten [1].

Obwohl sie unter dem Mikroskop homogen erscheinen, häuften sich in den letzten Jahren Hinweise darauf, dass Hepatozyten auf mehreren Ebenen heterogen sind. Zum Beispiel können Hepatozyten auf natürliche Weise ihr Genom duplizieren, was zu polyploiden Hepatozyten führt. Dabei führt unvollständige Zellkernteilung zu einzelnen Zellen mit duplizierten Genom im Zellkern (Tetraploidie), während unvollständige Teilung der ganzen Zelle nach abgeschlossener Zellkernteilung zu Zellen mit zwei jeweils diploiden Zellkernen führt (Bi-nukleare diploide Zellen) [2]. Ursprünglich dachte man Polyploidie könne gegen hepatozelluläre Karzinome (HCC) schützen indem mehrere Genkopien den Verlust von Heterozygotie in den Zellen verhindern [3]. Jedoch wurde sowohl in HCC als auch in nicht-alkoholischer Fettlebererkrankung (NAFLD) ein höherer Anteil an polyploiden Zellen beobachtet [4][5]. Daher hat das erste Projekt dieser Disseration das Ziel die Rolle von Polyploidie in einzelnen Zellen gesunder, erwachsener Mäusen zu erforschen um damit eine Grundlage für künftige Forschung zu schaffen.

Neben der Genomduplikation stellt die strukturelle Organisation der Leber eine weitere Quelle von Heterogenität zwischen Hepatozyten dar. Die Leber ist in Leberläppchen organisiert, die jeweils eine Zentralvene in ihrer Mitte und eine Lebertrias an jeder Ecke aufweisen. Das Organ wird von der Lebertrias mit Sauerstoff und Nährstoffen versorgt, während die Zentralvene dazu dient, sauerstoffarmes Blut aufzunehmen und zum Herzen zu transportieren. Zonierung beschreibt den daraus resultierenden interzellulären Unterschied der Genexpression entlang des Nähr- und Sauerstoffgradienten [6]. Die Analyse von Sequenzierdaten des Transkriptoms einzelner Zellkerne in dieser Dissertation

führte zur Erkenntnis, dass ein Zusammenhang zwischen Genomduplizierung und Zonierung besteht.

Darüber hinaus wurde in aktuellen Studien beobachtet, dass Fettleber die Zonierung einzelner Hepatozyten beeinflusst [7]. Fettleber ist das erste Anzeichen sich entwickelnder NAFLD, die 25 % der weltweiten Population betrifft [8]. NAFLD ist assoziiert mit Veränderungen in den genetischen Mechanismen, die für den Stoffwechsel von Medikamenten verantwortlich sind [9]. In klinischen Studien, in denen neue Medikamente auf ihre Wirksamkeit und Sicherheit getestet werden, dient die 2D Zellkultur von primären humanen Hepatozyten (PHH) als Goldstandardmodell. Durch das Fehlen der 3D Struktur der Leber geht jedoch die Zonierung der Zellen in diesem Modell verloren. Das zweite Projekt dieser Dissertation erforscht daher in welchem Ausmaß Heterogenität zwischen morphologisch homogenen Hepatozyten vorhanden ist in Abwesenheit von Zonierung. Um die Auswirkungen von intrazellulärer Lipidanreicherung auf den Stoffwechsel der Zellen zu untersuchen werden in dieser Dissertation Veränderungen in der Genexpression analyisiert, besonders im Bezug auf die Stoffwechselkapazität von Medikamenten.

Contents

1 Introduction

The liver is responsible for the metabolism of endogenous and exogenous compounds. Among these are lipids, carbohydrates, and proteins, as well as xenobiotics, such as alcohol and drugs. Through the metabolism of lipids, the liver is responsible for maintaining lipid and cholesterol, and also energy homeostasis [1]. Moreover, the liver involves in the regulation of blood volume and endocrine control of growth signaling pathways [1]. These metabolic functions are mainly performed by the predominant liver cell type: the hepatocytes. As the liver is composed of 70-85 % hepatocytes in humans and 50-70 % hepatocytes in mice, the tissue appears histologically homogeneous when inspected under the microscope suggesting functional homogeneity [10][11]. Despite these morphological observations, evidence on cellular heterogeneity has been accumulating over the last years [12][4][11][13][14][15][16][17][18][19].
Thus, this thesis investigates layers of cellular heterogeneity in a tissue that has traditionally been considered homogeneous. The goal is to understand these layers in their role and relationship on tissue function.

1.1 Liver cell type composition

Hepatocytes are the predominant cell type of the liver, performing major metabolic functions [10][11]. In addition, liver is also composed of other, non-parenchymal cell types (NPCs) contributing to functional heterogeneity in the liver. These cell types include biliary epithelial cells (cholangiocytes), Kupffer cells and other immune cells, hepatic stellate cells, and sinusoidal endothelial cells (Figure 1.1)[1].

As part of the non-parenchymal compartment, cholangiocytes line the lumen of the bile duct where they involve in bile modification and secretion 1.1. At only 3-5 %, they represent a small proportion of all cells in the human liver [20]. Together with hepatocytes, cholangiocytes stem from primitive hepatocytes, often referred to as hepatoblasts [21]. A recent study has described a hybrid cell type sharing hepatocyte and cholangiocyte expression profiles that has the potential to regenerate liver tissue upon injury, highlighting the functional similarities between cholangiocytes and hepatocytes [22].

Kupffer cells are liver-specific macrophages that reside permanently in the organ, making up roughly one third of the non-parenchymal population [1][23]. They detect pathogens coming from the gut as well as apoptotic cells within the liver and regulate inflammatory mechanisms upon exposure to these stimuli [24]. Classically, macrophages in the liver are classified in two groups based on their pro-, and anti-inflammatory gene expression pathway profiles, respectively [25]. These two populations were further characterized at single cell resolution by MacParland *et al.* 2018 [18]. However, data in Su *et al.* 2021 led to no clear separation of macrophages into those two groups [26]. This study instead distinguished macrophages by classifying them as either Kupffer cells or monocyte-derived macrophages and argue that Kupffer cells have higher expression levels of pro-inflammatory markers than monocyte-derived macrophages, and vice versa [26]. Moreover, monocyte-derived macrophages have been observed to increase in proportion upon fat accumulation and upon 2,3,7,8-tetrachlorodibenzo-ρ-dioxin (TCDD)-treatment [26][27][28]. These findings indicate the presence of unapprehended cellular heterogeneity within a singular cell type that is dependent on the conditional context of the tissue. Other immune cells in the liver include neutrophil granulocytes, B- and T-lymphocytes, and natural killer (NK) cells. While these cell types play important roles during inflammation and pathogen infiltration in the tissue, they are not specific to the liver [26].

Hepatic stellate cells are pericytes making up around 5 % of the cells in the liver and exist in two states, a quiescent and an active state [23]. In their quiescent state they store vitamin A droplets, whereas upon activation, they rid themselves of these droplets to enter a proliferative state and contribute to the formation of scarring tissue during liver fibrosis [29][6].

Lastly, the most abundant non-parenchymal cell types are liver sinusoidal endothelial cells (LSECs), comprising around 50 % of all NPCs [23]. LSECs form fenestrated sieve plates that facilitate exchange of metabolites and oxygen between hepatocytes and the blood plasma while also functioning as a barrier between the plasma and the liver cell types [1]. Liver endothelial cells have furthermore been shown to be intertwined with other NPCs, and hepatocytes. For instance, Su *et al.* have observed chimeric cell populations of endothelial-chimeric Kupffer cells, endothelial-chimeric stellate cells, and endothelial-chimeric hepatocytes [26]. The latter have also been observed in a study by Xiong *et al.* [27]. Furthermore, the co-culture of endothelial cells and hepatocytes has been observed to improve the liver-metabolic functionality of hepatocytes [30]. Together this indicates that liver cell types are functionally intertwined and highly heterogeneous.

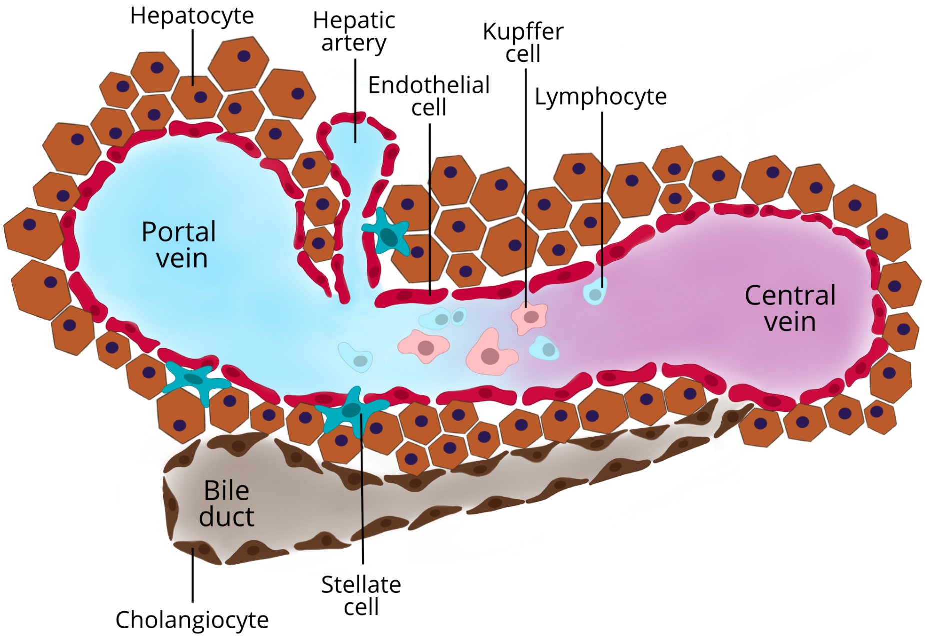

Figure 1.1: Schematic overview of the cell types within the liver. Image adapted from Ben-Moshe *et al.* [6]

1.1.1 Cell type identification

Traditionally, cell types are defined by their morphology and surface marker expression, which can also be used to isolate a given cell type from a tissue through fluorescent-activated cell sorting (FACS) building on established cell type characteristics [31]. However, cell-to-cell variability has been observed in populations that were previously thought of as homogeneous [32]. Hence, sequencing a bulk of pre-selected cells relying on established marker genes limits the opportunities to identify new cell types and cellular sub-types due to lack of resolution. Technological advances of the last decade have enabled the study of cellular heterogeneity in depth in an unsupervised way through the sequencing of single cells [33][32]. A nowadays common practice is to sequence the single cell transcriptome as a readout of cellular function [33]. This is possible because RNA expression has been shown to serve as an approximation of protein expression [34]. In particular, while the whole transcriptome levels have been shown to correlate poorly to protein expression [35][36], higher correlations have been observed for differentially expressed genes (DEGs) [34]. Hence, the cellular transcriptome represents an easy to obtain, approximate measure of cellular identity [34]. Single cell RNA-sequencing (scRNA-seq) has thus vastly enhanced the characterization of functional cell types within tissues and led to the further detection of sub-populations in existing cell types across multiple species, in both, health and disease [37][38][39][40]. As a result, reference atlases are being built to map and characterize all cell types present in a species. For example, mouse has been used as a model organism for humans to extensively study and characterize the cell types in the mammalian brain [38]. Since 2018, a transcriptomic atlas named the *Tabula muris* exists as a reference for the cell types in 20 mouse organs and tissues, a data set that is constantly growing with addition of further cell types and organs [37]. In humans, the collaborative work of scientists worldwide coordinated in the project of the *Human Cell Atlas* (https://www.humancellatlas.org/) is aiming to achieve a comprehensive map of all cell types within the human body and their respective characteristics [39]. As part of this, recent atlas studies have provided a deep characterization of cell types in the heart [40] and in the lung [41][42]. Furthermore, analogous to the efforts in mouse, the *Tabula Sapiens* presents an overview of the cell types in 24 human organs and tissues [43]. Independently of species and tissue, the single cell portal (https://singlecell.broadinstitute.org/single_cell) provides an overview of newly generated single cell sequencing studies. Specifically for the liver, the liver cell atlas (https://www.livercellatlas.org) stores data sets of scRNA-seq for both, mouse and human. These reference atlases provide a powerful resource of characteristic gene expression profiles in populations of functionally similar cells.

This is of particular interest when annotating cell types in a scRNA-seq experiment. A key step in identifying and annotating cell types and states from scRNA-seq data is the correct assignment of cells into functionally similar groups. Computationally, this is usually done by constructing a neighborhood graph based on the k nearest neighbors of each single cell before applying a clustering algorithm to group cells together [44]. Deciding on the number of clusters to correctly capture the full heterogeneity of a given sample while keeping cells with same functionality grouped together hence represents an important, non-trivial task. Approaches to solve this problem include the calculation of similarity scores to separate distinct groups and minimize intra-cluster variability, e.g. through hierarchical clustering [45][46]. While calculation of similarity scores and subsequent hierarchical clustering represents an unsupervised approach, the final decision on the number of clusters is still subjective [46]. Thus, determining a reasonable number of clusters often relies on incorporating additional information, e.g. marker gene expression [45][46]. Additionally, performing power analysis can help to identify the minimal informative number of cells within a cluster, hence setting a limit to the maximum number of clusters. For instance, Vieth *et al.* have developed an R-based tool to explore the power to detect differential expression between groups of interest based on the experimental design [47]. Expanding on this, Schmid *et al.* further developed a method to calculate the power for detecting cell types contributing at a given proportion to a tissue [48]. These methods can therefore be used to assess the minimum portion of a cells that can be reliably compared to the rest and provide a guidance for filtering steps [48].

After the cells are clustered together based on their transcriptomic similarity, cell type annotation is usually performed. To that end, known reference marker gene sets can be used to identify and annotate groups of cells in scRNA-seq experiments. Advances in the computational analysis nowadays make it possible to directly use sets of reference genes to score single cells for a given cell type based on their gene expression profile [49]. Additionally, efforts have been and are being made to FACS-sort known population of cells based on surface markers to then deeply characterize them through scRNA-sequencing for building reference gene data bases [37][50][43]. As these reference atlases are composed of a large amount of cells, they can be used for transfer learning, i.e. leveraging their cell type annotation onto clusters of cells from an unlabeled scRNA-seq experiment [51]. Expanding the knowledge about cell types, the combined information of clusters and marker gene expression further facilitates downstream identification and characterization of previously unknown cellular subtypes and cell states [16]. For instance in this thesis, markers of metabolic pathways are used to annotate subgroups of hepatocytes in a single cell RNA-sequencing (scRNA-seq) experiment.

1.1.2 Plate-based and droplet-based sequencing approaches

The first step of any scRNA-seq experiment is the isolation of single cells. Generally, two major approaches for this step of scRNA-seq technologies exist: plate-based and droplet-based methods. In plate-based approaches, single cells or nuclei are sorted into the wells of microtiter plates resulting in each well containing one single cell or nucleus [52][53][54]. Contrary to this, droplet-based approaches encapsulate the single cell or nucleus into a lipid droplet [55][39]. In both techniques, individual cells or nuclei receive a unique barcode allowing their later identification when cells are pooled for sequencing [56]. Plate-based and droplet-based techniques come with their own advantages and disadvantages. In general, plate-based approaches offer a lower throughput of sequenced single cells due to the restriction to individual plates [57][56]. However, on average, plate-based sequencing approaches generate high transcriptome coverage, with more reads being mapped to the reference genome and the detection of lowly abundant transcripts, especially the commonly used plate-based method SMART-seq2 [57][56]. Therefore, plate-based approaches are generally known to achieve a larger number of read counts across genes per single cell, referred to as the library size [57][56]. Furthermore, SMART-seq2 allows the sequencing of full-length transcripts, additionally increasing the yield of reads mapping to the reference genome [53]. This also enables the investigation of allele-specific and isoform expression [58].

On the other hand, droplet-based approaches, such as the platform of 10X Genomics, yield a high throughput of cells, although with the cost of a loss in sequencing depth [39][58][56]. These techniques are therefore especially suitable for studies exploring rare cellular sub-populations in a given context, e.g. an organ or a disease state [56][40]. As shown in this thesis, the choice of experimental approach has to be considered carefully as it impacts the results and the application of analysis methods. The high throughput in droplet-based techniques has greater power for the detection of cell sub-types. Contrary to that, full-length plate-based techniques yield high quality transcripts [56]. Hence, apart from the ability to identify and characterize new cellular subtypes, plate-based scRNA-seq techniques offer a powerful tool to perform in-depth comparisons of the functionality of groups of interest.

1.2 Polyploidization

A characteristic feature of hepatocytes is their tendency to undergo whole genome duplication, referred to as polyploidization [59][60][61]. Polyploidization occurs naturally in some mammalian tissues including, but not limited to, heart, muscle, and liver [62][63]. In the liver, the process starts after weaning and happens through the failure of cytokinesis [2][64]. As a result, polyploidy occurs in the adult mammalian liver and increases with age [61]. In mice, up to 80 % of hepatocytes are polyploid in adult animals, whereas in humans polyploid hepatocytes make up to 50 % of the adult liver [64]. Failure of nuclear division leads to cells containing a single nucleus with duplicated genome content, i.e. tetraploidy. Likewise, complete nuclear division before failure in cytokinesis results in cells containing two nuclei each containing one set of chromosomes, i.e. bi-nucleated diploid cells (Figure 1.2). Further incomplete divisions lead to mono- and bi-nucleated tetraploid, octoploid, and hexadecaploid hepatocytes. Nevertheless, the concrete role of these polyploid sub-populations remains unclear. Initially, polyploidy has been proposed to act protective against hepatocellular carcinoma (HCC) as more genome copies can prevent loss of heterozygosity events and buffer against mutational damage [65][3]. However, an enrichment of hepatocytes featuring high ploidy levels (>8n) has been observed in HCC and was associated to poor prognosis [5]. Moreover, tetraploid hepatocytes have been associated to intrahepatic lipid accumulation, therefore potentially accelerating progression of non-alcoholic and alcoholic fatty liver disease (NAFLD/ALFD) [4]. As part of this thesis, the effect of ploidy is explored in the healthy liver of young adult mice to establish a healthy reference for future disease studies.

1.2.1 Computationally accounting for different library sizes

Polyploid hepatocytes contain at least double the amount of genome and are larger in size than their diploid counterparts [65]. Furthermore, transcript concentration has been shown to be constant between cells of different sizes, which results in higher transcript abundance in larger cells [66][67]. When performing RNA-seq experiments, this can therefore lead to higher molecule counts in polyploid cells that do not necessarily translate to functional differences between diploid and polyploid cells [67]. The amount of mRNA molecules captured per single cells in a scRNA-seq experiment is referred to as the cell's library size [57]. Different library sizes between individual cells in a scRNA-seq experiment

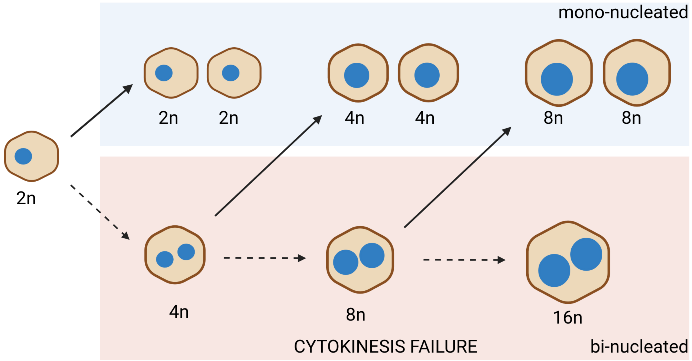

Figure 1.2: Failure in cytokinesis leads to polyploidization. Image adapted from Celton-Morizur *et al.* [13] and created in Biorender.com

can stem from technical artifacts, such as differences in reverse transcription efficiency between cells or differences in gene length contributing to bias during molecular sampling [68][69]. However, they can also reflect true biological variation [70][57]. Hence, computational normalization approaches are needed to address differences in library size between cells. The goal of normalization is for the normalized counts to represent true gene expression that is not influenced by extrinsic, technical factors. Library size correction is therefore a standard pre-processing step during the analysis of RNA-seq data, aiming to distinguish technical from biological differences [71][44]. Read counts per gene and sample are to be scaled in such manner that differences in sequencing depths between replicates, or single cells, are adjusted for and differential expression analysis reveals true biological variation. Therefore, finding the correct scaling factors for normalization can be complicated by true biological differences in overall mRNA content between samples, and by asymmetry of differential expression, i.e. one sample having an intrinsically higher amount of differentially expressed genes than the other [72]. Generally, there are two different approaches for read count normalization with different underlying assumptions. The first approach makes use of the read distribution for normalization purposes and comes with the assumption that there are the same amount of differentially expressed genes in all conditions and their expression is equally influenced by technical noise [72]. Counts from RNA-sequencing experiments are best modeled to follow a negative binomial distribution [73]. Anders *et al.* have developed an algorithm for bulk RNA-seq experiments assuming that most genes are not differentially expressed [74]. In their approach, library size normalization between samples is performed by calculating a sample-wise size factor. This is done by first calculating the ratios between read counts in a sample and the geometric mean across samples. The sample-wise size factor is defined as the median of these ratios [74]. For distribution-based normalization approaches in single cells, read counts from scRNA-seq experiments are often modeled following a zero-inflated negative-binomial distribution (ZINB) due to data sparsity [75]. As an example for such a method, the R-package *ZINB-WaVE* uses a global normalization factor based on the sample-level intercept in the ZINB, allowing to include gene-level as well as sample-level covariates in factor analysis [75]. However, it can be argued that an unconstrained negative binomial model can over-fit scRNA-seq data [69]. To overcome this, Hafemeister *et al.* have proposed a generalized linear model (GLM), in which the single cell library size is used as covariate [69]. Their method *sctransform* pools genes with similar expression levels across single cells and uses the Pearson residuals of their GLM of pooled genes against unique molecular identifier (UMI) counts to normalize the counts [69].

The second, and most commonly used normalization approach assumes the same total amount of mRNA between samples or individual cells. In its simplest form, this library size normalization aims to remove differences in sequencing depth by dividing the individual read counts per gene by the total number of reads in each sample [72]. Additionally, if full-length transcripts have been sequenced, it is advisable to correct for gene-length biases, i.e. longer genes have a higher potential to be sequenced more deeply [76]. As such, the method of calculating reads per kilobase per million mapped reads (RPKM) has been established for bulk RNA-sequencing experiments, where counts get divided by respective gene length in kilobases (kb) before they are summed up per sample and divided by the total library size in millions of reads. In single cell

RNA-sequencing experiments, normalization is often performed by summing the reads per cell and dividing by either the total library size across cells scaled by a factor or the average library size [77][78]. An adaptation of this approach for the single cell world is the adjustment of the scaling factor. As a single cell contains much fewer reads than a bulk sample, read counts are usually divided by the sum of reads per 10,000 to 200,000 instead of one million [77][79]. In general, this normalization approach is based on the assumption that all cells in an experiment have equal library sizes and differences between cells are purely technical.

However, in scRNA-seq, differences in the amount of transcripts have been observed between cell types, violating the assumption that each cell shares the same amount of total mRNA and rendering simple library size correction insufficient [70]. One of the most prominent methods accounting for differences in the amount of mRNA between cell types, is the one by Lun *et al.*, who have developed an R-based tool based on pooling cells with similar library sizes and calculating pool-wise size factors before deconvoluting these size factors back to the single cell level through random sampling [70]. This method also allows the use of external controls to model the relationship between a gene's average expression level and its corresponding variance in order to distinguish technical noise from true biological variation [70].

The addition of external controls represents another way to separate true biological differences from technical noise is. The most commonly used control is the addition of a regulated amount of synthetic spike-in molecules to each sample or single cell [80][81]. Initially created for bulk RNA-seq experiments, these molecules are designed in such manner that they do not align to mammalian reference genomes. In bulk experiments, it is advised to adjust the spike-in concentration such that spike-ins make up between 5 and 10 % of the library size [82]. Since the amount of mRNA molecules can differ from cell to cell, and it is not feasible to measure the amount of endogenous transcripts in every cell beforehand, the same amount of spike-ins is usually added to each single cell [82]. Between-sample differences in their coverage are then supposed to be purely technical [81]. The use of spike-ins therefore comes with the assumption that the spike-ins are unaffected by the biological condition and are subject to the same technical effects as the endogenous genes [72]. However, studies have shown that technical noise affects spike-ins differently than endogenous transcripts, violating the assumption that both are influenced in the same way [83][84].

Moreover, when sequencing to saturation, highly expressed genes represent a greater fraction of total sequenced molecules than less highly expressed genes [72][84]. Hence, it can be postulated that with a higher amount of endogenous mRNA in a cell (e.g. through polyploid cells producing more mRNA), the addition of spike-in molecules leads to a relatively higher ratio of endogenous transcripts to spike-ins than in cells with lower endogenous starting material [67][84]. To address this issue, a new normalization technique has been developed in this thesis, in which the scaling factor is adapted to minimize differences between cells with different cell size and genome content based on the ratio of spike-in molecules to endogenous transcripts. This is used to more accurately identify biological differences between diploid and tetraploid hepatocytes in the young adult mouse liver under physiological conditions (Chapter 3).

1.3 Liver zonation

While the specific functional roles of diploid and polyploid hepatocytes remain to be explored in depth, studies have shown that they show distinct positional preferences within the liver [85]. The liver spatial organization adds another layer of complexity to cellular heterogeneity. Supply of oxygen and metabolites to the liver happens through two large blood vessels: the hepatic artery and the portal vein [1]. The hepatic artery carries oxygen-rich blood from the aorta to the liver, whereas the portal vein transports metabolites from the gastrointestinal tract. Both blood vessels subdivide into smaller vessels that supply the cells within the organ with oxygen and metabolites. The liver is organized into roughly hexagonal tissue structures called hepatic lobules [1][6]. Each of the lobules' vertices consists of a capillary stemming from the portal vein (PV), surrounded by a capillary from the hepatic artery, and a bile duct. This structure of the three co-localized vessels is called the portal triad. The center of each lobule features a central vein (CV) collecting oxygen-deprived blood to transport back to the heart (Figure 1.3). It has been observed that diploid hepatocytes reside closer to the PV whereas tetra- and octoploid hepatocytes are found closer to the CV [85]. Zonation describes the difference in expression profiles along the gradient of nutrients and oxygen between central and portal vein [1][6]. Hence, the cells' transcriptomic profiles differ in relation to their position in proximity to central or portal vein, respectively [6].

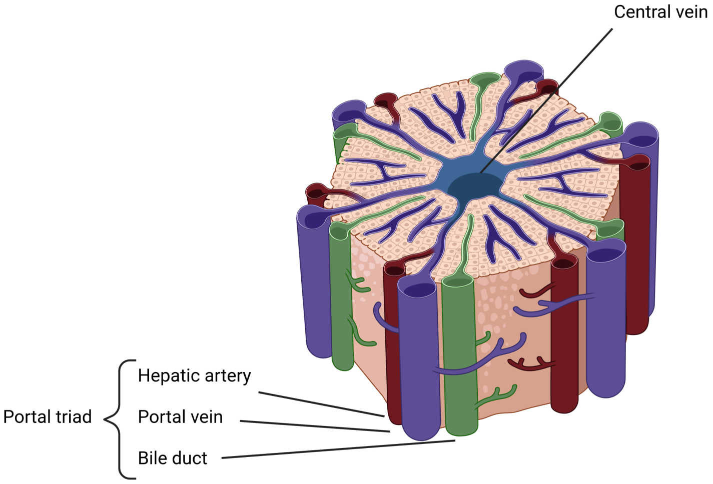

Figure 1.3: Hexagonal structure of the liver lobule with the central vein in the middle and the portal triad at the vertices. Image created in Biorender.com

1.3.1 Metabolism in the light of zonation

The labor division along the hepatic lobule leads to distinct functionality within the 3D architecture [16]. Blood from the nutrient-rich portal vein makes up 75 % of the blood volume in the liver, while only 25 % of it stems from the oxygen-rich portal artery, oxygen represents a limiting factor for hepatocytes [86][6]. Because of that, energy-demanding tasks are therefore rather performed periportally where oxygen is available in greater abundance [6]. For instance, periportal hepatocytes involve in lipid β-oxidation, urea- and gluconeogenesis, and the further secretion of glucose, proteins, hormones and bile. Meanwhile, pericentral hepatocytes are involved in glycolysis, lipogenesis, the uptake of cholesterol, bile and glutamine synthesis, and xenobiotic metabolism [6].

Traditionally, zonation has been studied by using (immuno-)histochemistry and *in situ* hybridization (ISH) [87]. Technological advances based on transcriptomic profiling have made it possible to study the gene expression profiles of individual hepatocytes with respect to their zonal location. As a first study to explore zonation on the transcriptomic level in mice, Braeuning *et al.* have isolated populations of pericentral, and periportal hepatocytes and used microarrays to address differential gene expression between these two groups [88]. Among the genes involved in the metabolism of carbohydrates, *Akr1b1* and *Idh1* were found to be enriched pericentrally whereas *Pck1* was found periportally, highlighting the spatial separation of glycolysis and gluconeogenesis in the liver lobule. Moreover, this early transcriptomic study found that *Cyp7a1*, involved in bile synthesis, and *Apoc2*, involved in lipid metabolism, were enriched pericentrally. In the pathways responsible for ammonia metabolism *Glul* and solute carriers 1A2 and 1A4 were enriched pericentrally, while *Ass1* was enriched periportally [88]. More recently, Halpern *et al.* have used a combination of *in situ* hybridization (ISH) and scRNA-seq to study zonation in mice at single cell resolution. As an example from this study, the spatial coordination of the bile metabolic pathway was confirmed in single cells, where genes involved in bile synthesis are expressed most pericentrally, followed by downstream genes in consecutive layers [16]. Overall, this comprises an in depth characterization of the single cell gene expression in nine zones along the mouse hepatic lobule, producing a comprehensible reference gene annotation for future studies investigating zonation [16]. Building the bridge from transcriptomic to proteomic information, a study by Berndt *et al.*, applying quantitative shotgun proteomics to sorted populations of pericentral and periportal hepatocytes, has revealed how protein expression behaves between these spatially separated populations [89]. While for both populations the uptake of free fatty acids (FFA) depends on plasma glucose concentration, periportal hepatocytes have a higher capacity to produce FFA. Moreover, ammonia metabolism also differs significantly on the protein level between pericentral and periportal hepatocytes. Periportal hepatocytes have a higher uptake of ammonia, which they administer by irreversible fixation in urea, whereas pericentral hepatocytes involve into glutamine synthesis [89]. Exploring the cellular heterogeneity of liver in humans, MacParland *et al.* have performed scRNA-seq and identified clusters of hepatocytes featuring gene expression profiles that were comparable to the zonation profiles described in mouse [18]. Moreover, Aizarani *et al.* established an atlas of the human liver cell types, in which they performed a thorough comparison of the single cell gene expression profiles between mouse and human in the light of zonation

[19]. These data sets serve as a reference for gene expression profiles along the zonation gradient.

1.3.2 Inferring spatial information from gene expression

The anatomy of zonation suggests grouping the cells into a periportal, a pericentral, and one to several groups between both. However, the gradient nature renders it difficult to define a discrete number of zones. Hence, up to 35 zones along the zonation gradient have been reported in scRNA-seq studies [16][19]. Several strategies have emerged to identify transcriptional areas and assign cells to them based on the single cell gene expression profile. Traditional approaches feature the incorporation of external information. For example, in the developing zebrafish embryo, Satija *et al.* have used *in situ* hybridization (ISH) RNA patterns to identify "landmark genes". These genes are then used to transfer spatial annotations onto single cells by leveraging information across co-expressed and co-regulated genes to impute missing values of landmark genes [77]. Adapting this approach to the mouse liver, Halpern *et al.* have used single-molecule ISH (smISH) to determine landmark genes for nine zones between CV and PV. After defining the relevant genes for each zone, they performed scRNA-seq, assigned cells to the respective nine zones based on their gene expression profile, and established a reference of marker gene expression in these zones [16].

Nowadays, computational approaches exist to explore the relationship between single cells based on their gene expression, independent of external information. Initially designed to infer cellular differentiation processes, e.g. as observed in hematopoiesis, the calculation of "diffusion pseudotime" (dpt) has helped to better understand dynamic changes between cell states that cannot be disentangled by clustering [90]. The idea behind this approach is that, despite scRNA-seq methods only providing a snapshot of gene expression in a single cell at a given time, the process of sampling many cells at once will yield cells in different states [90]. Harvesting cells that are at different stages along a developmental or differentiation process allows to order them based on similarities in their expression profiles [90][91]. In the context of liver zonation, this approach can be used to infer changes in gene expression along the trajectory of zonation, hence constructing a pseudospacial relationship between cells. For the human liver, Aizarani *et al.* applied dpt in this context and annotated 35 zones from CV to PV that were used to compare zonation markers between mouse and human

[19]. This thesis also makes use of dpt to study how the transcriptional profiles of diploid and tetraploid hepatocytes are related to the trajectory of zonation.

1.4 Impact of hepatic steatosis on cellular heterogeneity

Zonation might be affected by several environmental factors. For instance, the metabolic profile that hepatocytes exhibit in response to intracellular lipid accumulation is dependent on their spatial location [7]. During the span of life, hepatocytes can naturally accumulate intracellular lipids, known as hepatic steatosis [92]. Hepatic steatosis is a hallmark of non-alcoholic fatty liver disease (NAFLD) and increases over the lifetime [92]. Additionally, an increased incidence of NAFLD has been reported for the elderly [92]. Nevertheless, in relation to diet or genetic factors, NAFLD can occur already in younger individuals and currently affects 25 % of the world-wide human population [8]. During its progression, steatosis leads to inflammation and downstream fibrosis, stating the phenotype of irreversible non-alcoholic steatohepatitis (NASH). In 20 % of the cases, NASH progresses to cirrhosis and can potentially lead to the development of hepatocellular carcinoma (HCC) [93]. Studies using *in vivo* models fed with a high-fat diet have shown that NAFLD affects the cell type composition of the liver [27][7][26]. For example, NAFLD has been observed to lead to higher proportions of immune cells in the liver and simultaneous up-regulation of immune-related gene expression within endothelial and, to a lower extent, Kupffer cells [26]. In hepatocytes, accumulation of fat alters their metabolic capacity as can be observed through changes in gene expression *in vivo*. Similar as in non-parenchymal cells, also hepatocytes up-regulate inflammatory pathways and genes generally associated to inflammation, such as chemokines [27][26].

While *in vivo* models offer a great system to study the impact of intracellular lipid accumulation on all cell types present in the liver, genetic differences between animals and different choices for dietary NAFLD-induction can limit reproducibility [94]. Furthermore, inter-species differences can restrict to which extent findings from mouse and rat studies are applicable to humans [19]. For these reasons, primary human hepatocytes (PHH) are an established model most closely resembling the human liver [95][94]. The culture of PHH additionally allows to study the impact of different conditions on their metabolic profile [95][94]. For instance, hepatic steatosis can be modelled in this *in vitro* system by incubating the cells with free fatty acids (FFA)[96][94]. The common way to do so is incubating the cells with a mixture of oleatic and palmitic fatty acids at a ratio of 2:1 as this resembles hepatic steatosis *in vivo* without inducing toxicity [96]. This approach is also used in this thesis to dissect the impact of intracellular lipid accumulation on the metabolic heterogeneity of PHH in the absence of zonation (Chapter 4). A drawback of the PHH culture system is that the characteristic expression profiles associated to zonation are lost by taking the cells out of their context of the liver lobule[97]. Nevertheless, PHH retain their drug-metabolic profile *in vitro* [96][98]. As they are considered to most closely resemble *in vivo* human liver, the culture of PHH *in vitro* therefore represents the "gold standard" model for assessing drug safety and efficacy during pre-clinical trials [96][98][97]. .

1.5 Drug metabolism

A major role of hepatocytes *in vivo* is the metabolism and detoxification of exogenous compounds. Drug metabolism is a process consisting of three phases, that are regulated by key hepatic transcription factors, such as *HNF4α, SRC1*, and *PGC1α* [99][100][101]. During the first phase (phase I), members of the cytochrome P450 (CYP450) super-family of monooxygenases catalyze oxidation, reduction, hydrolysis, and cyclization reactions in order to increase the compounds' electrophilicity [102]. The *Human Genome Project* currently lists 57 members of the P450 super-family of enzymes [103][104]. Out of these 57 isoforms, the five members CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 are responsible for the phase I metabolism of around 70-80% of all nowadays available drug on the market [12][105]. Enzymes performing phase II of drug metabolism include glutathione-S-transferases (GST), sulfotransferases (SULT), uridine diphosphate glucuronyltransferases (UGT), and arylamine N-acetyltransferases (NAT). During phase II, these transferases add moieties to the metabolites from phase I to hydrophilize the compounds, and thereby facilitate their cellular excretion [106]. Subsequently, in phase III, transmembrane transporter proteins export the conjugated compounds from the cell [107]. For example, caffeine is subject to N-3 demethylation by CYP1A2, followed by sulfamethazine N-acetylation by phase II enzyme NAT2, resulting in paraxanthine that is subsequently excreted from the body through urine [108][109]. Generally *in vivo*, xenobiotic metabolism is mainly taking place in the pericentral region, where key phase I and phase II enzymes have been found to be up-regulated on both the transcriptomic and the proteomic level. These include, for instance, CYP1A2, CYP2E1, GSTA3, and SULT1B1, whereas GSTA2 and CYP2F2 are expressed periportally [88][16][89]. While many pharmacological studies focus on the overall metabolic capacity of liver to detoxify a xenobiotic compound, a recent single cell transcriptomics study has found that hepatocytes *in vivo* respond in a location-specific manner to the hepatotoxicant 2,3,7,8-tetrachlorodibenzo-ρ-dioxin (TCDD). Essentially, pericentral hepatocytes were shown to up-regulate nuclear receptors relevant in bile and lipid metabolism, whereas periportal hepatocytes up-regulated amino acid metabolism upon TCDD treatment [28]. Given these findings, it can be argued that the cellular context impacts the toxicity and efficacy of a given drug *in vivo* and the cellular heterogeneity within the tissue has to be considered.

However, the 2D culture of PHH represents the gold standard to assess drug efficacy and toxicity during early phases of pre-clinical trials of drug development [95][97]. Despite the loss of zonation-dependent gene expression in this standard *in vitro* culture of PHH, the cells have been shown to retain their drug-related metabolism [95]. Namely, PHH express the enzymes related to the three phases of drug metabolism similarly to cells *in vivo* and respond to substances inducing the expression of these enzymes [95][97]. PHH therefore represent a suitable model for the study of drug efficacy, toxicity and drug-drug interactions [110]. For instance, phase I cytochrome induction or inhibition is an indicator for the safety of a drug in clinical trials [111]. The expression of phase I enzymes can be induced in PHH by the presence of their substrates and measured at the protein and the transcriptome level [112]. These substrates can thereby be used to assess enzymatic activity, which indirectly serves as an indicator of the metabolic capacity of PHH [110][113][112]. Herein, simultaneous targeting of several cytochromes by a mixture of their respective substrates is formally known as the "phenotyping cocktail approach" [110][113][112]. In this thesis, the Sanofi-Aventis cocktail is used to measure drug-related metabolic capacity in PHH [110]. This phenotyping cocktail consists of substrates of the five cytochromes CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4, involved in the clearance of 70-80 % of available drugs in the market [105]. For the substrates in this cocktail, no drug-drug interactions have been reported [110]. Together with the model of hepatic steatosis, this allows to investigate in this thesis to which extent cellular heterogeneity *in vitro*, either reminiscent in culture or gained through hepatic steatosis, impacts the drug-related metabolic capacity of PHH (Chapter 4).

1.6 The aim of this thesis

The general aim of this thesis is to explore cellular heterogeneity of hepatocytes *in vivo* and *in vitro* through the analysis of single cell transcriptomic data. The first chapter hence addresses the principles of analysis for this data type. In the liver *in vivo*, intrinsic factors, such as polyploidization and zonation, contribute to the heterogeneity in the tissue and are so far poorly understood at single cell resolution. These factors have been shown to influence the cellular responses to external stimuli, such as the accumulation of lipids leading to NAFLD, or the admission of drugs. In relation to this, the third chapter in this thesis establishes a reference of healthy cellular heterogeneity in the young adult mouse liver under physiological conditions. The focus specifically lies on investigating the impact of polyploid hepatocytes on liver functionality, and the relationship between ploidy and zonation in mouse. As described above, differences in size and genomic content between nuclei of differing ploidy status present a challenge in distinguishing true biological variation from noise and effects generated by the expected differences in mRNA content. Therefore, a new normalization technique is developed in this project, using spike-in read counts to minimize the effect of library size differences between diploid and tetraploid hepatocytes. Furthermore, the co-expression levels of stem cell marker genes are calculated to explore the regenerative potential of tetraploid hepatocytes. Furthermore, using reference marker genes and the diffusion pseudotime algorithm allows to order nuclei according to their inferred spatial context based on their gene expression profile.

For this project to study polyploidy and zonation, mice offer a commonly accepted and suitable *in vivo* model system, from which results have been shown to be applicable in humans [16]. However, the species differences between humans and mice limits their usability as a model in other cases, such as exploring drug effects and NAFLD [94]. Herein, hepatic steatosis can be modeled in a more reproducible fashion using cellular *in vitro* models [95][98][94]. Additionally, in the context of the discovery and development of new drugs, toxicity and efficacy are usually tested in culture systems of primary human hepatocytes [95][96][98]. For these reasons, primary human hepatocytes are used as a model in the fourth chapter of this thesis to study the impact of intracellular lipid accumulation on cellular heterogeneity and their drug-metabolic capacity. Primary human hepatocytes have been shown to retain their drug-metabolic gene expression profile in culture [98][95]. However, they are known to lose their characteristic hepatocyte-like expression profile over the time in culture [114][115]. A challenge in the data analysis is therefore to distinguish true loss of expression from technical effects resulting in smaller library sizes to correctly identify cells losing their expression. Moreover, the 2D culture of primary human hepatocytes *in vitro* is characterized by the absence of tissue-specific zonation patterns [116]. Therefore, the major aim of this project is to explore the reminiscent cellular heterogeneity *in vitro* in response to external stimuli. Data integration of *in vivo* data sets is used to perform comparisons of cellular heterogeneity between *in vivo* and *in vitro*. After identifying metabolic profiles that are present in both, *in vivo*, and *in vitro*, their response to metabolic challenges are tested. As such, this thesis investigates how intracellular lipid accumulation affects the gene expression profile and transcriptional variability between individual cells. Moreover, the transcriptomic responses to a five drug cocktail are assessed to explore to what extent individual hepatocytes deviate from each other in their drug-metabolic response, normally and under hepatic steatosis.

2 A brief guide to scRNA-seq analysis

This chapter contains a comprehensive description of the methods used for analyzing single cell RNA sequencing data from two different research projects. The first project, hereafter Ploidy&Zonation, aims to dissect the impact of ploidy status on metabolic functionality and its relationship to zonation in the young adult mouse liver. In the second project, hereafter PHH diversity, the role of cellular heterogeneity is studied in a seemingly homogeneous population of cells in response to environmental challenges.
Different methodologies were used to address the biological questions for each of the two projects. Nevertheless, both projects share similar computational challenges, such as pre-processing, quality control, normalization, and batch integration. Therefore, this chapter covers the basic steps involved to proceed from raw transcriptomic read counts towards understandable output and explains how choices along this processing were made to match the respective biological questions.

In this chapter, the term *experiment* refers to all sequencing data generated for a given project. The donors used for both projects (mice in the Ploidy&Zonation project; humans in the PHH diversity project) are referred to as (biological) *replicates*. In case an experiment is comprised of different sequencing libraries, these libraries are referred to as *batches*.

2.1 Read alignment

2.1.1 Plate-based single nucleus RNA-sequencing data

During data generation for the Ploidy&Zonation project, single nuclei were sorted into wells of 384-well plates. A modified SMART-seq2 protocol was used for library preparation, resulting in the recovery of full-length transcripts [117]. This approach is referred to as single nucleus RNA-seq version 2 (snRNA-seq2). The first step in the analysis of any single cell or single nucleus RNA-seq (s*RNA-seq) data is the alignment of reads to a reference genome. This is done to bring the reads into meaningful genomic context. For the Ploidy&Zonation project, ERCC spike-in molecules had been added along the nuclear extracts into the wells of 384-well plates [81]. Differences had been observed for the recovery of ERCC-reads when diluting ERCCs from different lots the same way[117]. Therefore, to generate libraries with overall similar proportions of ERCC-reads, ERCCs had been diluted more if preliminary results otherwise showed untypically high amounts of spike-ins [82][117]. To count both, endogenous and ERCC reads, the reads sequenced in this data set were aligned to a combined reference genome of mm10 and ERCC92 using STAR [118]. STAR is a broadly used alignment software for RNA-seq data that has been shown to achieve high mapping accuracy and robust handling of splice junctions while requiring a low amount of computational resources in comparison to other mappers [119]. In this project, read alignment resulted in an average of between 213,377 and 728,614 aligned reads per nucleus for the individual plates. Moreover, the proportion of ERCC reads among the total number of reads ranged from 33.4 % to 79.9 % (in plate SNI-116(R2) and plate SNI-194, respectively). This can partially be explained by the use of different ERCC dilutions (ERCCs were diluted 1:300,000 in SNI-116(R2), and 1:100,000 in SNI-194). Overall, the average ERCC proportion was 41.4 % in plates with 1:300,000 dilution, and 66.3 % in plates with 1:100,000 dilution. However, among all plates for which ERCCs were diluted 1:100,000, the share of ERCC reads still ranged from 45.9 % to 79.9 % as shown in table 2.1.

Table 2.1: Dilution of ERCCs used in each plate together with recovered proportion of ERCC reads among all uniquely mapped reads

Plate	ERCC dilution	Percentage ERCC reads
SNI-116(R2)	1 in 300,000	33.4 %
SNI-160(R2)	1 in 100,000	45.9 %
SNI-192-p1	1 in 100,000	77.2 %
SNI-192-p2	1 in 100,000	74.0 %
SNI-193	1 in 100,000	73.9 %
SNI-194	1 in 100,000	79.9 %
SNI-234(R2)	1 in 100,000	47.1 %
SNI-235(R2)	1 in 300,000	34.8 %
SNI-626	1 in 300,000	51.6 %
SNI-634	1 in 300,000	46.3 %
SNI-635	1 in 300,000	41.0 %

In general, after alignment to a reference genome, count matrices are constructed by putting the reads into genomic context. For scRNA-seq experiments, where the library is composed of mRNA molecules from both, nucleus and cytoplasm, the majority of transcripts only contain exons. Therefore, it is a standard procedure to count reads falling into exonic regions [120]. However, in the Ploidy&Zonation project presented here, single nuclei were isolated, for which the proportion of unspliced transcripts is higher. Moreover, the SMART-seq2 technology used to generate the Ploidy&Zonation data allows the recovery of full-length transcripts. The software and reference genome to build the count matrix for this data set was therefore chosen with regards to the expected reads stemming from intronic regions. In brief, to retain and use all information, *htseq-count* was used to count all reads mapping to transcripts in a single nucleus, thereby considering both, exonic and intronic reads. The reads associated to transcripts were then aggregated into genes (Methods).

2.1.2 Droplet-based scRNA-seq data

For the PHH diversity project a droplet-based approach developed by *10X Genomics* was used to generate scRNA-seq data in two batches. To optimize costs while obtaining a high yield of information, the first batch was aimed to retrieve a high sequencing depths whereas more cells were sequenced in the second batch. In this project, whole single cells were isolated. Furthermore, in the *10X Genomics* approach the first 91bp at 3' end of a given transcript get sequenced. Therefore, the common practice of counting the reads falling into exons is suitable for this type of data[121]. The company providing the library preparation kit also provides software to directly build count matrices for a given experiment, named *cellranger count* [121]. This software includes an adapted version of the STAR alignment method that can align *10X Genomics* scRNA-seq data. Hence, the scRNA-seq data from the PHH diversity project was aligned to the human genome version GRCh38 using *cellranger count.* After mapping to the reference genome, the average number of raw reads per single cell differed between 38,575 for the first batch, and 7,045 for the second batch.

2.2 Quality control

The number of raw reads per cell that map to the reference genome gives a measure of library size homogeneity between cells. This can be of particular interest when comparing cells from different batches. However, that number provides no context how informative these reads are, i.e. how many different transcripts or genes were detected. In principle, methods achieving greater sequencing depths have a higher probability to capture lowly expressed genes and therefore should recover an overall greater number of genes [56]. Thus, the total number of detected genes in a s*RNA-seq data set serves as a measure for the quality of the data set.

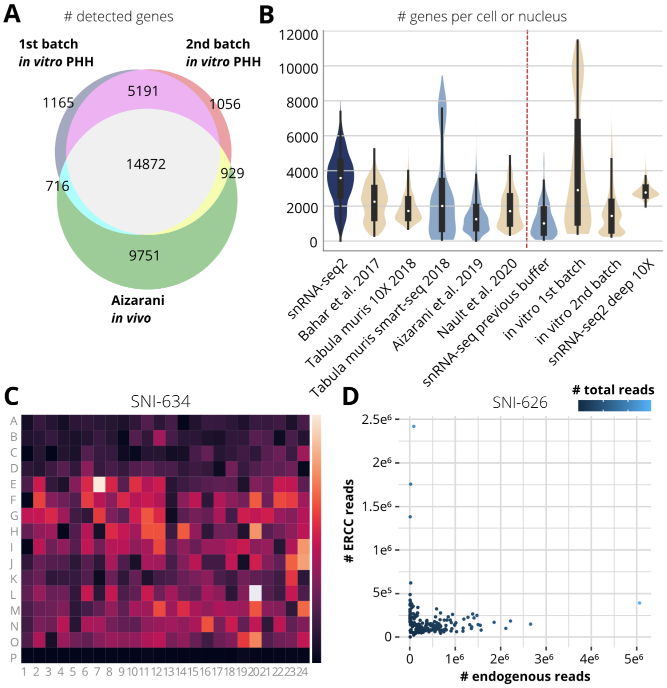

Figure 2.1: A Total number of genes detected in three different scRNA-seq experiments. **B** Number of expressed genes per cell or nucleus in different data sets. **C** Number of transcript counts in a plate of snRNA-seq2. **D** Scatter plot depicting the number of ERCC reads against the number of endogenous reads colored by the total mapped reads for a plate of the snRNA-seq2 data set.

Different experimental approaches show different sensitivity to detect genes. While plate-based approaches usually aim at higher sequencing depths with low cellular throughput, droplet-based approaches, such as from 10X Genomics, yield high cellular throughput but at a usually more shallow sequencing depth [57][56]. Moreover, even when using the same experimental approach, the total number of detected genes can differ substantially between s*RNA-seq data sets, depending on differences in the chemistry used or when experiments are performed in different laboratories. For example, figure 2.1 A depicts the total number of detected genes in a Venn diagram between three scRNA-seq data sets all performed on human liver cells [19][122]. While 14,972 genes were found to be commonly detected in all three data sets, between 1,056 and 9,751 genes were found individually in only one of the data sets. In line with plate-based approaches yielding higher coverage, the number of detected genes was highest in the roughly 10,000 cells from the Aizarani *in vivo* data set [19]. The two batches from the PHH diversity project were comprised of a total of roughly 60,000 cells and had a higher share of detected genes between each other than to the *in vivo* data set [19]. This can also be due to the *in vivo* data set consisting of different cell types, hence yielding a broader variety of genes.

The difference between methodologies becomes even more evident when transferring the information on gene detection back to the single cell level. For example, the number of genes detected in each single cell or nucleus is depicted in figure 2.1 B for different s*RNA-seq data sets done on liver figure. This allows the comparison of plate-based and droplet-based s*RNA-seq data sets that were generated wither within the same laboratory or in different international laboratories. On average, the snRNA-seq2 data set had the highest number of genes detected per nucleus. However, both, the first batch from the PHH diversity project, and the plate-based data set from the *Tabula muris* consortium, featured a fraction of cells with a higher number of genes (Figure 2.1 B) [37].

The technical heterogeneity within a singular experiment can be seen in figure 2.1 C, where the number of endogenous transcripts recovered per single nucleus is shown based on which well of a 384-well plate the nucleus was in. In the depicted plate SNI-634 from the Ploidy&Zonation snRNA-seq2 data set, lower transcript counts are observed in the wells located at the edges of the plate (Figure 2.1 C). Previous studies have shown that the outer rows of microtiter plates can experience poorer library quality in comparison to wells at the center of the plate due to higher evaporation rates at the edges, which could explain this pattern [123]. Moreover, more than 220,000 endogenous transcript reads were detected for two nuclei on this plate while the plate average was at 59,089 endogenous transcript reads per nucleus.

Overall, these results indicate that data quality and metrics for suitable filtering procedures depend on the sequencing approach. Moreover, library size fluctuations within the same plate and between plates showcase the need for appropriate normalization techniques.

2.3 Normalization

As described in the introduction of this thesis, the goal of normalization is for the normalized counts to represent true gene expression that is not influenced by extrinsic, technical factors. The addition of ERCC spike-ins offers another way to adjust read counts for technical variation [81]. Assuming that ERCC spike-ins are affected by technical variation the same way as endogenous transcripts, adding the same amount of ERCCs to each single cell can help to distinguish this purely technical variation from biological variation [82].

2.3.1 Adjusting for polyploidy using spike-ins

Amplification bias has been observed leading to differences in amplification rates between ERCC molecules and endogenous transcripts, violating the assumption that both are subject to the same technical variation [83][72][84]. Especially for cells expressing low amounts of endogenous transcripts, this can lead to the proportion of ERCC reads being artificially inflated [83]. The Ploidy&Zonation project is comprised of diploid and tetraploid nuclei, with higher transcript counts being expected in tetraploid cells based on the differences in nuclear size [67][66]. Therefore, an artificial inflation of spike-in reads in the diploid fraction can bias differential expression analysis between the populations of interest. The occurrence of this bias is exemplified on plate SNI-626, where some nuclei experienced a large number of ERCC reads and little amount of endogenous reads, and vice versa (Figure 2.1 D). Thus, the ratio of ERCC and endogenous reads per nucleus has to be taken into account for correct normalization in this data set and a new normalization technique was developed to do so. Briefly, nuclei with a proportionally high fraction of ERCC to endogenous reads were divided by a smaller number than nuclei with a low fraction of ERCC to endogenous reads (Methods). This was done to avoid differences in gene expression between diploid and tetraploid nuclei to be purely driven by their respective amounts of endogenous transcripts. This normalization technique will hereafter be referred to as ERCC ratio normalization (ERN).

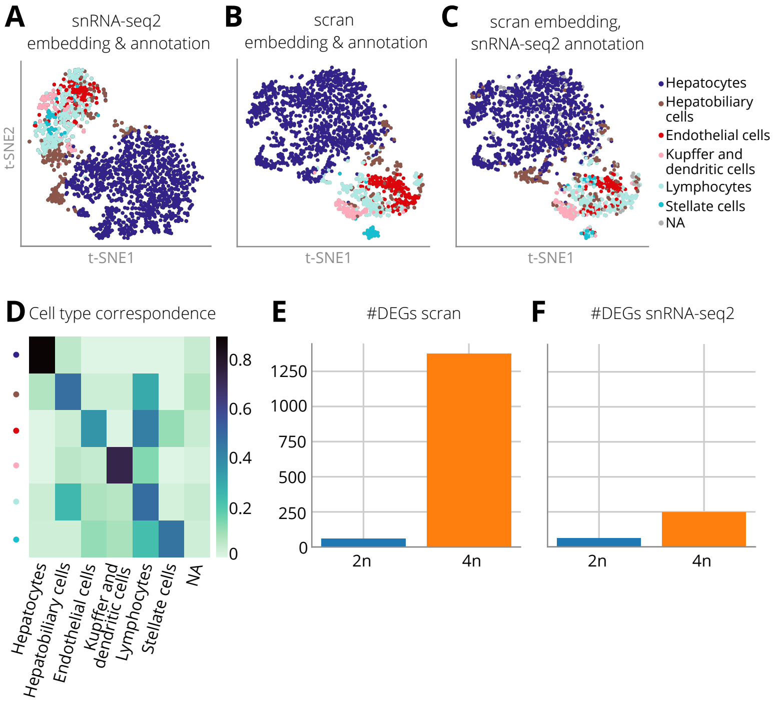

Figure 2.2: A t-SNE of the snRNA-seq2 data normalized by ERN. **B** t-SNE of the snRNA-seq2 data normalized using scran. **C** t-SNE of the snRNA-seq2 data normalized using scran but colored by the cell type annotation established after applying ERN. **D** Heatmap depicting cell type annotation correspondence between the different normalization techniques. **E** Barplot depicting the number of DEGs in 2n and 4n hepatocytes after scran normalization. **F** Barplot depicting the number of DEGs after snRNA-seq2 normalization.

Comparison to *scran*

Normalization techniques have previously been bench-marked by Vieth *et al.* with the method *scran* outperforming other techniques in terms of recovering true differential expression between conditions of interest [124]. Thus, to test the performance of ERN in comparison to a commonly used normalization protocol, *scran* was applied to the Ploidy&Zonation snRNA-seq2 data set. After applying both normalization approaches (ERN and *scran*) independently to the data set, nuclei were clustered by *Louvain* clustering and embedded in a t-SNE for visualization. For each cluster, marker gene expression was investigated to correctly annotate cell types present in the tissue (Methods). When using ERN, the non-parenchymal cells showed generally poor separation on the embedding (Figure 2.2 A). However, *Louvain* clustering led to the identification of all relevant non-parenchymal cell types (Figure 2.2 A). In comparison, *scran* normalization revealed a visually more evident separation of non-parenchymal cell types on the t-SNE embedding (Figure 2.2 B). Between the two methods, the cell type annotation of each nucleus showed an overall great correspondence. This can be seen by visualizing the cell type annotation achieved after ERN on top of the *scran* embedding (Figure 2.2 C). These results were further quantified by calculating the percentage of nuclei assigned to a specific cell type between the two normalization techniques (Figure 2.2 D). For Hepatocytes, and Kupffer- and dendritic cells, more than 80 % of nuclei were assigned to the same cell type between the two normalization approaches. The lowest correspondence was found for Lymphocytes and Endothelial cells, because many of the nuclei annotated as lymphocytes after ERN were labeled as endothelial cells after *scran* normalization (Figure 2.2 D). Lymphocytes have been observed to be subject to dissociation-related changes in gene expression making them prone to feature an overall lower level of mRNA expression than other cell types, and harder to capture through snRNA-seq [125]. Therefore, the differences between normalization techniques could potentially be explained by *scran* aiming to preserve differences in mRNA levels between cell types whereas ERN tries to minimize them.

The goal of this Ploidy&Zonation project is to identify defining gene expression characteristics between diploid (2n) and the tetraploid (4n) hepatocytes that are independent of their respective mRNA content. As shown in figure 2.2 E, using *scran* as a normalization technique led to the detection of more than 1,250 up-regulated genes in 4n hepatocytes. This number of up-regulated genes is potentially confounded by the greater amount of mRNA molecules occurring in larger nuclei [67]. Meanwhile, the specifically designed ERN led to the detection of 241 up-regulated genes in 4n against 2n. Hence, by using ERN for normalization, the differences identified between the two groups of interest are less likely to be purely confounded by nuclear size and better represent true functional differences between the two groups.

2.3.2 Droplet-based scRNA-seq data

The goal of the PHH diversity project was to identify functionally heterogeneous groups of hepatocytes in response to treatment conditions. The data set from this project was comprised of primary human hepatocytes (PHHs) with unknown ploidy status. For data sets in which it is not known whether cells contain the same mRNA levels, *scran* has been shown to produce robust results under different scenarios [124]. Furthermore, *scran* has been reported to outperform simple library-size normalization techniques by better preserving biological variance while removing technical variations [124]. Due to the ploidy status being unknown and no ERCCs being present, *scran* was therefore used for normalization in this project (Methods).

2.4 Batch correction

As shown above, normalization aims to remove technical variation between cells or nuclei within an experiment to only retain true biological differences between the populations of interest. Apart from the technical differences within one experiment, data sets can be comprised of several experiments resulting in batch effects. In general, batch effects can stem from different laboratories or experimentalists performing the experiments, e.g. by imposing variations in sample acquisition or handling, used reagents and protocols [126]. Even when experiments are performed in the same laboratory by the same person, batch affects can arise from different flow cells or sequencing lanes being used, or in the case of plate-based approaches, the respective plates [126]. Additionally, biological factors also contribute to batch affects. These include the individual or animal a sample was taken from, the spatio-temporal context of sample acquisition, and stochastic differences in cell type composition [126].

2.4.1 Removing plate-dependent differences

In the Ploidy&Zonation project, batch effects can be observed by plates displaying different number of transcripts, which represents a source of unwanted technical variation that cannot be attenuated by normalization alone (Figure 2.3 A). Applying batch correction methods to adjust for this heterogeneity is therefore crucial [71]. Several batch correction methods have been developed for the integration of s*RNA-seq data sets aiming to remove unwanted variation by i) changing the neighborhood graph or ii) the embedding to bring together similar cells from different batches, or iii) directly adjusting the counts inside the count matrix [126]. For example, *comBat* applies a linear regression to remove coverage differences between plates, leading to similar reads counts between plates (Figure 2.3 B)[127]. Due to its fast and easy usability, *comBat* was used to remove batch effects in the Ploidy&Zonation snRNA-seq2 data set (Methods).

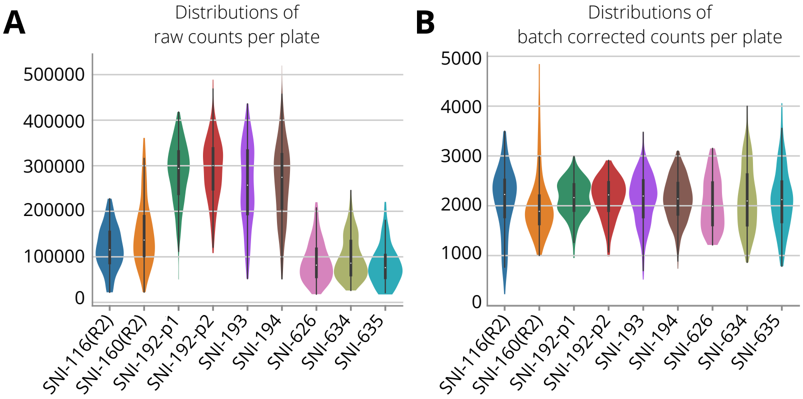

Figure 2.3: **A** Violin plot depicting the number of raw reads per single nucleus for each of the analyzed plates **B** Violin plot depicting the number of batch corrected, normalized reads per single nucleus in each of the analyzed plates.

2.4.2 Donor-specific analysis and batch integration

The PHH diversity data set is composed of single cells from four male human donors, whose clinical characteristics are summarized in table 2.2. As described above, the data set was sequenced in two batches, each containing cells from two of the donors. While the first batch (donors HUM180812 (1) and HUM4152 (2)) yielded an average of 38,575 reads per cell, the second batch (donors HUM181641 (3) and HUM4190 (4)) yielded an average of 7,045 reads per cell. Because of this technical variability between the batches, initial analysis was done separately per donor. The goal of this analysis was to identify biological variation that was present in all four donors. In particular, this was done because batch integration methods can potentially over-correct the data [126]. Therefore, analyzing the donors individually helps to obtain an overview of biological variation across donors before applying batch integration approaches. As depicted in figure 2.4 A-D, biological variation was present within the primary human hepatocytes (PHHs) in all four donors, independently of the cells' treatment conditions. In particular, every donor showed at least four clusters of hepatocytes (Figure 2.4 A-D). To compare the identified clusters between donors, the top 1,000 differentially expressed genes were obtained for each cluster in every donor, and their overlaps between donors were calculated. This led to the detection of three transcriptionally similar clusters that were present in all four donors, hereafter named shared cluster 1, 2, and 3 (Figure 2.4 E, Methods).

Table 2.2: Selected clinical characteristics of the four male human donors used in the PHH diversity project

Donor ID	Age	Drug/Tobacco/Alc. use	Ethnicity
HUM180812 (1)	57	No/No/Social	Hispanic
HUM4152 (2)	18	Yes/No/No	Caucasian
HUM181641 (3)	56	Yes/No/Social	Caucasian
HUM4190 (4)	26	No/Yes/No	Caucasian

However, some clusters were only detected in one of the four donors (e.g. cluster5 in donor1 and cluster1 in donor3, hereafter named "icluster1" and "icluster2", respectively). Additionally, cluster2 and cluster3 in donor1 showed similarity to cluster2 in donor2. These clusters were therefore grouped into shared cluster4, which was not observed in the donors of the second batch (Figure 2.4 A, C, and E). The detection of shared clusters between donors was used as basis for the joint analysis of all four donors together (Methods). After performing joint filtering and library size normalization, the cells from shared cluster3 already showed relative proximity on a UMAP embedding before batch correction was applied (Figure 2.5 A). Moreover, cells from the first batch (donors 1 and 2) were embedded more closely to each other between the two donors than cells from the second batch (donors 3 and 4), in which only a group of cells from donor3 appeared in proximity to the cells from donor4 on the UMAP (Figure 2.5 B). To choose the most suitable option for overcoming the difference in sequencing depth between the two batches, batch integration methods were applied and compared.

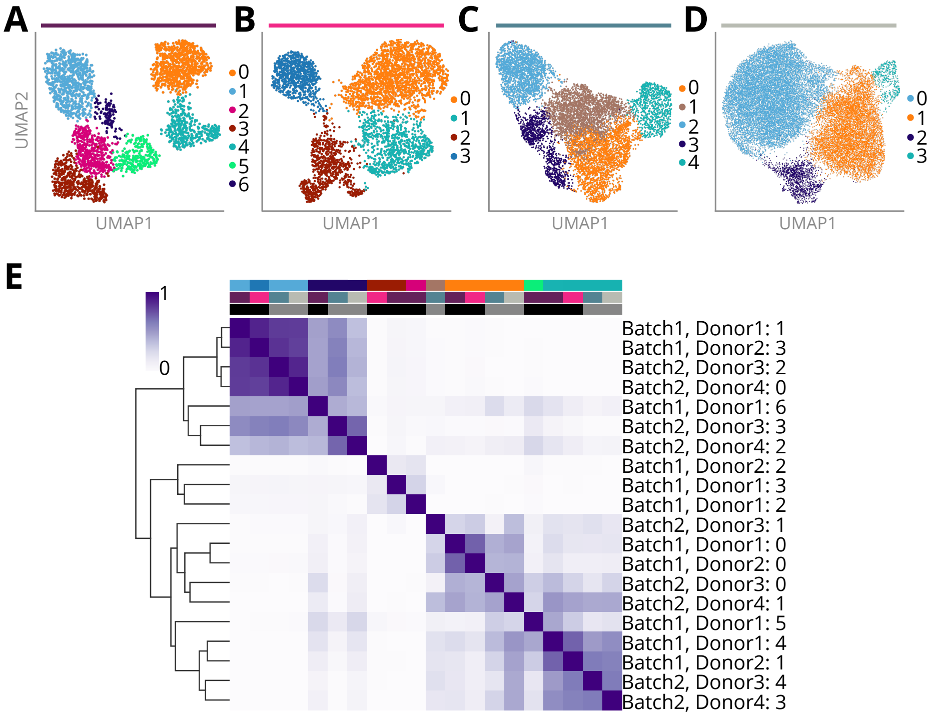

Figure 2.4: A-D UMAPs showing *Louvain* clustering for the four human donors from the *in vitro* scRNA-seq data on PHH diversity **E** Heatmap depicting the percentage of overlap for the top 500 genes per *Louvain* cluster between donors. Top row: *Louvain* clusters, second row: individual donors, third row: batch; tree showing hierarchical clustering based on similarity between clusters.

Data integration using Harmony

Harmony is an unsupervised integration method aiming to create a joint embedding for integrated data sets [128]. This joint embedding puts cells from different batches into the same groups based on their gene expression [128]. As such, Harmony has been shown to achieve high data integration while preserving biological relationships between the cells [126]. Applying this method to the PHH diversity data set brought together those clusters that had previously been identified to be shared between donors (Figure 2.6 A). Therefore, *Louvain* clustering on the *Harmony* embedding was used to finalise the annotation of the clusters that were unique per donor, resulting in the identification of four PHH subgroups that are based on the shared clusters. For instance, cells from "icluster2" (specific to donor3) were separated into different clusters on the integrated embedding, and thus assigned to subgroup I, and II, respectively. Overall, cells from donor1 and donor2 were integrating well with both, donor3 and donor4, whereas cells from donor3 and donor4 were more separated on the UMAP (Figure 2.6 B). Subgroup III was the only group represented by only one *Louvain* cluster and had the most equal proportions of all four donors (Figure 2.6 A-D).

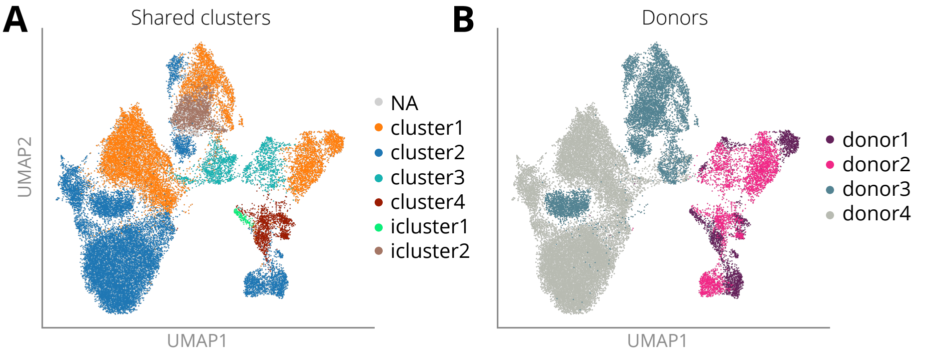

Figure 2.5: UMAP of the unintegrated PHH diversity scRNA-seq data comprised of four human donors sequenced in two batches, colored by **A** shared clusters between donors and **B** donors.

Additionally, some cells from the second batch were initially assigned to shared cluster2 based on the individual donor analysis. However, after integration, these cells were clustering with cells from shared cluster4 that had only been identified in donors from the first batch. As later downstream analysis showed, these subgroup IV cells were losing their hepatocyte-characteristic expression along culture time [114][115]. Their identification in donor1 and donor2 is likely due to the higher sequencing depth achieved in the first batch, enabling a better cluster separation [47][48]. Furthermore, cells from "icluster1" (specific to donor1) were also annotated as subgroup IV after integration. Despite the overarching biological similarity of the subgroups, donor-specificity was still present within some of the identified subgroups. For example within subgroup II, *Louvain* clusters 8 and 11 were mainly composed of cells from donor3, while the majority of cells in clusters 0, 3 and 10 were stemming from donor4 (Figure 2.6 A-C). Nevertheless, in summary, this unsupervised integration approach and the subsequent *Louvain* clustering allowed the annotation of PHHs into four major subgroups that were present in all four human donors (Figure 2.6 D).

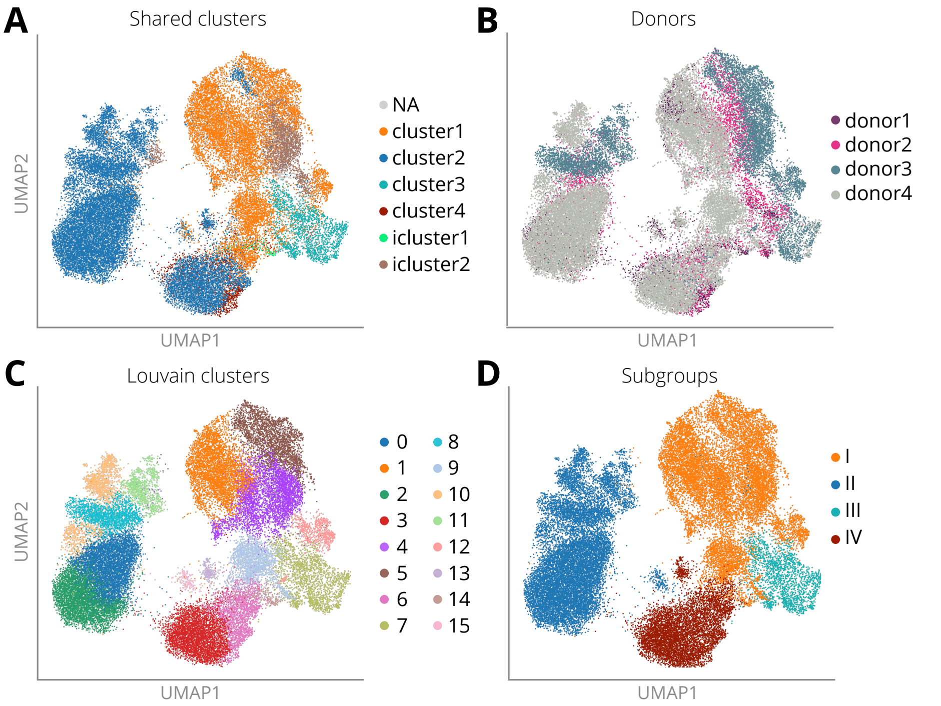

Figure 2.6: UMAP of the cells from the PHH diversity project after batch integration was performed using *Harmony*, colored by **A** shared clusters, **B** human donors, **C** *Louvain* clusters, and **D** characterized hepatocyte subgroups.

Confirmation through scGen

As a guided integration software, *scGen* uses a variable autoencoder (VAE) based on the low-dimensional representation of the cells to predict the impact of a batch on gene expression [129]. To work effectively, this method relies on the annotation of cell types prior to integration [129].

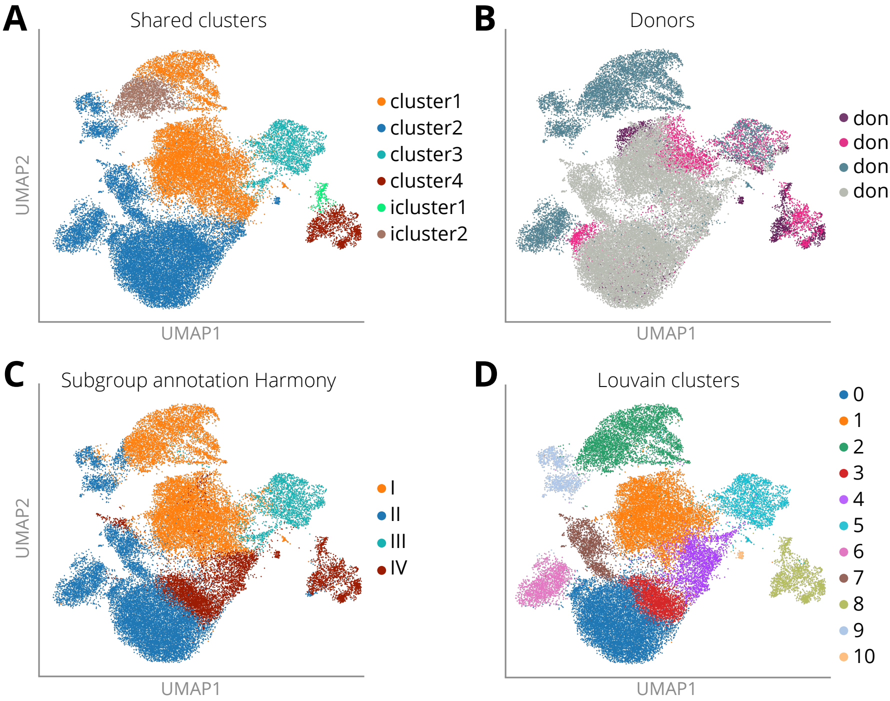

Figure 2.7: UMAP of the cells from the PHH diversity project after batch integration was performed using *scGen*, colored by **A** shared clusters, **B** human donors, **C** subgroup annotation inferred from the *Harmony*-integrated data, and **D** *Louvain* clusters on the *scGen*-integrated data.

Hence, to test this integration method, the previously identified shared clusters between donors were used as pre-annotated cell types. While cells from shared cluster2 are brought together better than when using *Harmony*, a group of cells from shared cluster3 stemming from donor4 were separated from the rest of cells in that group (Figure 2.7 A and B). Overall, a higher degree of separation was observed between donors for *scGen* in comparison to *Harmony* (Figure 2.7 B), especially for donors 1 and 2 that were integrated better through *Harmony*. The subgroup annotation inferred after applying *Harmony* batch correction was put onto the *scGen* embedding to compare the findings to the *Louvain* clustering resulting from *scGen*. Consistent to the results from *Harmony*, cells from "icluster1" were grouping with cells from shared cluster4. However, because *scGen* makes use of the pre-defined cell type annotation gained from the individual donor analysis, this method had a bias to group together cells with same shared cluster annotation. Hence, "icluster1", "icluster2", and shared cluster4 were more separated from the other clusters on the UMAP than when using *Harmony* (Figure 2.7 A). Moreover, the cells from the second batch that were found to belong to subgroup IV after *Harmony* were embedded more closely to cells from shared clusters 1 and 2 than to shared cluster4 (Figure 2.7 A, C, and D). Therefore, using *scGen* could result in these subgroup IV cells getting wrongfully assigned to subgroup I, and II. Overall, it can be concluded that both integration methods led to a consistent identification of four hepatocyte subgroups. Due to the unsupervised nature of *Harmony* leading to a more intuitive subgroup annotation, this was the final method of choice.

2.5 Transcriptional variability

After the identification of cell types or groups of interest in the whole data set, the next step in single cell RNA-seq analysis is often calling differential gene expression between these groups of interest. Differential expression of mRNA has been shown to be usable as approximate measure for protein production [34]. Hence, differences in transcript abundances give information about how groups of cells differ in their functionality [34]. Apart from the total differences in transcript abundance, cells also fluctuate in their stochasticity to express a given gene. This is because mRNA and protein production are stochastic processes resulting in intrinsic variation in their respective yields [130]. The variation in mRNA yield between cells of the same type is often referred to as transcriptional variability, or noise. In s*RNA-seq data, several factors can contribute to transcriptional variability, including technical noise, sampling bias, and biological variability. Technical noise can be introduced during the generation of scRNA-seq data and affects the accuracy of gene expression measurements [54]. Sampling bias can occur when the cells in a sample are not representative of the entire population, resulting in skewed results [131]. Biological variability can be caused by differences in the cellular environment, such as the transcriptional regulators active within a cell, or external factors [131][130]. For instance, cells from elderly people have been shown to increase in background gene expression, i.e. more noisy transcription, which is in turn associated to a less targeted response to a stimulus [132][133]. Hence, measuring this transcriptional variability is a tool to assess the uniformity of gene expression in population of cells.

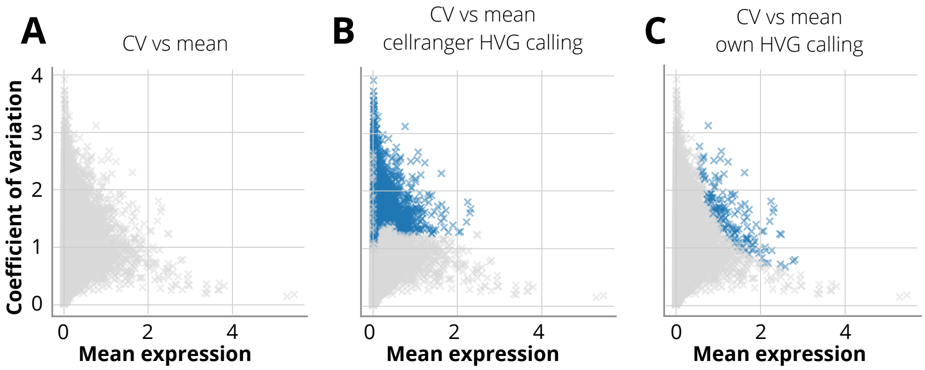

Figure 2.8: A Scatter plot showing the relationship between the traditionally calculated coefficient of variation (CV) and the mean gene expression, colored by **B** highly variable genes (HVG) identified by applying cellranger's method, and **C** HVG identified by applying a cutoff on the CV calculated on log-transformed data.

While traditionally, transcriptional variability is often measured through single molecule fluorescence in situ hybridization (smFISH), single cell sequencing techniques offer to dissect differences on the whole transcriptome level across individual cells [131]. In s*RNA-seq, transcriptional variability is often measured as the *coefficient of variation* (CV), defined as the standard deviation divided by the mean [77]. However, genes with a low mean expression have been shown to have an artificially high CV [134][130]. This effect is exemplified here in figure 2.8 A for the snRNA-seq2 data set of the Ploidy&Zonation project. Several computational approaches exist to adjust for this effect [135][134][136][137]. One of the earliest computational techniques works by computing the distance between the squared CV and a rolling median along expression levels [135]. Later techniques implemented approaches to model the noise in ERCC reads to correct the endogenous counts accordingly [134][136]. A recent approach by

Dominic Grün constructs a neighborhood to identify homogeneous cell states, allowing to disentangle transcriptional noise in a given neighborhood [137]. Based on the approach by Kolodziejczyk *et al.* other modern and commonly used computational tools also apply a rolling median to the linear fit between mean and CV to correctly identify highly variable genes [78][77]. For example, figure 2.8 shows highly variable genes identified by applying the "cellranger" method implemented in *scanpy* revealing that very lowly expressed genes do not get called highly variable despite their high CVs. Logarithmic transformation of the count data decreases the issue of high differences in mean expressions between genes. However, Canchola *et al.* have shown that calculating the traditional CV (standard deviation divided by the mean) on log-transformed count data is not precise because it often results in underestimation of variability [138]. Instead they report a formula specifically adjusted to calculate the CV on log-transformed data (Methods) [138]. Hence, this formula was used to calculate the CV for both data sets in this thesis. In the Ploidy&Zonation project, highly variable genes were identified based on this CV calculation by making use of the ERCCs (Methods). Arguing that variability in the spike-in molecules, that were added in the same quantities to all wells, should only stem from technical sources, the CV was calculated for the ERCC reads. Endogenous genes were therefore called highly variable if they deviated more than one standard deviation from the median observed for the ERCCs (Methods). As depicted in figure 2.8 C, the thereby identified highly variable genes are independent of mean the expression, deeming this an appropriate approach of identifying highly variable genes (Figure 2.8 C).

3 Ploidy in young healthy mouse liver

While the previous chapter contained an overview of common decision points during the analysis of scRNA-seq data, this chapter goes in depth on the downstream analysis of the in vivo snRNA-seq2 data set used for the Ploidy&Zonation project. The focus of this analysis is to elucidate how gene expression differs between diploid hepatocytes and tetraploid hepatocytes. Moreover, the follow-up question addressed in this chapter is how hepatocytes featuring different levels of ploidy relate to liver zonation, i.e. whether the gene expression of both hepatocyte populations can be linked to zonation patterns, indicating that they show spatial preferences. The majority of the results presented in this chapter are published in Richter et al. 2021 [117].

3.1 Method overview

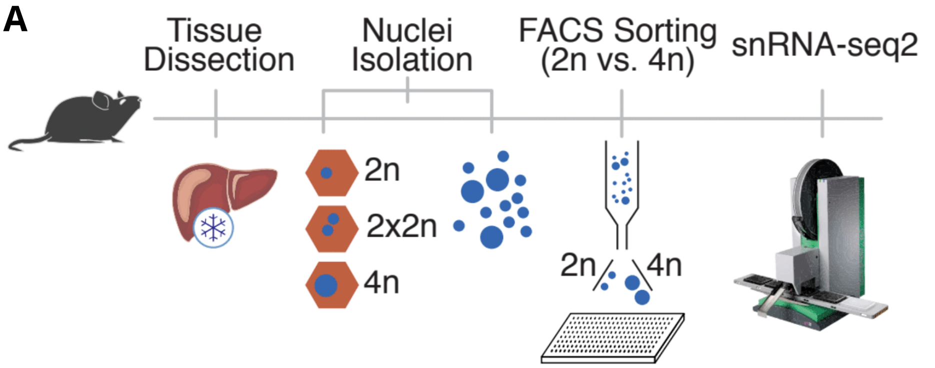

Figure 3.1: A Illustration of the workflow performed by the experimentalist to obtain snRNA-seq2 data. Briefly, flash-frozen liver tissue was dissected, nuclei were isolated and subjected to FACS sorting based on their genome content. A liquid minimization robot was used during the generation of full-length cDNA libraries. This illustration was taken from Richter *et al.* [117]

For the Ploidy&Zonation project described in this chapter, data was generated using snRNA-seq2 that employs FACS to sort nuclei by their genome content. SnRNA-seq2 is a method developed by the Martinez-Jimenez lab that enables data generation from frozen samples. Thus, it allows the exploration of long-term archived samples of both, healthy and diseased conditions [117]. The method is plate-based and follows a modified SMART-seq2 protocol, allowing the recovery of full-length transcripts. As shown in figure 2.1 B, the number of genes per single nucleus obtained through the snRNA-seq2 method outperforms other single cell and nucleus RNA sequencing protocols. In the Ploidy&Zonation project presented in this chapter, the snRNA-seq2 method was used and led to the acquisition of on average 557,385 reads and a median of 3,599 genes per nucleus. The high number of detected genes per nucleus represents an advantage of the snRNA-seq2 method over other approaches. To investigate whether this high number is achieved through the improved chemistry of the protocol or can be obtained by deep sequencing regardless of chemistry, cells were deeply sequenced in a droplet-based 10X Genomics experiment. This attempt, however, failed with 1,000,000 reads per single nucleus only resulting in a median of 2,776 genes per nucleus. Extrapolation on the saturation curve revealed that a hypothetical average sequencing depth of 2.3 million reads per nucleus would be required to reach a median of 3,600 genes per nucleus. (Methods, Figure 2.1 B, right). In line with plate-based approaches achieving higher sequencing depths than droplet-based approaches, this provides further evidence that plate-based approaches yield higher library complexity [57][56].

Working on flash-frozen tissue complicates the isolation of intact whole cells as thawing induces breaks in the cytoplasmatic membrane [139] [140][141]. The isolation of single nuclei is therefore a compromise between being able to work on archived samples and obtaining single cell information [142]. The gene expression from single nuclei and single cell extracts has been compared to determine to which extent data from single nuclei represents the transcriptional landscape of single cells [143]. For instance, Lake *et al.* have reported correlations in the range of 0.53 to 0.74 between nuclear and whole cell gene expression for all detected genes in brain samples, depending on the cell type [143]. In the liver, the transcriptional profile captured by single nucleus RNA-seq has been shown to correspond to bulk measurements [28]. However, it remained to be explored to which extent the nuclear transcriptome of hepatocytes isolated from liver tissue correlates with the whole cell transcriptome (mRNA). To address this question, in the Ploidy&Zonation project, the gene expression obtained through snRNA-seq2 was correlated to the *Tabula muris* data set. This publicly available data set is comprised of whole liver cells obtained using the plate-based smart-seq method [37]. Only hepatocytes were used for this comparison (Figure 3.2). The nuclear transcriptome was found to correlate at a *Pearson* correlation of 0.62 to the total mRNA expression from single cells. In summary, this shows that hepatic nuclei extracts capture the transcriptional landscape of whole hepatocytes in agreement with the previously reported correlation in brain cells [143].

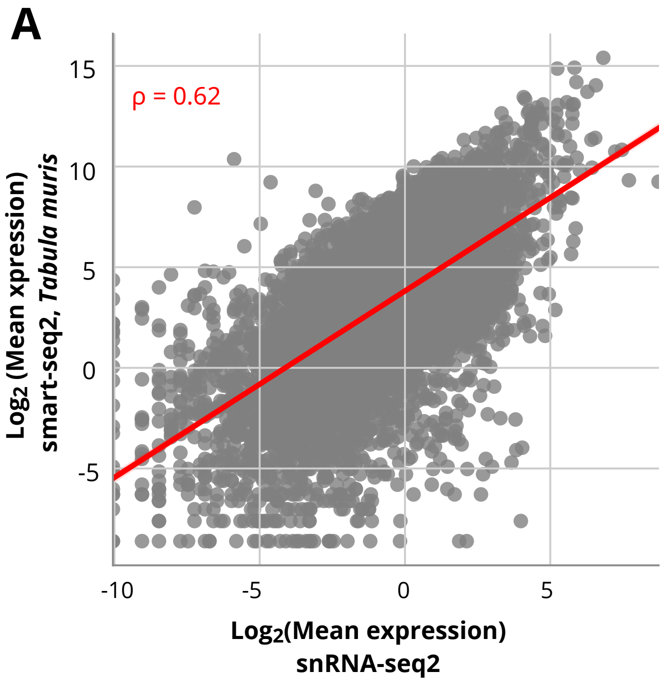

Figure 3.2: A Scatter plot showing the logarithmic mean expression for all genes shared between the snRNA-seq2 data set from this thesis and the *Tabula muris* scRNA-seq data set [37]. Red line indicates the regression line, $\rho = 0.62$. This figure was adapted from Richter *et al.* [117].

3.2 Cell type identification

The snRNA-seq2 data set for the Ploidy&Zonation project stems from young healthy mouse livers that are comprised of a variety of cell types as described in the introductory section of this thesis. Apart from hepatocytes, cell types expected to be present in this data set from liver tissue include Kupffer and Dendritic cells (APCs), endothelial cells, cholangiocytes, hepatic stellate cells (HSCs), and immune related cells. To computationally separate them from each other and annotate them, *Louvain* clustering was performed and publicly available marker genes were used in form of curated marker gene lists (Methods) [50].

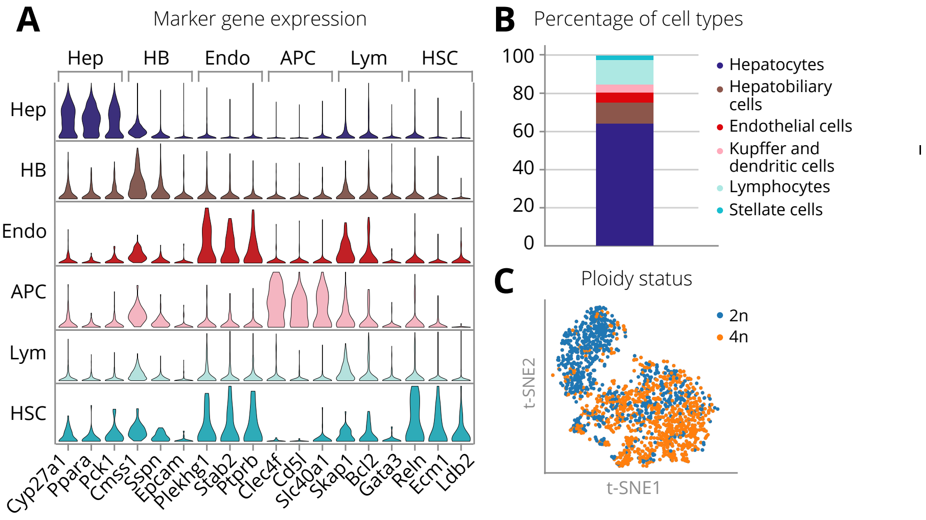

Figure 3.3: **A** Stacked violin plot depicting three representative marker genes per identified cell type in the *in vivo* snRNA-seq2 data set. **B** Bar plot adding up the percentage of different cell types detected from single nuclei sequencing of frozen mouse liver. **C** t-SNE embedding of the single nuclei colored by their ploidy status.

In brief, low resolution *Louvain* clustering (resolution = 0.2) separated the non-parenchymal cells (NPCs) from hepatocytes, a split that can also be observed on the t-SNE embedding (Figure 2.2 A). Marker genes such as *Hnf4α*, *Ces3a*, and *Cyp27a1* were used to identify hepatocytes. In line with what is expected from histological reports, 64.3 % of the sequenced nuclei were assigned to hepatocytes while the rest were stemming from NPCs (Figure 3.3 B) [10][11]. The FACS sorting of nuclei based on genome content allows to assess the ploidy levels of individual nuclei. As shown in figure 3.3 C, the majority of NPCs were found to be diploid (upper left population on the t-SNE) while hepatocytes were split into diploid and tetraploid nuclei. However, NPCs featured a small population of tetraploid cells, making up 6.6 % of the data. These could either be nuclei clumping together during sorting, or nuclei from dividing cells that contain a duplicated genome due to replication. To address this question, cell cycle analysis was performed using *cyclone*. This software assigns cells to either G1, S, or G2M based on scores for cell cycle marker genes [144]. Overall, the majority of nuclei were assigned to G1, indicating their gene expression profile was not associated to active division (Figure 3.4 A and B). Hepatocytes showed the lowest proportion of nuclei in division with only 5.6 % of hepatocyte nuclei being assigned to either S or G2M phase. This is in line with studies showing that the majority of hepatocytes reside in the liver in a non-divisive, quiescent state (G0) with an estimated turnover of 1 in 10,000 to 40,000 [145][146][147]. The highest proportion of dividing nuclei was found in hepatic stellate cells (HSCs), where 21.1 % of the nuclei were computationally identified to be either in S or G2M phase (Figure 3.4 C). Out of the 109 nuclei sorted as tetraploid in the NPC fraction, only 11 nuclei were dividing. The rest of these nuclei were therefore most likely sorted as tetraploid due to nuclei clumping together during sorting.

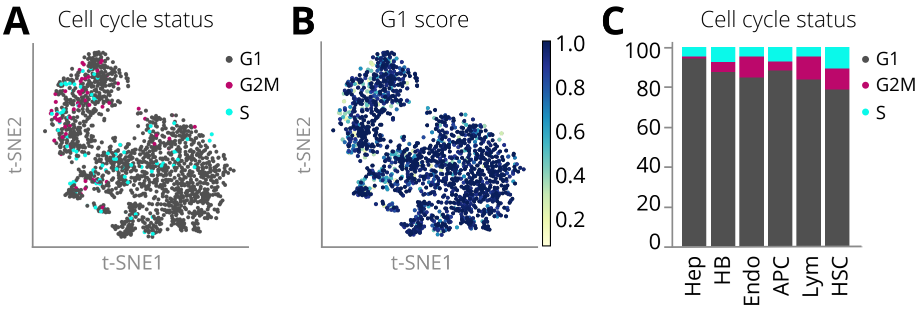

Figure 3.4: A t-SNE embedding of single nuclei from young adult mouse livers, colored by assigned cell cycle phase using *cyclone*. **B** t-SNE embedding colored by G1 score calculated by *cyclone*. **C** Bar plot showing the percentages of nuclei computationally assigned to G1, S, and G2M phase per cell type.

3.3 Transcript expression

Due to different chemistries being used, plate-based methods generally achieve higher sequencing depths per cell/nucleus (at the expense of sequencing fewer cells/nuclei) compared to droplet-based methods that in contrast have a high cellular throughput with low sequencing depth [57]. This allows for plate-based methods to achieve higher sensitivity in capturing lowly expressed genes. For instance, transcription factors tend to be more lowly expressed and therefore harder to capture than other genes in s*RNA-seq approaches [148].

The snRNA-seq2 method that was used for data generation in this chapter is a plate-based method and therefore enables the capture of lowly expressed genes, including key hepatic transcription factors. One of the captured transcription factors, for instance, is Mlxipl, also known as ChREBP, which regulates the pathway responsible for β-oxidation of fatty acids whereas Hnf4α regulates expression in carbohydrate-related pathways [100][149]. As depicted in figure 3.5 A, both *Mlxipl* and *Hnf4α* as well as other key hepatic transcription factors were evenly expressed between diploid (2n) and tetraploid (4n) hepatocyte nuclei (hereafter called 2n and 4n hepatocytes, respectively).

As described above, the modified SMART-seq2 chemistry used here also generates full-length transcripts comprised of intronic and exonic regions. Moreover, nuclear extracts contain a higher ratio of unspliced transcripts than whole cell extracts. This gives the opportunity to study the expression of different isoforms in hepatocytes. Figure 3.5 B shows the respective isoforms that were detected for the eight transcription factors depicted in figure 3.5 A. For the genes *Rxra*, *Nr1i2*, and *Cebpα* one isoform dominated gene expression across all hepatocytes in the snRNA-seq2 data, whereas for *Mlxipl*, *Hnf4α*, *Pparα* several isoforms contributed in different intensities to gene expression (Figure 3.5 B). Differences in isoform expression for *Mlxipl* have previously been linked to metabolic activity [150]. No preference of isoforms for a given transcription factor was found between 2n and 4n hepatocytes. Nevertheless, some isoforms and transcription factors showed slightly higher levels of expression in 4n hepatocytes. This could be due to the higher number of endogenous counts observed in 4n in comparison to 2n hepatocytes (Figure 3.5 C).

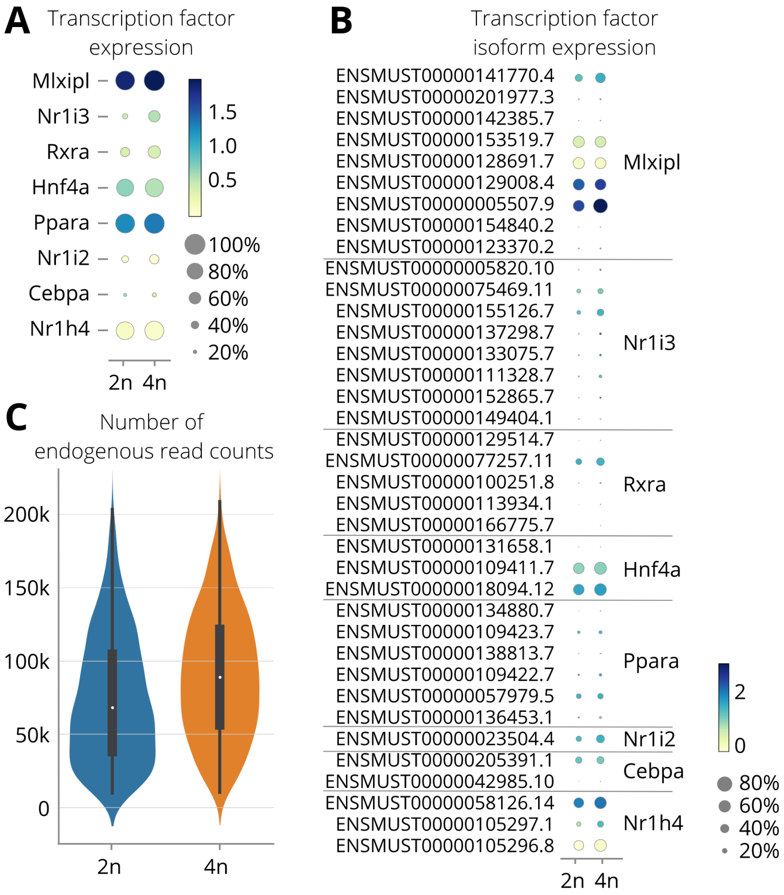

Figure 3.5: A Dot plot depicting the average gene expression (color bar) of selected transcription factors in 2n and 4n hepatocytes and the percentage of nuclei in which they are expressed (dot size) **B** Dot plot depicting the average isoform expression (color bar) for the transcription factors shown in **A**. **C** Violin plot showing the number of endogenous read counts in 2n and 4n hepatocytes.

3.4 Transcriptional variability

The higher number of endogenous read counts in 4n hepatocytes is likely due to the larger genome content resulting in larger transcript abundances linked to mechanisms related to gene dosage compensation and expression homeostasis [65][66][67][151]. The role of polyploidy on the tissue function under physiological conditions remains largely unexplored. In hepatocellular carcinoma (HCC), polyploidy has initially been postulated to diminish tumor suppressor loss-of-heterozygosity (LOH) [65][3]. However, an enrichment of higher ploidy levels (>8n) has been observed in HCC and is associated to poor prognosis [5]. When comparing 4n to 2n hepatocytes, Halpern *et al.* have found lower transcriptional variability levels in 4n hepatocytes [14].

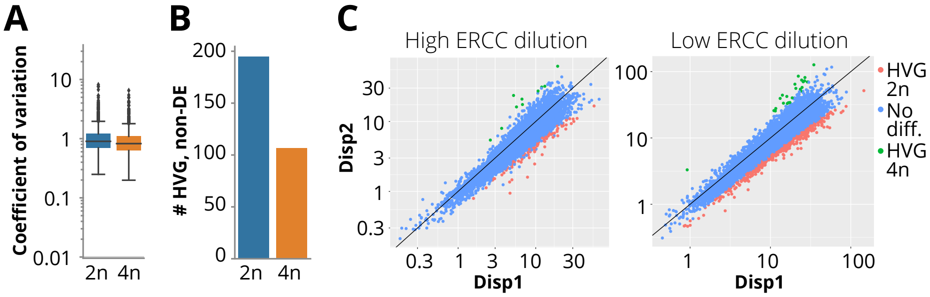

Figure 3.6: A Boxplot showing the coefficient of variation for genes in 2n and 4n hepatocytes. **B** Bar plot illustrating the number of highly variable genes (HVGs) that were not differentially expressed in 2n and 4n hepatocytes. **C** Scatter plot depicting the dispersion for genes in 2n and 4n hepatocytes, calculated using BASiCS. Left: nuclei with high ERCC dilution were used, right: nuclei with low ERCC dilution were analyzed. Color indicates level of variability (red: HVG in 2n, blue: not highly variable, green: HVG in 4n)
Panels **A** and **B** adapted from Richter *et al.* [117]

To address transcriptional variability in the snRNA-seq2 data set, the coefficient of variation was calculated as described in the Methods section of this thesis according to Canchola *et al.* [138]. As shown in figure 3.6 A, 2n hepatocytes indeed had significantly higher transcriptional variability compared to 4n hepatocytes. Moreover, focusing the analysis on non-differentially expressed (non-DE) genes, a higher amount of highly variable genes (HVGs) was found in 2n hepatocytes. This results confirm the single-molecule study by Halpern *et al.*

who also found higher transcriptional variability in 2n hepatocytes compared to 4n [14]. Additionally to calculating the coefficient of variation, the software BASiCS was used to address variability between 2n and 4n hepatocytes taking into account the variation of the ERCCs [134]. BASiCS uses the ERCCs to separate technical from biological variation through a Bayesian hierarchical model and allows to annotate highly or lowly variable genes. As two different ERCC dilutions were used in the scRNA-seq2 data set, this was performed separately for the two dilutions. For both dilutions, more highly variable genes were identified in the 2n population (Figure 3.6 C). Thus, 2n hepatocytes showed higher transcriptional variability than 4n hepatocytes regardless of the method used.

3.5 Differential expression between 2n and 4n hepatocytes

In comparison to 2n hepatocytes, 4n hepatocytes featured both, a higher total amount of genes detected and more differentially expressed genes (DEGs, Figure 2.2 F and Figure 3.7 A). Therefore, the questions arose why there were more DEGs in 4n hepatocytes and whether genes that were highly expressed in 4n hepatocytes were also detected in 2n hepatocytes. Not detecting these genes in 2n hepatocytes could artificially bias differential expression towards 4n hepatocytes. Across all genes detected in hepatocytes, genes that were not present in 2n hepatocytes showed low to moderate expression levels in 4n (Figure 3.7 B). Hence, the lower amount of detected genes in 2n hepatocytes did not bias the differential expression analysis, meaning that the DEGs were driven by biological differences.

To get a comprehensive overview of what pathways are up-regulated in 2n and 4n hepatocytes, respectively, gene ontology (GO) analysis was performed based on the DEGs between 2n and 4n hepatocytes. As depicted in figure 3.8 A and B, in comparison to 4n hepatocytes, 2n hepatocytes up-regulate pathways related to glucose metabolism, including the pathway of gluconeogenesis, the hexose biosynthetic process, and the monosaccharide biosynthetic process. In contrast, 4n hepatocytes up-regulate pathways involved in sterol, cholesterol, lipid, and xenobiotic metabolism. Apart from performing unsupervised GO analysis, the expression of individual genes was investigated. The majority of the investigated key hepatic transcription factors were not differentially expressed between 2n and 4n hepatocytes (Figure 3.8 C). However, for some transcription

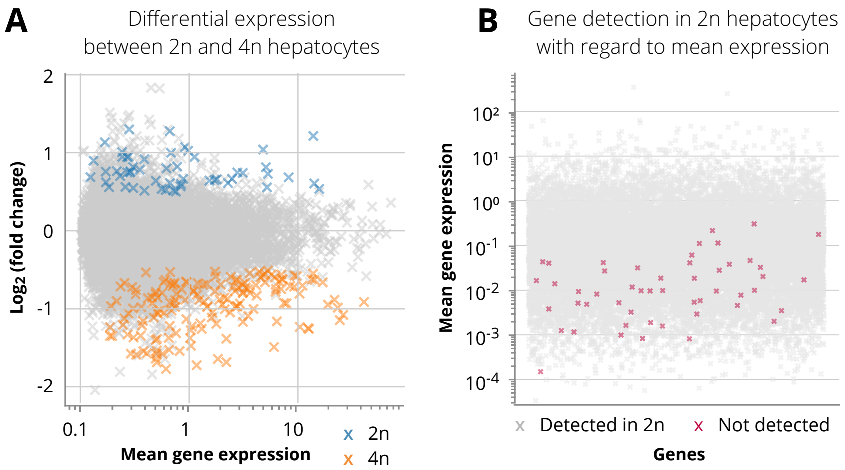

Figure 3.7: A MA-plot showing the $\log_2$-fold change against the mean gene expression for genes with mean expression > 0.1 (colors indicate differential expression, blue: up-regulated in 2n, orange: up-regulated in 4n) **B** Scatter plot depicting the mean expression of all genes detected in hepatocytes and whether they are detected in 2n hepatocytes (grey: detected in 2n and 4n, red: not detected in 2n).
Figure adapted from Richter *et al.* [117]

factors, a change in distribution was observed, indicating heterogeneity between the 2n and 4n population of nuclei for these genes. *Pck1* was significantly up-regulated in 2n hepatocytes, and, while not significantly, gene expression of *G6pc* was also increased in 2n hepatocytes, providing further evidence of their increased involvement in glucose metabolism (Figure 3.8 D). All the selected genes responsible for lipid metabolism showed changes in the distribution of gene expression (Figure 3.8 E). Moreover, while *Acaca* and *Alb* were significantly up-regulated in 2n hepatocytes, expression of *Acox2* was significantly increased in 4n hepatocytes, suggesting that lipid metabolism is heterogeneous in hepatocytes. In line with up-regulation of xenobiotic metabolism related pathways, 4n hepatocytes showed significant up-regulation of several cytochromes from the cytochrome P450 pathway responsible in phase I drug metabolism (Figure 3.8 F). Out of the selected cytochromes, only *Cyp2f2* was significantly up-regulated in 2n hepatocytes. In contrast, 2n hepatocytes showed up-regulation of genes involved in protein metabolism, including *Gls2*, *Hal*, *Hpx*, and *Sds* (Figure 3.8 G).

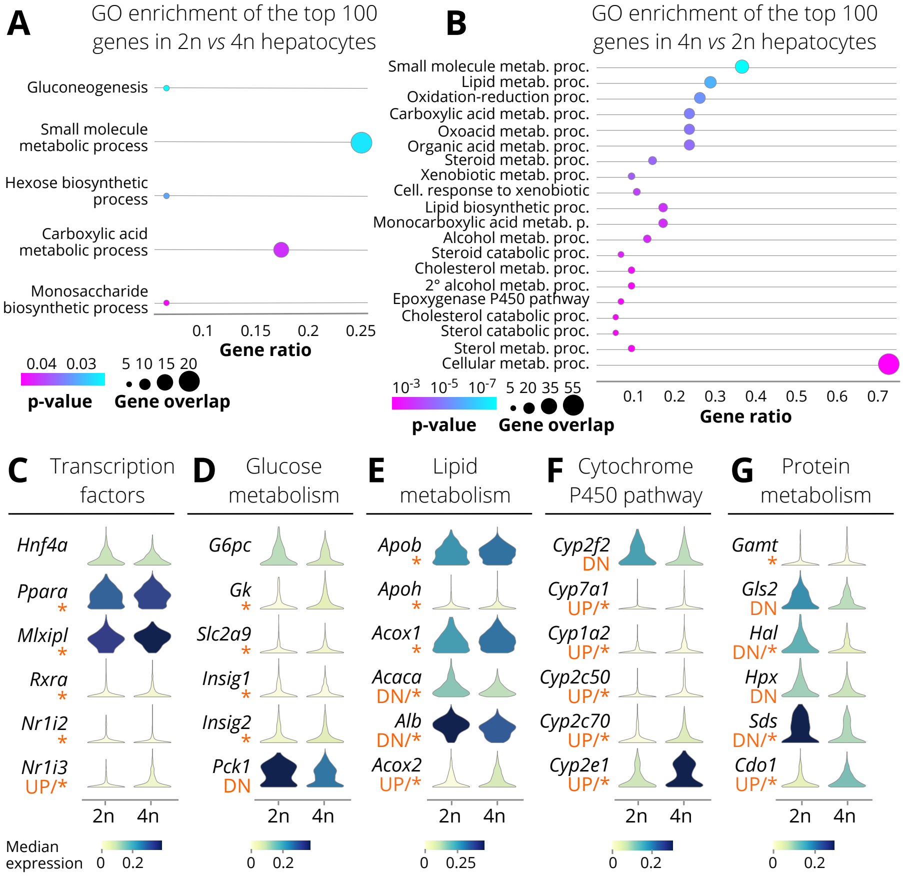

Figure 3.8: A Scatter plot depicting the GO terms the top 100 genes in 2n vs. 4n hepatocytes are enriched in. Dot size represents overlap between DEGs and genes in pathways, color refers to p-value. **B** Scatter plot depicting the GO terms the top 100 genes in 4n vs. 2n hepatocytes are enriched in. Dot size represents overlap between DEGs and genes in pathways, color refers to p-value. **C-G** Stacked violin plots showing gene expression between 2n and 4n hepatocytes of transcription factors (**C**), markers for glucose metabolism (**D**), markers for lipid metabolism (**E**), genes in the cytochrome P450 pathway (**F**), and markers for protein protein (**G**).

3.6 Higher levels of ploidy

As described in the introductory section of this thesis, further failures in cellular divisions can also give rise to higher levels of ploidy, such as octoploid (8n), and hexadecaploid (16n) nuclei [13]. To shed light on these populations, data from an experiment comprised of 4n, 8n, and 16n hepatocytes were analyzed. In line with the higher concentration of mRNA associated to bigger cell size [67], an increase in the number of detected genes per nucleus was observed with ploidy level (Figure 3.9 A).

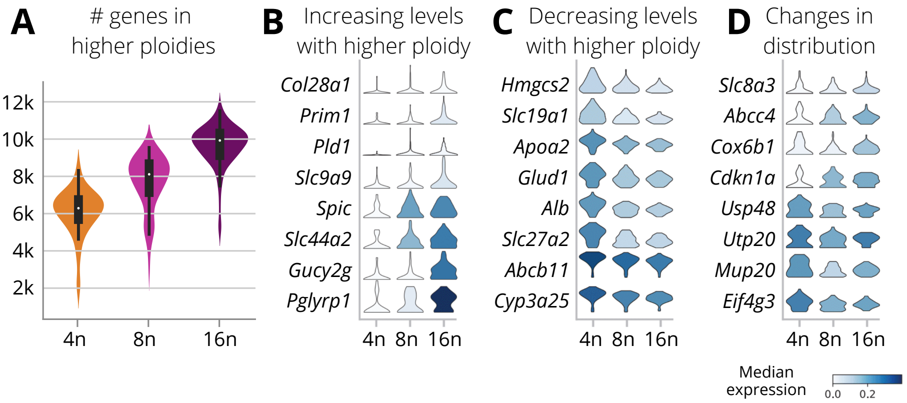

Figure 3.9: A Violin plot depicting the number of genes per nucleus depending on ploidy status for 4n, 8n, and 16n nuclei. **B** Stacked violin plot of representative genes that significantly increase in expression level with ploidy. **C** Stacked violin plot of representative genes that significantly decrease in expression level with ploidy. **D** Stacked violin plot of representative genes that remain unchanged in their mean expression with ploidy but change in distribution.
Figure adapted from Richter *et al.* [117]

As therefore expected, several genes were detected to significantly increase with ploidy level in their expression. For instance, the solute carriers *Slc9a9* and *Slc44a2* were found to increase with ploidy status (Figure 3.9 B). However, despite the overall higher transcript abundance in nuclei of higher ploidy levels, other genes were detected to decrease with ploidy status. For example, the solute carriers *Slc19a1* and *Slc27a2* were found to decrease (Figure 3.8 C). Moreover, *Alb* and the cytochrome *Cyp3a25* showed decreasing expression levels with ploidy (Figure 3.8 C). Additionally, several genes were identified with no significant changes in mean expression between nuclei of different ploidy but which showed significant changes in distribution between all ploidy levels (Figure 3.8 D). Overall, these results imply that ploidy introduces heterogeneity among hepatocytes and the ploidy status of nuclei has a complex effect on their gene expression.

3.7 Stem cell properties in polyploid hepatocytes

A question that has been sparking debate in the field is whether and to what extent polyploid hepatocytes can involve in tissue regeneration upon injury. For example, Wang *et al.* postulated the existence of a pericentrally located stem cell niche in liver prompting regeneration [152]. On the other hand, Lin *et al.* found *Tert*-positive cells, involving in tissue regeneration, to be evenly distributed along the liver lobule [153]. Moreover, other studies conclude that all hepatocytes along the liver lobule have regenerative potential and up-regulate *Axin2* and *Lgr5* while *Lgr5* has generally been found to be higher in the pericentral regions [154][155].

To investigate stem cell potential of 2n and 4n hepatocytes, the expression and co-expression of nine selected stem cell markers was explored [156]. As shown in figure 3.10 A, both 2n and 4n hepatocytes showed presence of the nine stem cell marker genes. Moreover, *Axin2*, *Tbx3*, *Lgr5*, *Itga6*, *Tert*, and *Notch2* were found to be co-expressed across nuclei, regardless of ploidy status (Figure 3.10 B). Especially *Tbx3* and *Lgr5* showed high co-expression levels, measured by Jaccard distance (Methods, Figure 3.10 B). In brief, using the presence or absence of a given marker gene within the nuclei without taking into account its expression levels allows to explore co-existence within a single nucleus. Moreover, it can be assessed what percentage of nuclei co-express several stem cell markers simultaneously. Table 3.1 shows that 43.0 % of 2n nuclei, and 33.7 % of 4n nuclei were co-expressing two of the nine stem cell markers. Slightly higher percentages of co-expression of more than three stem cell markers were found in 4n hepatocytes compared to 2n. While this indicates that indeed 4n hepatocytes express genes related to regenerative potential, lack of 2n hepatocytes co-expressing stem cell markers can also be due to dropouts in the nuclei with lower genome content. In summary, the percentage of nuclei co-expressing several stem cell markers was similar between 2n and 4n hepatocytes, indicating that both diploid and tetraploid hepatocytes most likely share regenerative potential at the transcriptomic level.

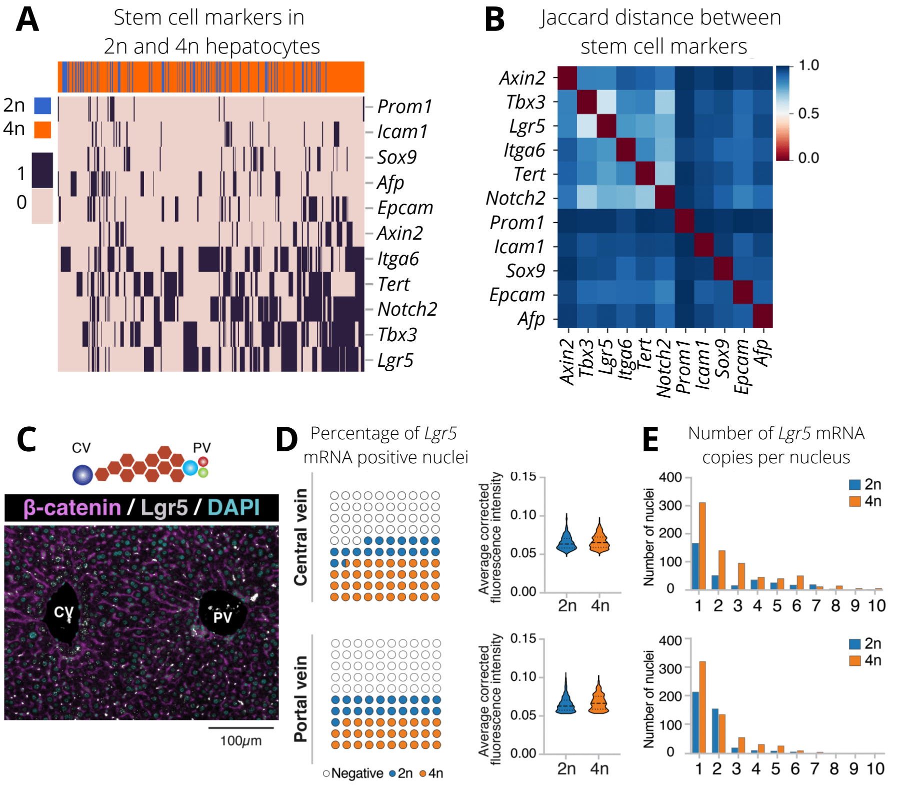

Figure 3.10: A Heatmap depicting binary expression levels of nine representative stem cell markers across all single nuclei. Columns represent single nuclei, rows represent genes, color bar refers to ploidy status of the single nuclei. **B** Heatmap illustrating the Jaccard distance between the nine stem cell markers as a measure of co-expression across nuclei the nuclei. Smaller distance values indicate higher co-expression. **C** Representative image of the (IF)/RNA-FISH co-detection analysis performed on a liver lobule by Dr. Kelvin Yin. The mRNA expression levels of Lgr5 are shown in gray, β-catenin is shown in magenta, and 4',6-diamidin-2-phenylindol (DAPI) in cyan. **D** Nuclei within 50μm radius of the pericentral or periportal vein were assigned CV, and PV, respectively. Ploidy status was assessed based on nuclear diameters, and the percentage of Lgr5-positive nuclei was calculated (left). Violin plots illustrate the fluorescence intensity of Lgr5 (normalized against the background) in 2n and 4n nuclei (right; ns=not significant, t-test) Top: CV, bottom: PV. **E** Bar plots depicting the number of Lgr5 mRNA copies per nucleus in 2n and 4n hepatocytes (top: CV, bottom: PV).
Figure adapted from Richter *et al.* [117].

Table 3.1: Percentage of nuclei co-expressing several stem cell markers

No. stem cell markers	% All	% 2n	% 4n
2	35.7	43.0	33.7
3	26.1	19.0	28.1
4	17.3	10.1	19.3
5	7.7	5.1	8.4
6	3.6	2.5	3.9
7	0.3	0	0.4

This finding obtained by analyzing the snRNA-seq2 data set was experimentally confirmed by Dr. Kelvin Yin using an immunofluorescence (IF)/RNA-FISH co-detection technique ("RNAscope"). Briefly, β-catenin and stem cell marker *Lgr5* mRNA were simultaneously measured in the liver lobule (Figure 3.10 C-E). In line with the findings of Chen *et al.*, the area surrounding the central vein (CV) featured a higher percentage of *Lgr5* positive hepatocytes than the area around the portal vein (PV) [155]. No significant differences were found in the *Lgr5* levels and copies per nucleus between 2n and 4n hepatocytes (Figure 3.10 D and E), again indicating that both populations share similar regenerative potential.

3.8 Zonation

Zonation describes the bi-directional gradient of gene expression resulting from the gradual supply of oxygen, nutrients, and hormones to the liver cells along the lobule [14][16][157]. Along the zonated liver lobule, 2n hepatocytes have been observed to reside more closely to the periportal vein (PV) while polyploid hepatocytes were reported to reside closer to the central vein (CV) [85]. In an early study exploring zonation at the single cell transcriptome level, Halpern *et al.* combined single-molecule ISH (smISH) with scRNA-seq and succeeded to assign several marker genes across the mouse liver to zones based on their spatial expression patterns [16]. The study by Halpern *et al* is therefore used as a resource for zonation marker genes in this thesis. As described in the introduction, spatial information can computationally be inferred from single cell or snRNA-seq data by applying diffusion pseudotime (dpt) algorithms [90][77].

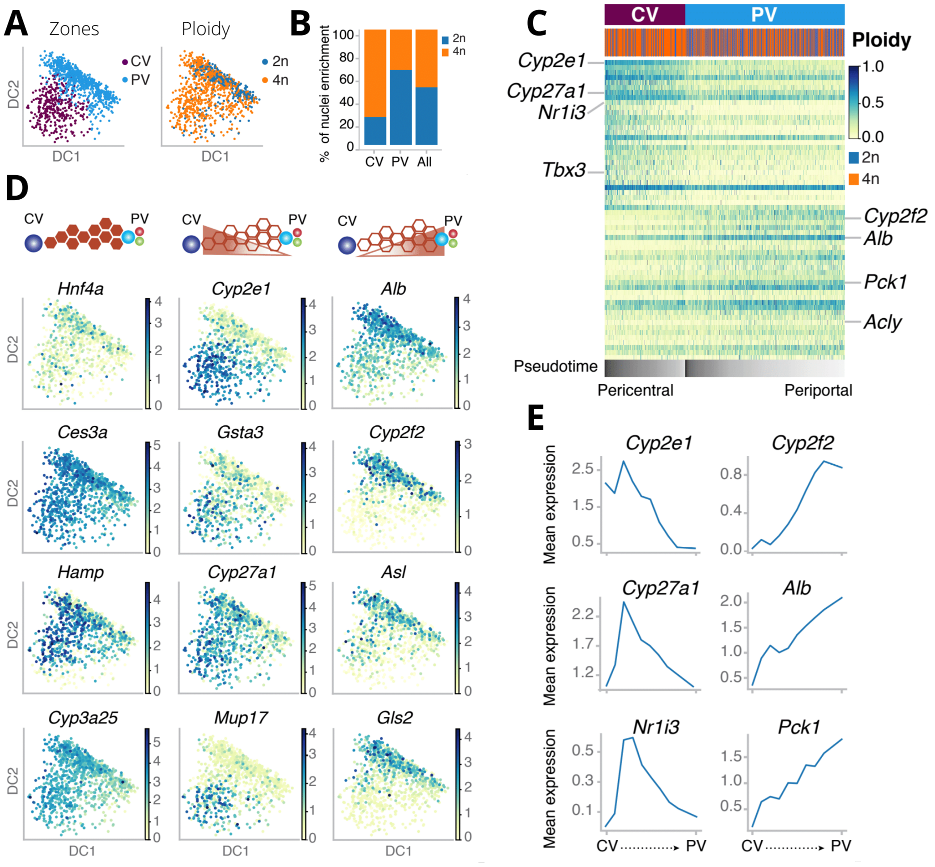

Figure 3.11: A Diffusion maps of the hepatocytes from the *in vivo* snRNA-seq2 data set, colored by assigned zones (left), or ploidy status (right). **B** Bar plot depicting the percentage of 2n and 4n nuclei in the CV and PV zone, respectively. **C** Heatmap illustrating the bi-directional expression gradient along the vector of pseudospace for the top 30 DEGs between CV and PV. Columns represent single nuclei, rows represent genes. **D** Diffusion maps of the hepatocytes from the snRNA-seq2 data set, colored by expression level of representative zonation markers (left: non-zonated markers, middle: CV markers, right: PV markers. **E** Line plots showing mean expression of representative zonation markers along the vector of pseudospace (left: CV markers, right: PV markers. Figure adapted from Richter et al. [117].

This approach is performed here, where calculating diffusion components (DCs) is used as means of dimensionality reduction (Methods, Figure 3.11 A).

For this analysis, the set of known zonation marker genes published by Halpern *et al.* was used to order the nuclei purely based on expression characteristics associated to zonation. After calculating dpt, *Louvain* clustering was performed and the zonation marker genes were used to assign clusters to a CV and a PV zone, respectively (Figure 3.11 A, left). As can be seen already in the diffusion map, calculating the percentage of 2n and 4n hepatocytes in the CV and PV zone showed that 2n hepatocytes were more periportally whereas 4n hepatocytes were more pericentrally enriched (Figure 3.11 A, right, and B). For the purpose of visualization, in figure 3.11 D, representative marker genes were selected and depicted on the diffusion map. Genes not supposed to show zonation patterns at the mRNA level, such as *Hnf4α*, *Ces3a*, *Hamp*, and *Cyp3a25*indeed showed a homogeneous distribution across the diffusion map [6] (Figure 3.11 D, left column). Other genes representing pericentral markers, such as *Cyp2e1*, *Gsta3*, *Cyp27a1*, and *Mup17* showed, as expected, higher expression levels in the nuclei assigned to CV. On the contrary, periportal marker genes like *Alb*, *Cyp2f2*, *Asl*, and *Gls2* showed higher expression levels in the nuclei assigned to PV. Furthermore, calculating dpt allows ordering the nuclei from CV to PV based on progressive changes in the expression levels of the used zonation marker genes. As depicted in figure 3.11 C, pericentrally expressed marker genes, such as *Cyp2e1*, *Nr1i3*, and *Tbx3* showed highest expression at the pericentral zone and gradual decrease in expression toward the periportal area. Likewise, periportally enriched genes, such as *Cyp2f2 Alb*, and *Acly* showed gradual increase in expression level in nuclei ordered from CV to PV. Overall, this heatmap represents in an easily interpretable way the bi-directional expression gradient of zonation markers in the mouse liver by visualizing the expression levels of 40 genes simultaneously that were differentially expressed between PV and CV. Looking at individual marker genes, however, allows to explore their changes in mean expression along the vector of pseudospace referring to zonation in more detail. For instance, figure 3.11 E shows that especially *Alb* and *Pck1* gradually increase in mean expression from pericentral to periportal, with the out-most assigned nuclei each representing the lowest, and highest expression level, respectively (Figure 3.11 E). *Pck1* shows a step-wise increase indicating the presence of zones along the lobule, whereas e.g. *Cyp27a1* is most highly expressed in nuclei assigned close to CV from where the expression continuously decreases.

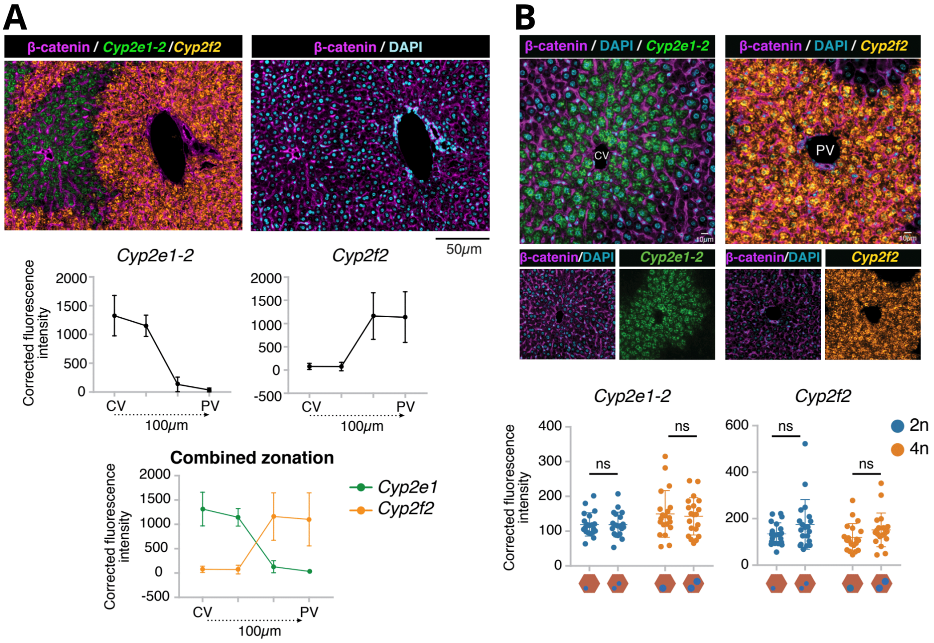

Figure 3.12: A Representative image of the (IF)/RNA-FISH co-detection analysis performed on a liver lobule by Dr. Kelvin Yin. The mRNA expression levels of Cyp2e1-2 are shown in green and of Cyp2f2 in yellow (top; the fluorescence intensity of each gene was normalized against the background). Line plots showing the normalized fluorescence intensity over a linear distance (100 μm) from CV to PV, separately for the two markers, and in combination (bottom, mean ± s.d. is shown). **B** The normalized fluorescence intensity of Cyp2e1-2 and Cyp2f2, respectively, was categorized in regards to 2n or 4n mono-nucleated or bi-nucleated hepatocytes. Ploidy status was determined by *β*-catenin (magenta) and DAPI (cyan) staining. No significant changes were detected, neither between ploidy levels nor the number of nuclei per cell (ns= not significant, unpaired t-test, lines in the scatter plots indicate mean ± s.d.).
Figure adapted from Richter et al. [117].

These findings from the snRNA-seq2 data were again experimentally validated by Dr. Kelvin Yin through immunofluorescence (IF)/RNA-FISH co-detection ("RNAscope"). As shown in figure 3.12 A, mRNA of Cyp2e1 was pericentrally enriched whereas mRNA of Cyp2f2 was periportally enriched. Through measuring fluorescence intensities along the liver lobule, the bi-directional expression gradient as seen in the nuclear transcriptomic data was further confirmed for the two selected marker genes. Moreover, the RNAscope method was used to investigate differences in the mRNA expression levels between mono- and bi-nucleated 2n or 4n cells. No significant differences were found between mono-nucleated and bi-nucleated cells (figure 3.12 B).

To further investigate the relationship between zonation and polyploidy, data was analysed from mouse livers where zonation was disrupted. A collaboration with Prof. Dr. Neil Henderson (University of Edinburgh) led to study an additional mouse model of liver fibrosis, which has been shown to disrupt zonation [158][159][160][161][162]. Prof. Henderson established a working mouse model of liver fibrosis by treating the mice with carbon tetrachloride (CCl_4), inducing hepatocyte necrosis, structural changes around the central vein, and activation of hepatic stellate cells [159][163]. In this model, CCl_4 is diluted in olive oil, and control mice are thus treated with olive oil alone. As shown in figure 3.13 (left), the analysis of the snRNA-seq2 data from mice treated with olive oil confirmed the findings of a pericentral enrichment of 4n hepatocytes. *Louvain* clustering again was able to separate and assign to either a PV or a CV region (Figure 3.13 A). The percentage of 2n and 4n hepatocytes in each of the two zones reflected the previous findings in wild-type, untreated mice with 2n being periportally, and 4n hepatocytes being pericentrally enriched. Visually, this can be inspected when comparing figure 3.11 B and figure 3.13 B. The spatial distribution of marker gene expression, depicted in figure 3.13 C, showed similar zonation patterns between CV and PV as had been observed in the wild-type untreated mice (Figure 3.11). This additionally indicated that in wild type animals, the relationship between zonation and ploidy is comparable between mice from different labs and is not affected by the administration of olive oil. However, when zonation patterns were disrupted through fibrosis, gene expression of zonation markers was affected. As shown in figure 3.13 (right), assignment of zones through *Louvain* clustering became more challenging as zonation marker genes followed less clear zonation patterns (Figure 3.13 F). For instance, the pericentral markers *Cyp27a1* and *Gsta3* were expressed in the same nuclei (located in the lower left corner of the diffusion map) as the periportal markers *Alb* and *Cyp2f2*. Overall, fewer nuclei were assigned to CV in the fibrosis model (Figure 3.13 A). Additionally, fewer 2n hepatocytes were found in the fibrotic liver (Figure 3.13 E).

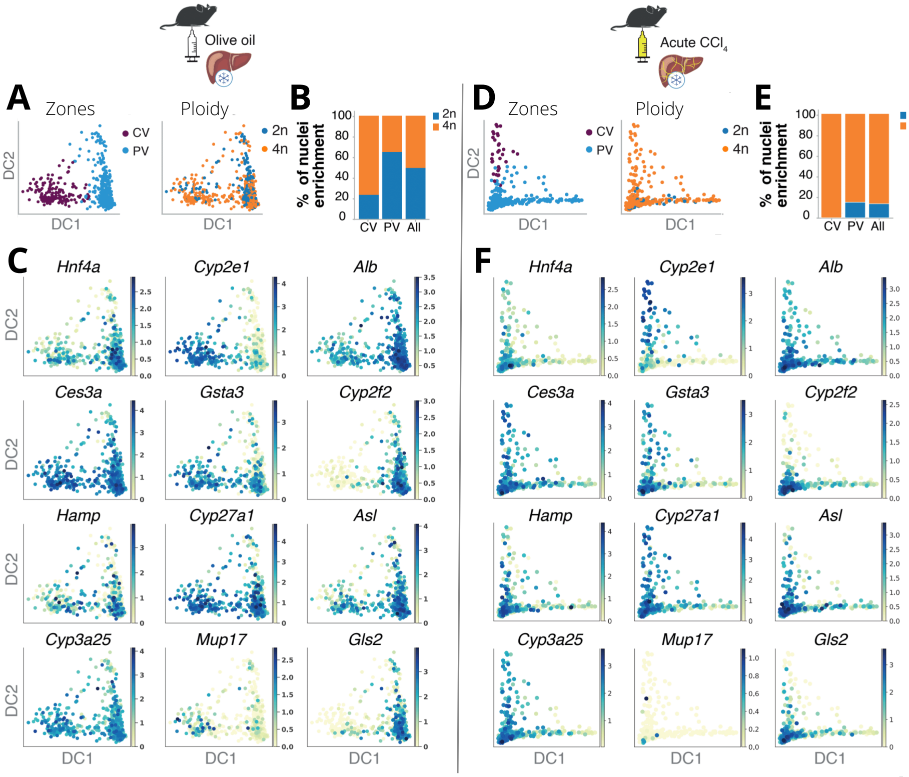

Figure 3.13: **A** Diffusion maps of a control mouse model treated with olive oil, colored by assigned zones (left) and ploidy status (right). **B** Bar plot showing the percentage of 2n, and 4n nuclei in CV, and PV, respectively. **C** Diffusion maps of a control mouse model treated with olive oil, colored by representative zonation markers (left: not-zonated markers, middle: CV markers, right: PV markers. **D** Diffusion maps of a mouse model treated with CCl_4 to model liver fibrosis, colored by assigned zones (left) and ploidy status (right). **E** Bar plot showing the percentage of 2n, and 4n nuclei in CV, and PV, respectively. **F** Diffusion maps of a mouse model treated with CCl_4 to model liver fibrosis, colored by representative zonation markers (left: not-zonated markers, middle: CV markers, right: PV markers.
Figure adapted from Richter et al. [117].

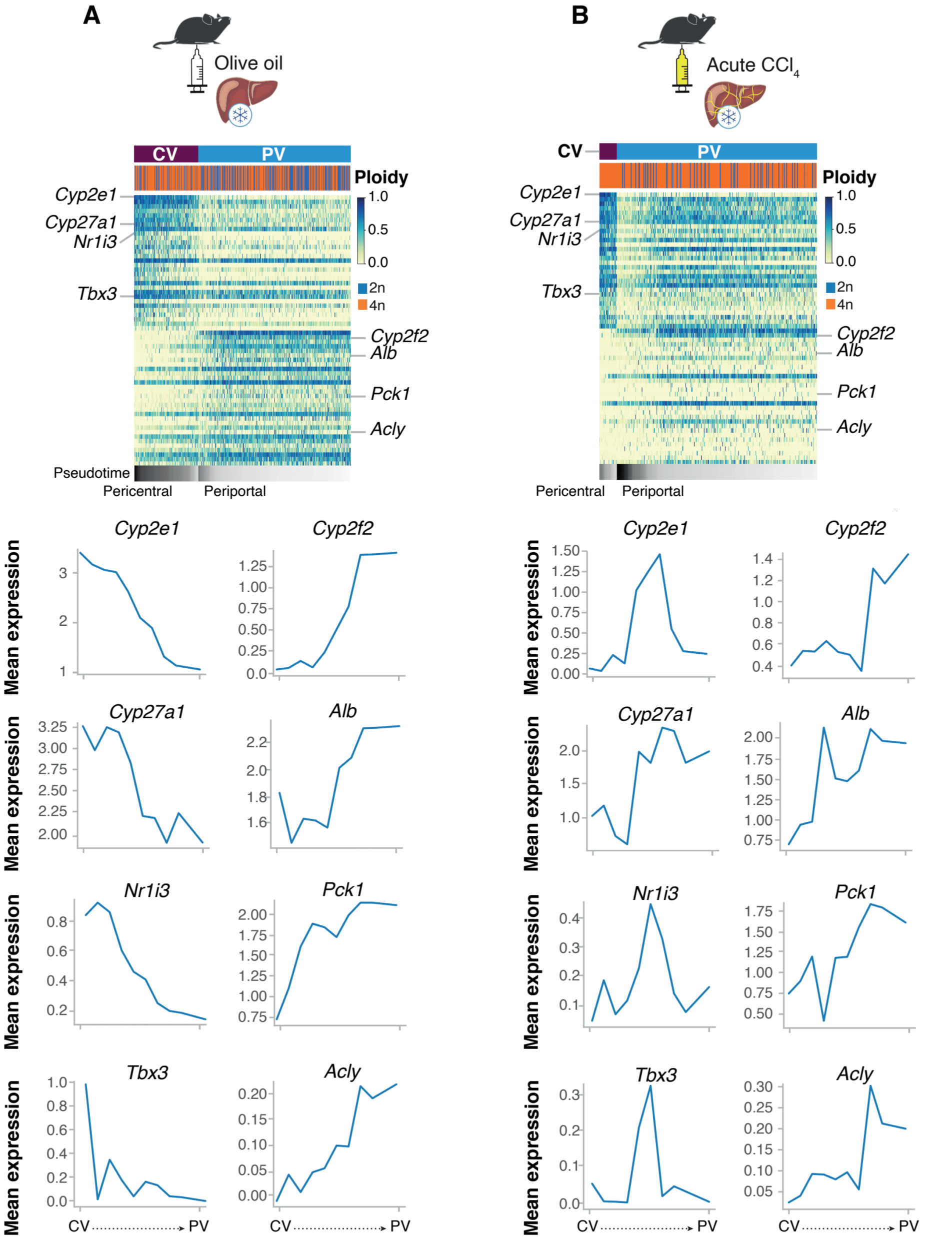

Figure 3.14: A Heatmap showing the bi-directional gene expression gradient for the top 30 DEGs between CV and PV, for a control mouse model treated with olive oil (top). Line plots show the mean expression of representative zonation markers along the vector of pseudospace (bottom; first column: CV markers, second column: PV markers). **B** Heatmap showing the bi-directional gene expression gradient for the top 30 DEGs between CV and PV, for a mouse model treated with CCl_4 to induce liver fibrosis (top). Line plots show the mean expression of representative zonation markers along the vector of pseudospace (bottom; first column: CV markers, second column: PV markers). Figure taken from Richter et al. [117].

Disruption of zonation patterns in a fibrotic mouse model was even more evident when looking at the expression of individual marker genes along the vector of diffusion pseudospace for both the olive oil and the CCl_4 treated mice. Figure 3.14 shows the bi-directional expression gradient for 40 zonation markers between CV and PV (top), as well as for individual marker genes (bottom). For the olive oil treatment (Figure 3.14 A) the spatial expression changes followed the expected bi-directional zonation patterns from CV to PV. For the CCl_4 treated mice, however, the bi-directional expression gradient was less evident. While the expected increase in expression from CV to PV could partially be observed for *Cyp2f2*, *Alb*, *Pck1*, and *Acly*, pericentral marker genes were not found to gradually decrease along the vector of pseudospace (Figure 3.14 B).

In summary, this chapter provided evidence for functional differences between 2n and 4n hepatocytes with zonation being a factor that is interconnected with these ploidy-associated differences. Both 2n and 4n hepatocytes showed gene expression profiles associated with regenerative potential. In comparison to 2n hepatocytes, 4n hepatocytes showed lower transcriptional variability, and an enrichment in xenobiotic metabolism, which is related to their rather pericentral gene expression profile.

4 Zonation-independent hepatocyte heterogeneity

The focus of this chapter lies on studying what cellular heterogeneity is present in primary human hepatocytes (PHHs) in vitro. Since PHHs are taken out of the lobular environment of the liver, they lack zonation patterns in 2D culture. Moreover, they tend to lose their characteristic hepatocyte-like phenotype over time in culture. In summary, this chapter focuses on dissecting the cellular heterogeneity in PHHs, which are a seemingly homogeneous population of cells. In addition to this, this chapter addresses how this heterogeneity plays a role under chronic exposure to free fatty acids (FFA), and the effect of a cocktail of five drugs to functionally study the transcriptional responses of individual cells to a defined metabolic challenge. The majority of results presented here can be found in Sanchez-Quant et al. 2023 [164].

4.1 Method overview

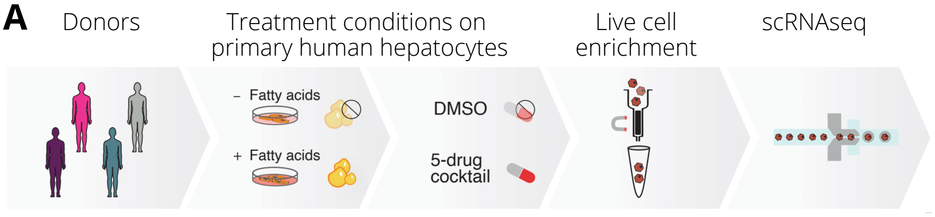

Figure 4.1: A Illustration of the workflow performed by the experimentalist to obtain the scRNA-seq data from *in vitro* PHHs.
Figure adapted from Sanchez-Quant *et al.* [164]

The experiments to generate the data for this project were performed by Eva Sánchez-Quant and Dr. Ioannis Deligiannis. In brief, as depicted in figure 4.1 A, cyropreserved primary human hepatocytes (PHHs) from four healthy human donors were cultured under different treatment conditions. To model hepatic steatosis, cells from the same donor were split into two groups and incubated with DMSO or free fatty acids (FFA), respectively. Furthermore, the cells were treated with a cocktail composed of five drugs, as will be described in greater detail below. For the assessment of the metabolic capacity of PHHs under different conditions. Thus, there are four conditions in the experimental design: (i) untreated (DMSO), (ii) incubated with FFA without cocktail treatment (FFA), (iii) treated with the five-drug cocktail (Cocktail), and (iv) incubated with FFA and the cocktail (FFA+Cocktail).

4.2 Characterization of PHH heterogeneity

Analysis of every donor individually revealed cellular heterogeneity in all four human donors that was independent of treatment condition. After determining that transcriptionally similar subgroups were present in all donors, the cells from all four donors were computationally combined by applying *Harmony*. Four subgroups of PHHs were identified (Methods). Figure 4.2 A depicts the four hepatocyte subgroups that were detected independently of the treatment condition in all four human donors. To exclude potential contamination of other liver cell types leading to subgroup detection, the expression of key hepatocyte marker genes was explored showing that all subgroups expressed hepatocyte marker genes. This analysis indicated no contamination by other cell types (Figure 4.2 B, left).

Studies on the opportunities and limitations of hepatocytes in culture as a model to assess drug metabolism have shown that PHHs tend to undergo loss of expression along culture time [114][115][165]. This leads to PHHs in culture gradually losing their characteristic hepatocyte-like transcriptomic phenotype [95]. From single cell transcriptomic data, this can be shown by the percentage of cells in a subgroup expressing a given gene. The idea behind this is that loss of expression would lead to more stochastic gene expression, i.e. genes would be expressed in fewer cells within a group of cells that is undergoing dedifferentiation. As shown in figure 4.2 C, genes in subgroup IV were expressed in a lower percentage of cells in comparison to the other three subgroups. Subgroup IV was thus labeled as the group of PHHs losing their expression whereas the other three subgroups were defined as metabolically active.

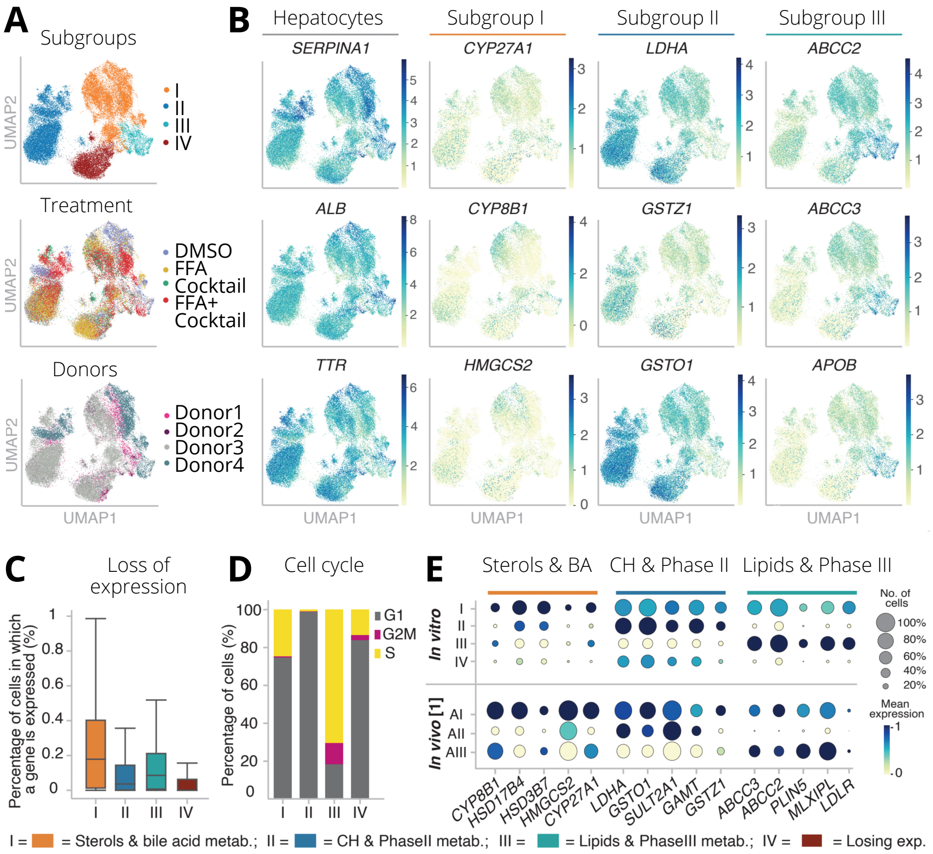

Figure 4.2: A UMAPs of the *in vitro* data set, colored by subgroups (top), treatment conditions (middle), and donors (bottom). **B** UMAPs of the *in vitro* data set, colored by marker gene expression of key hepatocyte markers (first column), marker genes for subgroup I (second column), subgroup II (third column), and subgroup III (fourth column). **C** Box plot showing the percentage of cells in which a given gene expressed for the four identified subgroups. **D** Bar plot for the percentage of cells in each subgroup computationally assigned to the different cell cycle phases G1, S, and G2M using *cyclone*. **D** Dot plot showing representative marker genes used to characterize the three metabolically active subgroups *in vitro*. Scaled mean expression is shown for the four subgroups *in vitro* (top) and for similar subgroups identified in a human *in vivo* data set from Aizarani *et al.* (bottom).
Figure adapted from Sanchez-Quant *et al.* [164]

Apart from identifying the cells losing their characteristic expression profile, cell cycle analysis was performed using *cyclone* [144]. The majority of hepatocytes reside in the liver in a non-divisive, quiescent state with an estimated turnover of 1 in 10,000 to 40,000 [145][146]. Therefore, only a small percentage of cells is expected to be in division. Nevertheless, this analysis was performed to check whether cell division contributes to the separation into subgroups. While indeed the majority of hepatocytes were computationally assigned to G1/G0 phase, subgroup III showed an enrichment of cells in S- and G2M-phase (Figure 4.2 D). To further investigate the functional specialization of each of the metabolically active subgroups, known marker genes for metabolic pathways were used. The list used was compiled from literature by Eva Sánchez-Quant and is available at the publication associated to this project [164]. Figure 4.2 E illustrates representative marker genes for each of the three metabolically active subgroups. Subgroup I was found to involve in the metabolism of bile acids and sterols, marked by the expression of genes, such as CYP8B1, CYP27A1, HSD17B4, HSD3B7, and HMGCS2. Subgroup II was characterized by carbohydrate and phase II metabolism, indicated by the expression of LDHA, GSTO1, SULT2A1, GAMT, and GSTZ1. Apart from the enrichment of S-phase associated cells, subgroup III was characterized by lipid and phase III metabolism, represented by expression of ABCC2, ABCC3, PLIN5, MLXIPL, and LDLR.

4.2.1 Zonation-independent heterogeneity *in vivo*

The presence of these three hepatocyte subgroups was further confirmed in publicly available *in vivo* data sets from human livers. Aizarani *et al.* have performed a study in which they analyzed the single cell profiles in the *in vivo* liver from nine human donors [19]. These donors were a mixture of male and female patients that covered a similar age range (34-77 years old) as the donors, from which the *in vitro* PHHs were taken (26-57 years old) [19]. Figure 4.2 E shows that after normalization and clustering of this data set (Methods), three subgroups were identified with similar gene expression profiles as the metabolically active subgroups in the *in vitro* data set. For this analysis, the marker gene expression was compared between *in vivo* clusters and *in vitro* subgroups (Figure 4.2 E).

As mentioned in the beginning of this chapter, the standard 2D culture model used for this project does not retain zonation patterns [95]. However, analyzing *in vivo* data allows exploring the relationship between the identified hepatocyte subgroups and zonation. In a first step to do so, it was hence assessed whether zonation could still play a role in the identification of subgroups *in vitro*. The PHHs were therefore scored for their expression of zonation marker genes, taken from Aizarani *et al.* [19]. As described in the method section of this thesis, cells were assigned to CV, mid-zone, or PV based on their zonation marker scores. Briefly, cells with simultaneous high CV and low PV scores were assigned to CV, and vice versa, whereas cells with moderate CV and PV scores were assigned to mid-zone (Methods). It was found that, *in vitro*, no particular subgroup showed enrichment of either pericentral (CV), periportal (PV), or mid-zonal expression profiles (Figure 4.3 A-C). This indicates that subgroups of PHH do not show zonation patterns after 72h in 2D culture.

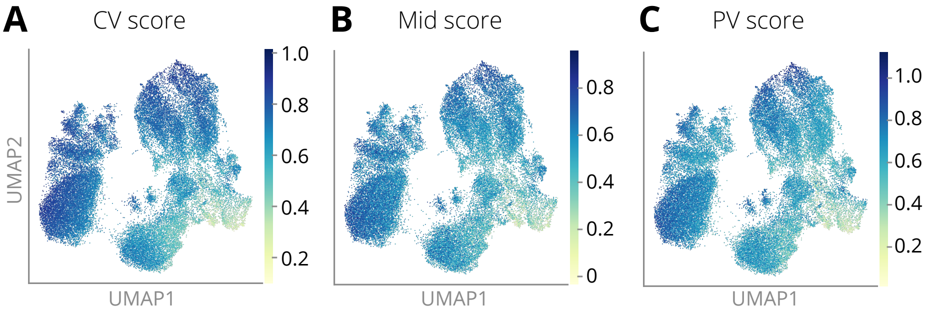

Figure 4.3: UMAPs depicting the *in vitro* data set, colored by **A** pericentral score, **B** mid-zonal score, and **C** periportal score.

To further explore this observation in a setting where zonation was present, the *in vivo* data from Aizarani *et al.* was again used. Figure 4.4 A-C illustrates that distinct patterns were detected *in vivo* for CV, mid-zonal, and PV marker genes, whereas these patterns were not found *in vitro*. Leveraging the information about the identified subgroups on the *in vivo* UMAP, it can be observed that the gene expression profiles of individual cells within subgroups are affected by zonation (Figure 4.4 D-E). For example, cells in subgroup I showed mutually exclusive high scores for either CV or PV markers. Calculating the percentage of cells assigned to each of these zones for the three subgroups revealed that, overall, the highest percentage of periportal-assigned cells was observed for subgroup I. In general, all subgroups were present in all three zones, with subgroup II having the highest percentage of mid-zone assigned cells and cells in subgroup III being rather assigned to mid- or CV than to PV (Figure 4.4 F).

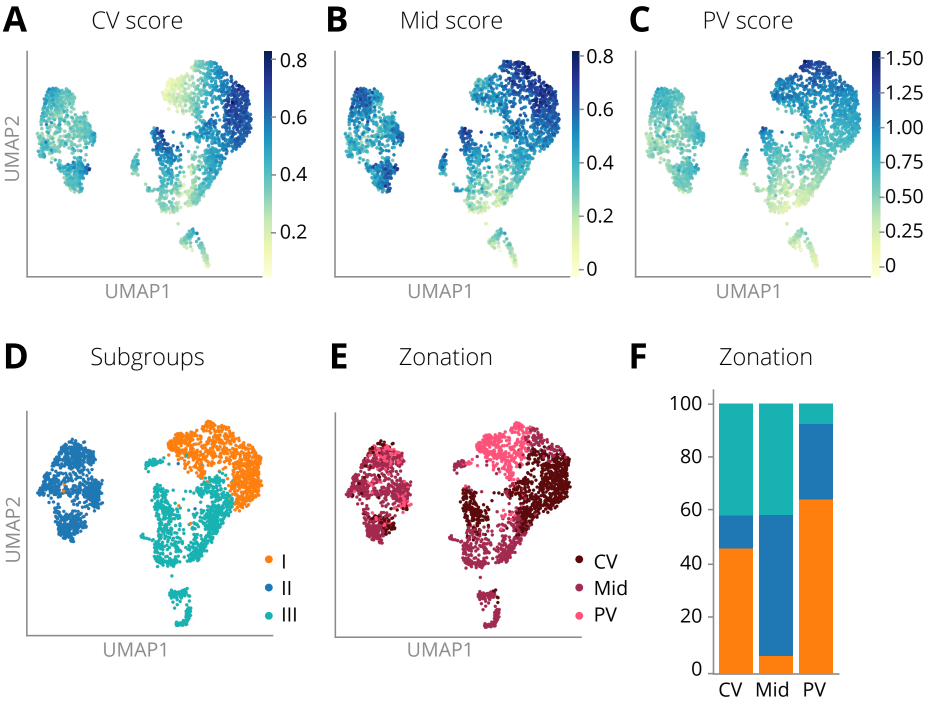

Figure 4.4: A-E UMAPs showing the Aizarani *in vivo* data set, colored by pericentral score (**A**), mid-zonal score (**B**), periportal score (**C**), subgroup annotation based on the findings *in vitro* (**D**), and assigned zones based on the scores from **A-C** (**E**). **F** Bar plot depicting the percentage of cells belonging to a given subgroup in each of the assigned zones (CV, Mid, PV).
Figure adapted from Sanchez-Quant *et al.* [122]

4.2.2 Comparison of *in vitro* and *in vivo*

After establishing the relationship between zonation and subgroup formation, the findings were further validated by assessing another *in vivo* data set from MacParland *et al.* [18]. Their study is comprised of single cell transcriptomic data from the livers of five deceased human donors (4 male, 1 female), covering a donor age range of 21 to 65 years. They report the presence of six distinct clusters of hepatocytes in the liver that could partially be linked to zonation [18]. To further investigate the role of zonation in the subgroups, the gene expression profiles of those *in vivo* clusters were compared to the three metabolically active subgroups identified *in vitro*. Figure 4.5 A shows the percentage of overlaps for the top 500 genes per subgroup between the *in vitro* subgroups and the *in vivo* clusters from MacParland *et al.*.

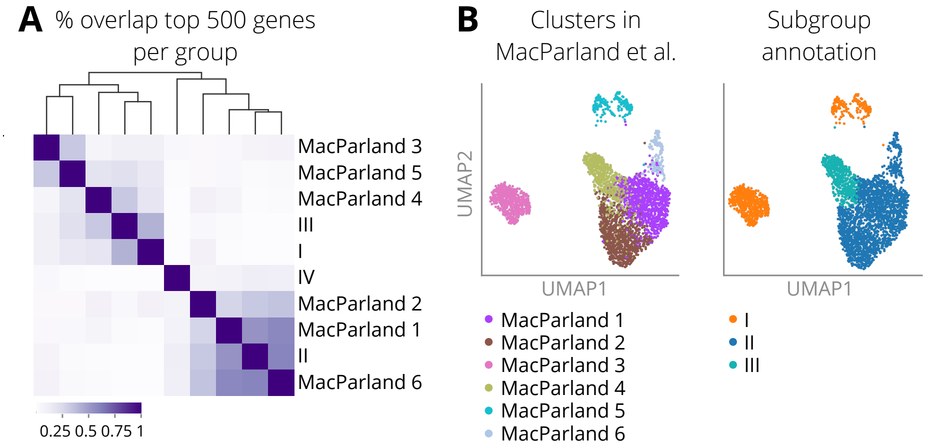

Figure 4.5: A Heatmap showing the percentage of overlap for the top 500 genes per *in vitro* subgroups I-III and the six *in vivo* clusters reporter in MacParland *et al.* [18]. **B** UMAPs showing the MacParland *in vivo* data set, colored by their reported cluster annotation (left), and the identified subgroups based on the *in vitro* findings (right).

Especially *in vivo* clusters 1, 2, and 6 showed high gene expression similarity to subgroup II. *In vitro* subgroups I and III showed higher similarity to each other than to any of the *in vivo* clusters, although subgroup III also had high gene expression similarity to cluster 4. Therefore, clusters 1, 2, and 6 were labeled as subgroup II; cluster 4 was labeled as subgroup III; and clusters 3 and 5 were labeled as subgroup I since they expressed some marker genes of subgroup I (Figure 4.5 B)[18]. The split between cluster 3 and 5 in this *in vivo* data set could be due to inter-donor heterogeneity as the majority of cells in cluster 3 were coming from one donor [18]. Together with the general differences between *in* *vivo* and the 2D *in vitro* culture of PHHs, this observation could partially explain the low similarity to *in vitro* subgroup I for these clusters.

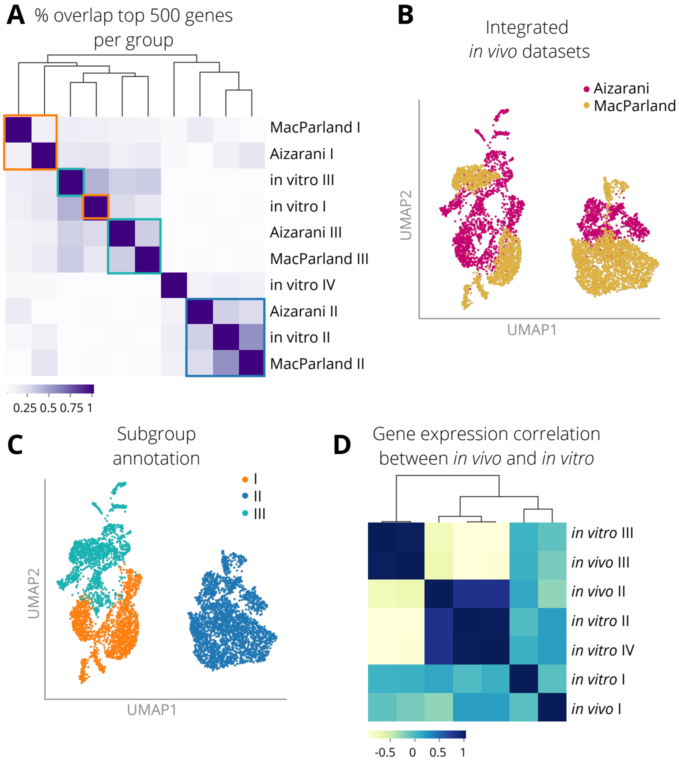

Figure 4.6: **A** Heatmap depicting the percentage of overlap between the top 500 genes per group in the two *in vivo* and the *in vitro* data after annotating the subgroups separately for the two *in vivo* data sets based on the findings *in vitro*. **B-C** UMAP of the two *in vivo* data sets after data integration using *scGen*, colored by data set (**B**) and annotated subgroup (**C**). **D** Heatmap showing the correlation of gene expression for the top 10 DEGs per subgroup between the integrated *in vivo* and the *in vitro* data.
Figure adapted and expanded from Sanchez-Quant *et al.* [122]

Combining both *in vivo* liver data sets, figure 4.6 A shows the percentage of overlap between the top 500 genes per subgroup for all three data sets (two *in vivo* and one *in vivo*). Overall, the *in vitro* subgroup II showed high gene expression similarity with the corresponding *in vivo* clusters (Figure 4.6 A). In contrast, subgroup I and III were more similar to each other within the two *in vivo* than to the *in vitro* subgroup III cells. Increasing the number of cells in an analysis has been associated to higher statistical power to draw robust conclusions [47][48]. Hence, to increase the power for the comparison between *in vitro* and *in vivo* while also decreasing the impact of inter-donor heterogeneity, the two publicly available *in vivo* data sets were computationally combined as described in the methods section. In brief, subgroups were annotated separately for the two *in vivo* data sets and *scGen* was used for data integration. The common embedding for both *in vivo* data sets, colored by subgroup annotation, can be seen in Figure 4.6 C. In line with the previous findings, cells in subgroup II formed a tightly integrated cluster, which located separately on the UMAP, whereas subgroup I and III were closer to each other (Figure 4.6 C). After the integration using *scGen*, the correlation of expression between *in vitro* and *in vivo* was calculated for the top 10 DEGs per subgroup (Figure 4.6 D). Despite the low percentage of shared genes previously observed between *in vitro* and *in vivo* for subgroup III (Figure 4.6 A), the top 10 genes were highly correlated between *in vivo* and *in vitro* (Figure 4.6 D).

Moreover, the three metabolically active subgroups annotated in the integrated *in vivo* data sets showed high relative expression of the marker genes that were used for the initial subgroup characterization *in vitro* (Figure 4.7 A). This indicated that the three metabolically active subgroups indeed showed similar functional specialization *in vitro* and *in vivo*.

Additionally, integrating the two *in vivo* data sets led to more power in analyzing the relationship between subgroups and zonation. To do so, the cells were ordered based on the vector of diffusion pseudospace starting at the cell with the lowest CV score per subgroup (Figure 4.7 B). In the integrated *in vivo* subgroup I, periportal marker genes, such as HMGCS2, decreased in their mean expression along the pseudospace vector whereas pericentral marker genes, such as CYP8B1, increased. The same trend could be observed in subgroup III. In subgroup II, however, gene expression generally decreased along the vector of pseudospace (Figure 4.7 B, middle). This indicated that for subgroup II, gene expression was not following zonation patterns as the pseudotime algorithm instead ordered cells by overall expression level, regardless of the marker gene. Overall, the subgroups identified *in vitro* could successfully be identified *in vivo*. While zonation patterns where absent *in vitro*, marker gene expression in subgroups of hepatocytes was observed to be influenced by zonation *in vivo*.

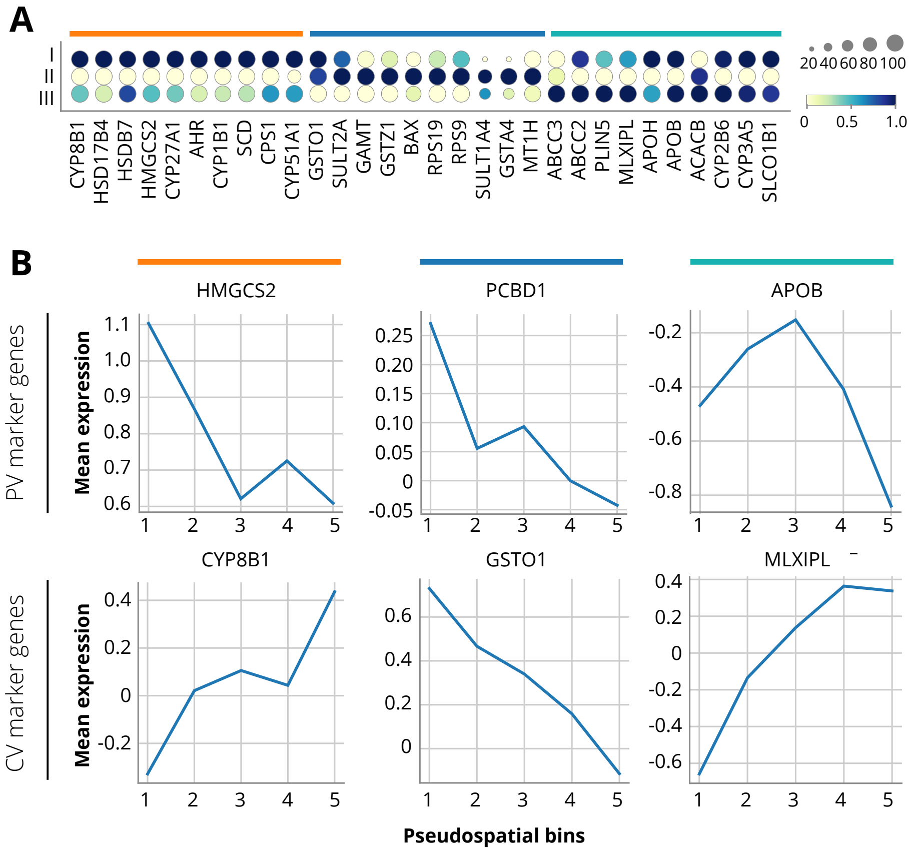

Figure 4.7: A Dot plot showing relative expression levels of the marker genes used to characterize the subgroups *in vitro* on the integrated *in vivo* data set. Dot size corresponds to percentage of cells in which a gene is expressed, color refers to scaled mean expression level. **B** Line plots showing the mean expression for selected zonated subgroup marker genes. Columns correspond to subgroups identified in the integrated *in vivo* data set, top row shows representative periportal marker genes, bottom row shows representative pericentral marker genes.

4.2.3 Lipid accumulation affects cellular heterogeneity in PHHs

After computationally confirming the presence of the *in vitro* subgroups in two *in vivo* data sets, the next question to answer was how this cellular heterogeneity is affected by internal and external factors. During the span of life, hepatocytes can accumulate lipids intracellularly, referred to as hepatic steatosis and a hallmark of NAFLD [166]. *In vivo* studies using single cell transcriptomics in mouse models have shown that hepatic steatosis changes the composition and transcriptomic profile of non-parenchymal cells [27][26]. For instance, hepatic steatosis has been observed to lead to higher infiltration of immune cells in the liver and a transcriptomic shift towards immune response pathways in several cell types, yet in-depth studies on hepatocytes remain challenging due to the difficulty of isolating fresh and intact hepatocytes from the liver tissue [27][26]. Bulk studies of hepatocytes have reported that lipid accumulation can lead to disruptions in lipid metabolism and increased chemokine production [166][167]. At single cell resolution, it has been shown that periportal hepatocytes down-regulate their expression profiles upon intracellular lipid accumulation [7]. For this project, the question was how different subgroups of PHHs in culture respond to the intracellular lipid accumulation in the absence of zonation. Apart from calculating changes in gene expression, cellular heterogeneity can also be measured in transcriptional variability. In general, higher transcriptional variability is associated to less coordinated responses to a stimulus [132]. As described for the snRNA-seq2 data set in the previous chapter, the coefficient of variation can be used to estimate transcriptional variability. Figure 4.8 A shows the comparison of the coefficient of variation between FFA- and untreated cells per subgroup. Overall, cells that were losing their characteristic expression in culture (subgroup IV) had the highest transcriptional variability regardless of their treatment with FFA. Accumulation of FFA led to increased transcriptional variability in subgroups I and II, whereas a significant decrease in transcriptional variability was observed for subgroup III. Taking into account that subgroup III is characterized by genes involved in lipid metabolism, the decrease in transcriptional variability in this subgroup suggests that these cells show a more coordinated gene expression in response to lipid accumulation. Aside from FFA accumulation changing the transcriptional variability in the subgroups, the proportion of treated and untreated cells captured differed between subgroups. The proportion of FFA-treated cells was particularly high for subgroup IV, suggesting that FFA accumulation could lead to increased loss of expression in culture (Figure 4.8 B). The ratio of palmitic to oleic acid used to model steatosis here has been shown to yield only minor lipotoxic and apoptotic effects [96]. Nevertheless, FFA treatment in general has been reported to trigger apoptosis, hence providing a potential mechanism how lipid accumulation could increase loss of expression in culture [168].

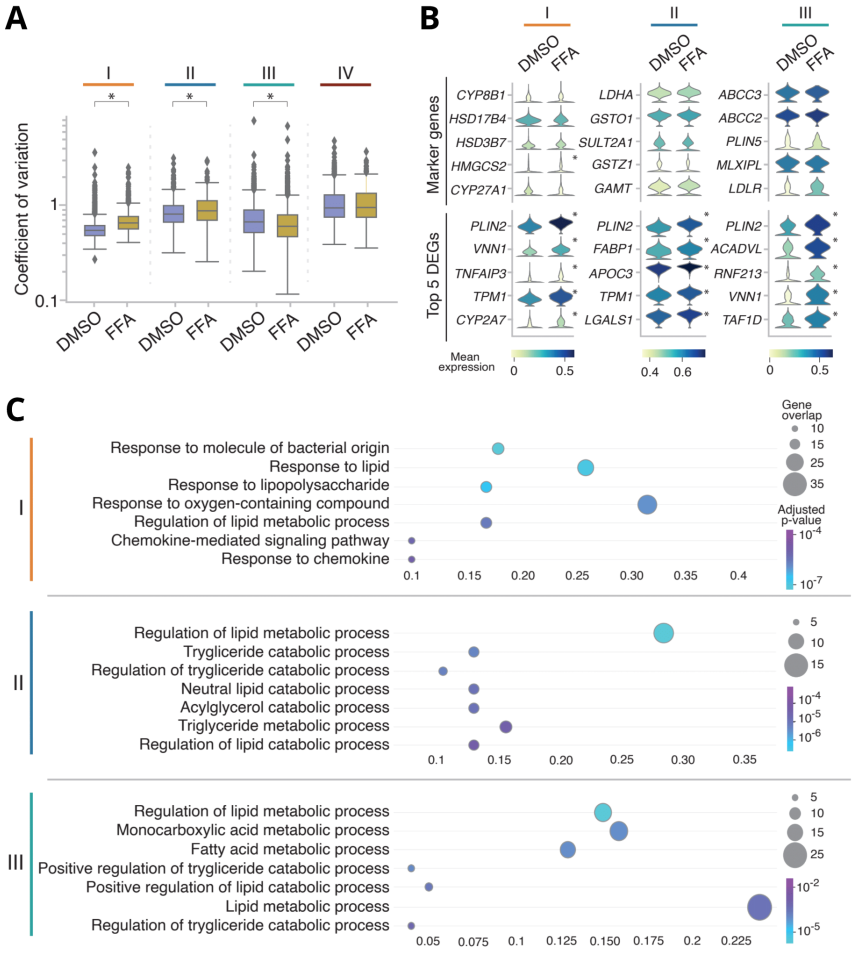

Figure 4.8: A Box plots showing the coefficient of variation for FFA- and DMSO treated cells in every subgroup (*=Bonferroni-adjusted p-value < 0.05, *MannWhitneyU*-test). **B** Violin plots depicting i) the expression levels of the marker genes used for subgroup characterization in figure 4.2 D (top), and ii) the top 5 DEGs between DMSO and FFA for each subgroup (bottom, *=Bonferroni-adjusted p-value < 0.05). **C** Scatter plot of the top 7 GO pathways the DEGs between FFA and DMSO in each metabolically active subgroup are enriched in. Figure adapted from Sanchez-Quant *et al.* [164]

To therefore answer the question whether intracellular lipid accumulation can influence the transcriptomic profiles of PHH subgroups, the expression of the subgroup-specific metabolic marker genes was investigated comparing untreated to FFA-treated cells (Figure 4.8 C, top). Among the representative markers only HMGCS2, a marker for bile metabolism used to characterize subgroup I, was significantly up-regulated upon accumulation of FFA. In the presence of intracellular lipids, in all subgroups perilipid protein 2 (PLIN2) was found among the top 5 significantly up-regulated genes. PLIN2 involved in the coating and storage of lipid droplets, therefore suggesting that the storage of lipids in droplets could be performed similarly across hepatocyte subgroups. Moreover, the top five DEGs per subgroup were mostly genes associated to either the metabolism of lipids, or the inflammatory stress response associated to the progression of NAFLD (Figure 4.8 C, bottom). For example, FABP1 in subgroup II, and ACADVL in subgroup III are both involved in the oxidation of fatty acids [169][170]. In contrast, TNFAIP3 in subgroup I, and LGALS1 in subgroup II have been shown to be up-regulated in NAFLD [171][172]. The expression of genes known to be involved in the different steps of lipid clearance were investigated per subgroup between untreated and FFA-treated cells (Figure 4.8 D). Furthermore, since hepatic steatosis leads to the progression of NAFLD mainly through inflammation, genes associated to inflammation and cellular stress were also investigated [166].

With the aim to functionally characterize the transcriptomic differences upon FFA in each of the subgroups, gene ontology analysis for biological processes was performed on the DEGs. As illustrated in figure 4.8 D, subgroup I up-regulated pathways related to chemokine signaling, indicating the potential relevance of this subgroup in the progression of NAFLD. Subgroup II up-regulated pathways related to the clearance of neutral lipids and involved in triglyceride metabolism. The pathways of lipid and carbohydrate metabolism have been shown to be intertwined, playing a role in type II diabetes [173]. These results hence further support the evidence that subgroup II's response stem from the initial involvement in carbohydrate metabolism. Finally, subgroup III up-regulated pathways involved in the lipid and fatty acid metabolism, again indicating a more coordinated response towards the intracellular lipid accumulation in comparison to the other subgroups.

4.2.4 PHH heterogeneity influences drug-metabolic capacity *in vitro*

Altered lipid metabolism has furthermore been shown to be associated to changes in the cytochrome P450 pathways [174][9]. Despite their tendency to undergo de-differentiation in culture, PHHs have been shown to express the majority of drug-metabolising enzymes, respond to enzyme inducers, and produce a metabolic profile similar to *in vivo* [95]. Moreover, the PHHs used in this thesis were certified by the supplier Lonza to be plateable and characterized for transporter and enzyme activity as well as their induction potential. As mentioned in the introduction of this thesis, PHHs are considered the gold standard model to study drug metabolism and potential drug-drug interactions *in vitro*. Generally, applying a drug cocktail allows to study the activity of several cytochromes simultaneously [110][112]. The Sanofi-Aventis cocktail used here for the assessment of drug metabolism in PHHs has been shown to be safe for *in vivo* studies and no pharmacokinetical interactions within the cocktail have been observed in the original study [110]. The compounds of this cocktail are selective for the five cytochromes, CYP1A2 (caffeine), CYP2C9 (S-warfarin), CYP2C19 (omeprazole), CYP2D6 (metoprolol), and CYP3A4 (midazolam). More recently, CYP2B6, CYP2C9 and CYP3A4 have been reported to contribute to the metabolism of metoprolol *in vitro*, but their contribution was concluded to be minor and not compromising the appropriateness of metoprolol as phenotyping probe for CYP2D6 [175]. Furthermore, the Basel cocktail using the same compounds for CYP1A2, CYP2C19, CYP2D6, and CYP3A4 (but losartam for CYP2C9 instead of S-warfarin) has been shown to be suitable for assessing metabolic capacity of PHHs in culture [112]. The expression level of cytochromes can be used to phenotype the metabolic activity of hepatocytes [97]. Here, in addition to focusing on the targeted cytochromes, measuring the whole transcriptome allows for in-depth characterization of the hepatocyte response to a metabolic challenge at single cell resolution.

To contrast which information is obtained through a single cell study in comparison to bulk, the transcriptomic response to the Sanofi-Aventis drug cocktail was first measured at the pseudobulk level. Figure 4.9 A, left, shows the expression levels of the five targeted cytochromes, averaged across all cells, in response to the cocktail in comparison to DMSO.

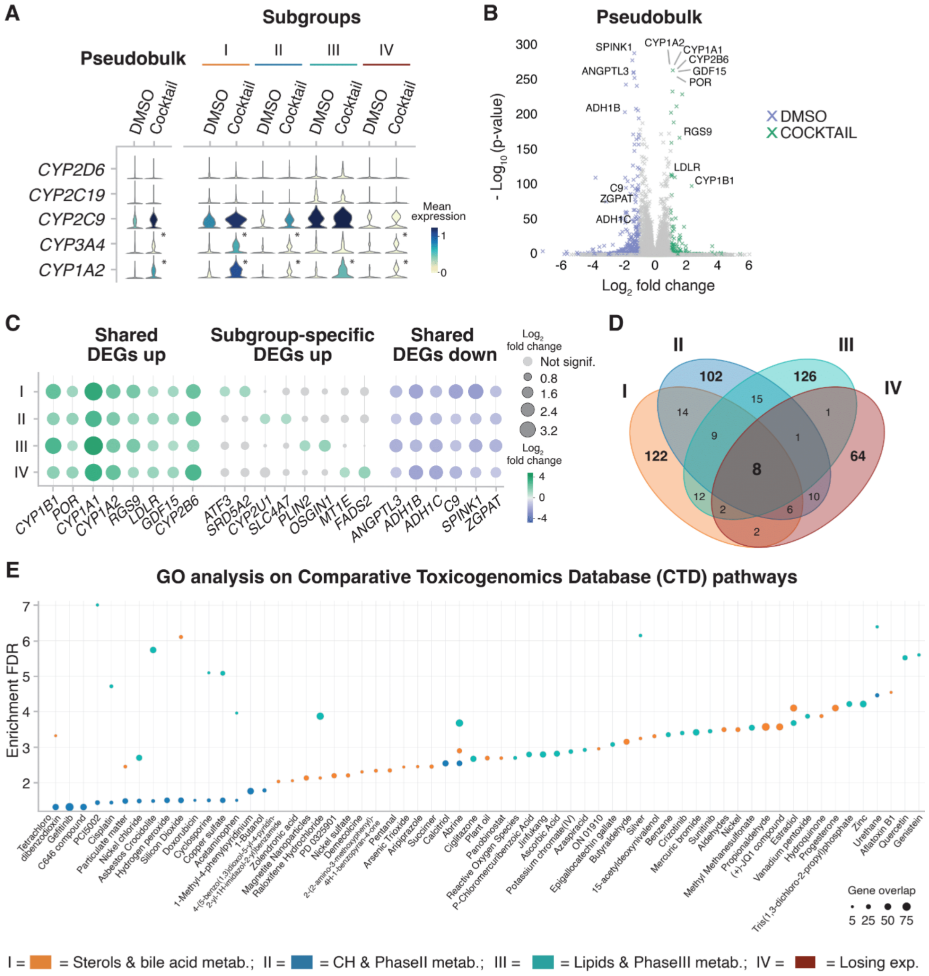

Figure 4.9: A Stacked violin plots showing the expression levels of the targeted five cytochromes in DMSO versus cocktail treatment for all cells (Pseudobulk, left) and separately for each subgroup (Subgroups, right) (*=Bonferroni-adjusted p-value < 0.05, |log$_2$-fold change| > 1, t-test). **B** Volcano plot depicting significance level (-log$_1$0(p-value)) against effect size (log$_2$-fold change) for the comparison between cocktail and DMSO in pseudobulk (green=up-regulated in cocktail, purple=up-regulated in DMSO, labels highlight genes up-regulated in all subgroups). **C** Dot plot showing per subgroup, the log$_2$-fold change (dot size and color intensity) between cocktail (green) and DMSO (purple) for genes that are significantly i) up-regulated in all subgroups (left), ii) up-regulated specifically in only one subgroup (middle), and iii) down-regulated in all subgroups (right). Grey indicates no significance. **D** Venn diagram showing the overlap between genes that are significantly up-regulated upon cocktail treatment between subgroups. **E** Scatter plot showing which drug-metabolism related pathways (CTD database) genes are enriched in that are specifically up-regulated in a given subgroup (color) upon cocktail treatment. Dot size represents number of genes overlapping in a given pathway. Figure taken from Sanchez-Quant *et al.* [164]

Significant differential expression of a gene was determined by a p-value below 0.05 and an absolute effect size ($\log_2$-fold change) of greater than 1. It has previously been reported that metoprolol can inhibit the expression of CYP2D6 and CYP3A4, but not affecting the metabolism of midazolam by CYP3A4 [176]. Furthermore, omeprazole can inhibit expression of CYP2C19 [177]. However, no down-regulation of any of the five targeted cytochromes was observable at both, the pseudobulk and individual subgroup level, indicating no inhibitory effects of the compounds at the transcriptome level in the studied PHHs (Figure 4.9 A). While the expression levels of CYP2D6 and CYP2C19 were comparable to baseline DMSO-levels, CYP3A4 and CYP1A2 were significantly up-regulated in the pseudobulk of all cells in response to cocktail. CYP2C9 was also up-regulated but had a low $\log_2$-fold change towards DMSO. Additionally, dissecting the effect of the cocktail per identified subgroup revealed subgroup-specific effects for the expression of the five cytochromes in response to the cocktail (Figure 4.9 A, right). For instance, the expression level of CYP3A4 was increased 3.4-fold in the pseudobulk of all cells upon cocktail treatment. Specifically in subgroup III, however, no significant up-regulation was observed for this gene in response to the cocktail whereas an 4-fold increase was observed in subgroup I.

Expanding the analysis from the five cytochromes mediating the metabolism of the five compounds of the cocktail, the effect of the cocktail on the whole transcriptome was analyzed. In the pseudobulk, the cocktail led to down-regulation of genes associated to alcohol metabolism, ADH1B and ADH1C (Figure 4.9 B). These genes were among six genes that were consistently down-regulated across all four subgroups (Figure 4.9 C, right). On the other hand, the cocktail led to increased expression of phase I drug-metabolism related genes, such as CYP1A2, CYP1A1, CYP1B1, and CYP2B6 as well as genes related to stress-response, such as GDF15 and RGS9 (Figure 4.9 B). Overall, eight genes were consistently up-regulated upon cocktail treatment across all four subgroups (Figure 4.9 C, left). However, each subgroup showed subgroup-specific up-regulation of a list of genes upon cocktail treatment. For example, ATF3 and SRD5A2 were only up-regulated in subgroup I, CYP2U1 and SLC4A7 were only up-regulated in subgroup II, PLIN2 and OSGIN1 were only up-regulated in subgroup III, and MT1E and FADS were only up-regulated in subgroup IV (Figure 4.9 C, middle). These representative genes again reflect the unique metabolic profiles of the individual subgroups. For instance, PLIN2 and OSGIN1 are genes involved in the coating and metabolism of lipids, indicating that hepatocytes maintain their specialization and respond accordingly in a metabolic challenge.

Overall, the number of specifically up-regulated genes in response to the cocktail was similar between the three metabolically active subgroups and was 122 for subgroup I, 102 for subgroup II, and 126 for subgroup III. In line with the loss of characteristic expression, subgroup IV showed the least amount of specifically up-regulated genes, 64 (Figure 4.9 D). The transcriptomic differences observed between subgroups in the response to the five-drug cocktail suggests subgroup-specific pathway expression under metabolic challenges. To therefore measure this in an unsupervised manner, gene ontology analysis was performed. Instead of using the pathways related to biological processes that such analyses are usually performed on, a database was used containing pathways related to metabolism of xenobiotic compounds (Comparative Toxicogenomics Database, CTD) [178][179]. This led to the observation that, in response to the cocktail, the three metabolically active subgroups indeed expressed pathways related to the metabolism of different compounds in a subgroup-specific manner. For example, cells in subgroup I were the only ones up-regulating the pathways responsible for the metabolism of progesterone and plant oil, among others. Furthermore, specifically for subgroup II the pathways for Calcitriol and Cisplatin were up-regulated. Finally, only in subgroup III the pathways for Quercetin and Reactive Oxygen Species (ROS) were up-regulated (Figure 4.9 E).

4.2.5 Intracellular lipid accumulation diminishes drug-metabolism in a subgroup-specific manner

Changes in pathways related to drug metabolism, particularly in the cytochrome P450 pathway, have been observed in response to changes in lipid metabolism [174][9][180]. Furthermore, hepatic steatosis observed in NAFLD can increase drug-induced liver injury (DILI) [180]. Likewise, the co-administration of several drugs simultaneously also leads to a higher incidence of DILI [181][182]. Hence, combining and expanding the findings from the previous two sections of this thesis, the impact of intracellular lipid accumulation was assessed on the drug-metabolism related pathways. As a first step, the expression level of the five cytochromes targeted by the Sanofi-Aventis cocktail was measured in PHHs either treated with the cocktail alone or treated with the cocktail after incubation with FFA. In every of the characterized PHH subgroups, the $\log_2$-fold change of the expression towards baseline (DMSO) levels was measured for each targeted cytochrome. Figure 4.10 A shows that across all subgroups, the expression of the five cytochromes was generally lower in cells that were incubated with FFA before cocktail administration. For example, in subgroup I, CYP3A4 and CYP1A2 were significantly up-regulated upon cocktail incubation in otherwise untreated cells. However, when the cells were incubated with FFA before cocktail treatment, only CYP1A2 was still significantly up-regulated. Specifically in subgroup III, CYP2D6, CYP2C19, CYP2C9, and CYP3A4 were significantly down-regulated towards DMSO levels when cells were treated with the cocktail after FFA-incubation.

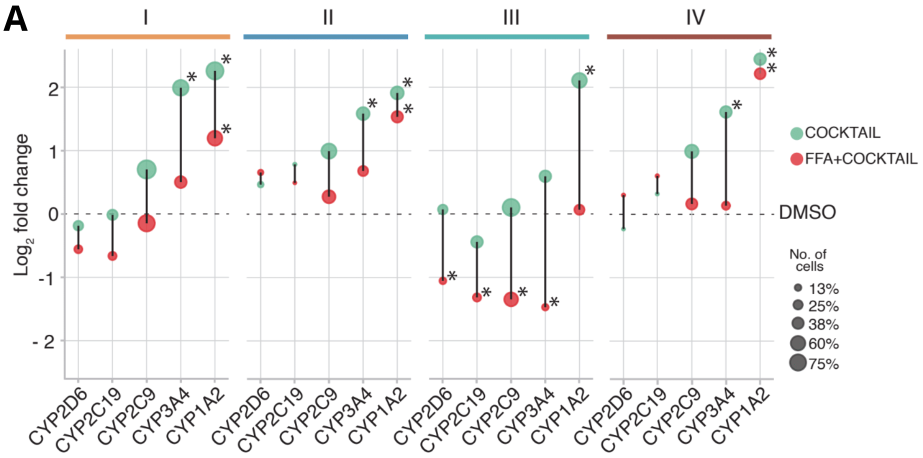

Figure 4.10: A Dumbbell plot depicting in each subgroup the $\log_2$-fold change of each of the targeted cytochromes to DMSO baseline levels. Colors refer to treatment condition (green=cocktail, red=FFA+cocktail), dot size represents percentage of cells in which gene is expressed. (*=Bonferroni-adjusted p-value <0.05 and $|\log_2$-fold change$| > 1$, t-test). Figure adapted from Sanchez-Quant *et al.* [164]

Measuring the whole transcriptome allows the quantification and characterization of which genes and pathways are differentially expressed in response to cocktail and in the response to FFA+cocktail, and their overlap (Figure 4.11 A). Of the 498 genes up-regulated across all cells upon cocktail treatment, 234 were also up-regulated in FFA+cocktail, whereas 264 were solely up-regulated under cocktail treatment. Likewise, 602 genes were specifically up-regulated when the cells were treated with FFA+cocktail. Gene ontology analysis in showed that the genes specifically up-regulated in cocktail-treated cells (but not in FFA+cocktail treated cells) were enriched in pathways responsible for xenobiotic metabolism (Figure 4.11 B, top, green). Genes that were up-regulated in both, cocktail and FFA+cocktail conditions, were enriched in pathways related to the response to stimuli, but less specific for xenobiotic metabolism (Figure 4.11 B, middle, beige). Finally, the genes that were specifically up-regulated in FFA+cocktail treated cells were enriched in pathways related to stress response and general biological processes (Figure 4.11 B, bottom, red). Subgroup I showed the highest percentage of genes specifically up-regulated in cocktail whereas subgroup II showed the highest percentage of genes specifically up-regulated in FFA+cocktail. Overall, similar percentages were observed for the genes in those categories for all subgroups (Figure 4.11 C). This indicated that intracellular lipid accumulation affected the expression of drug-related genes to a similar extent in all subgroups.

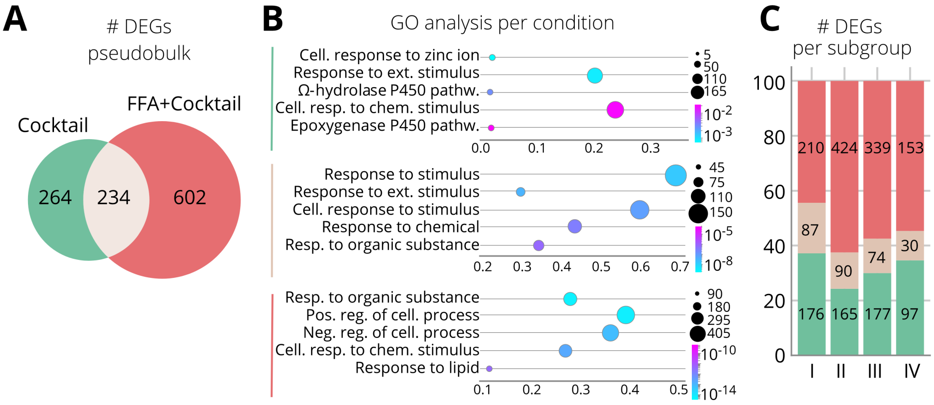

Figure 4.11: **A** Venn diagram depicting the overlap of genes up-regulated in cocktail and FFA+cocktail treatment across all PHHs. **B** Scatter plot showing the GO pathways the genes are enriched in that are i) specifically up-regulated in cocktail (top, green), ii) up-regulated both in cocktail and FFA+cocktail (middle, beige), iii) specifically up-regulated in FFA+cocktail (bottom, red). **C** Bar plot showing the percentage of genes in each category for all four PHH subgroups.
Figure adapted from Sanchez-Quant *et al.* [164]

Finally, it was investigated how lipid accumulation changes the transcriptional landscape in response to a five-drug cocktail. For this, gene set enrichment analysis (GSEA) was performed comparing the cells treated with FFA+cocktail to the cells treated with cocktail. The pathway for insulin resistance was among the pathways enriched in the cells treated with FFA+cocktail (Figure 4.12 A).

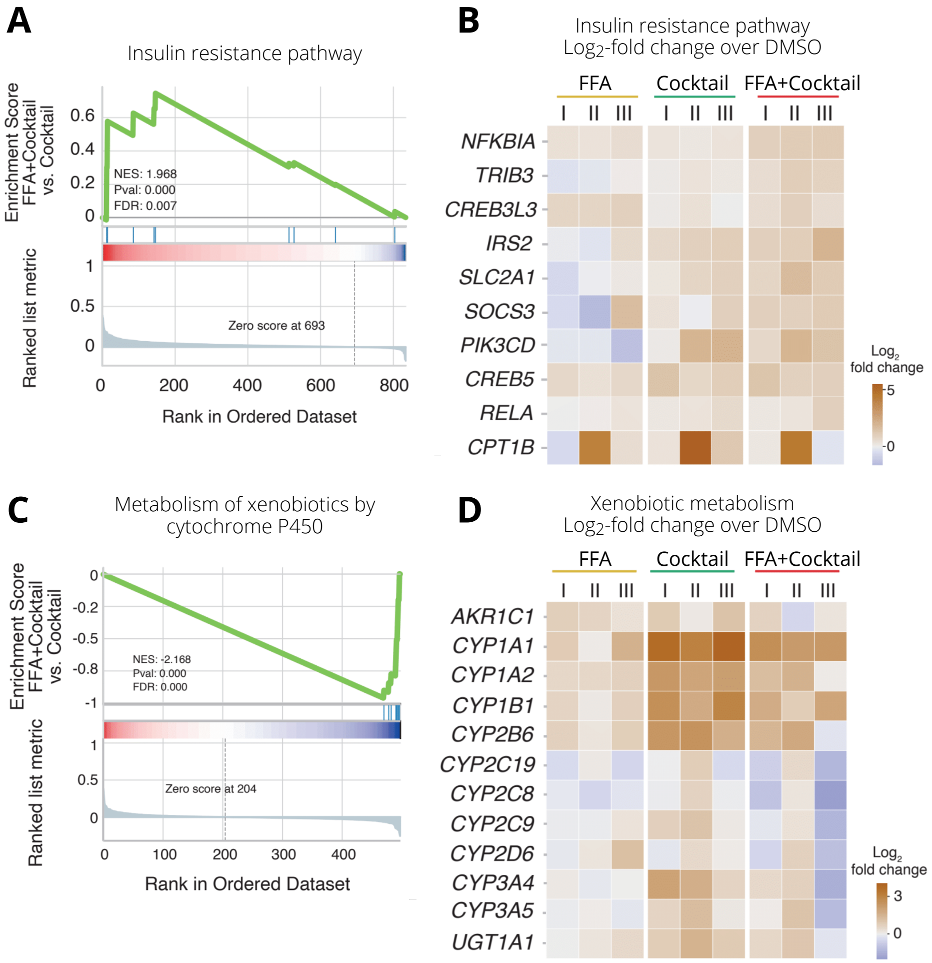

Figure 4.12: A GSEA plot for FFA+cocktail versus cocktail for the pathway "Insulin resistance pathway", that the genes specifically up-regulated in FFA+cocktail were enriched in. **B** Heatmap showing log$_2$-fold change towards DMSO levels for genes in the insulin resistance pathway. **C** GSEA plot for FFA+cocktail versus cocktail for the pathway "Metabolism of xenobitics by cytochrome P450", that the genes specifically up-regulated in cocktail were enriched in. **D** Heatmap showing log$_2$-fold change towards DMSO levels for genes related to drug-metabolic pathways.
Figure adapted from Sanchez-Quant *et al.* [164]

Measuring the $\log_2$-fold change towards baseline DMSO levels of the individual genes in this pathway revealed that incubation with either FFA or cocktail alone did not lead to high up-regulation of the genes in all subgroups (Figure 4.12 B). Only the combined treatment of FFA+cocktail led to consistent up-regulation of genes in the insulin resistance pathway. This indicated that, in lines with reported literature, hepatic steatosis can be linked with adverse drug reactions, for instance through insulin resistance [183][184][185]. Furthermore, the comparison of cocktail and FFA+Cocktail treated cells revealed that FFA+Cocktail led to down-regulation of the pathway "metabolism of xenobiotics by cytochrome P450" (Figure 4.12 C). In line with Aubert *et al.*, this provides evidence that intracellular lipid accumulation changes P450 expression and leads to diminished drug-metabolic capacity [180]. Investigating the $\log_2$-fold towards DMSO of genes in drug-related pathways showed that FFA treatment alone did not significantly change the expression of these genes, despite several cytochromes associated to both, FFA and drug metabolism [186]. Moreover, while cocktail treatment alone led to up-regulation of several drug-related genes, treatment with FFA+cocktail diminished the level of up-regulation across subgroups or even led to their down-regulation (Figure 4.12 D). Specifically for subgroup III, expression of CYP2C19, CYP2C8, CYP2D6, CYP3A4, and CYP3A5 was particularly diminished towards DMSO levels in FFA+Cocktail treated cells (Figure 4.12 D, right). These results indicate that intracellular lipid accumulation affects the gene expression network along multiple pathways. The combination of intracellular lipid accumulation and a five drug challenge diminished the drug-related response of PHHs and increased potentially cytotoxic cellular mechanisms, such as insulin resistance. Particularly subgroup III showed diminished drug-related metabolic capacity under FFA+Cocktail treatment.

Overall, this chapter shows that in the absence of zonation, previously unobserved cellular heterogeneity is retained in PHHs *in vitro* in the form of metabolically specialized hepatocyte subgroups. These subgroups respond differently to metabolic challenges, such as intracellular lipid accumulation and a five drug cocktail. When PHHs were treated with FFA and the cocktail, they increased expression of genes related to insulin resistance but decreased drug-related metabolism. This effect was particularly evident in subgroup III. In summary, these result indicate that the subgroup-composition of the tissue *in vivo* could influence the effect of hepatic steatosis, the drug metabolic capacity, and the risk of developing adverse drug reactions.

5 Discussion

5.1 Analysis of single cell RNA sequencing data

The technological advances of the last decade enable the in-depth characterization of cellular heterogeneity at single cell resolution [33][187][40]. Since their establishment, scRNA-seq experiments offer a powerful tool to study the transcriptome of individual cells [33]. However, computational analysis of the data from such an experiment needs to ensure that the obtained results reflect true biological differences between cells. Thus, the quality control measures and subsequent analysis decisions, e.g. data normalization and batch correction, need to be done carefully to reach meaningful conclusions on the respective biological question. Moreover, the biological questions explored in this thesis required different experimental approaches. The main goal of the Ploidy&Zonation project was to characterize transcriptomic differences between 2n and 4n hepatocytes in healthy adult mice, whereas the PHH diversity project aimed to study cellular heterogeneity of PHH in response to a stimulus. Depending on the experimental approach and chemistry used for library preparation, i.e., droplet-based or plate-based, the sequencing depth obtained per single nucleus or cell varies substantially [39][57]. To address the respective biological questions, the data presented in this thesis has been generated using two different experimental protocols, namely the plate-based method snRNA-seq2 developed for the Ploidy&Zonation project, and a droplet-based method by *10X Genomics* for the PHH diversity project. Plate-based methods are generally designed to achieve higher sequencing depths but comparatively low numbers of cells/nuclei whereas droplet-based methods yield a high throughput of single cells/nuclei at a shallower sequencing depth [57][56]. Because of the difference in cellular (or nuclear) throughput between these two approaches, close to 2,500 nuclei were analyzed in the plate-based Ploidy&Zonation project, and an overall amount of roughly 65,000 single cells in the droplet-based PHH diversity data set. Furthermore, the average number of reads differed at two orders of magnitudes, at around 7,000 per single cell in the PHH diversity project, and 700,000 per single nucleus in the

Ploidy&Zonation project. These results are in line with droplet-based methods yielding lower coverage than plate-based ones. These differences between a plate-based approach such as SMART-seq2 and a droplet-based approach such as 10X Genomics have also been reported to result in higher number of detected genes for SMART-seq2 and more cellular clusters in 10X [56].

In summary, choosing the SMART-seq2 protocol to study the impact of ploidy in the young adult mouse liver allowed the in-depth characterization of the two populations, including the detection of usually lowly abundant transcription factors and the comparison of isoform expression. Contrary to this, the high cell throughput achieved by a 10X Genomics experiment enabled the detection of hepatocyte subgroups.

Apart from the choice between plate- and droplet-based sequencing techniques, a limiting factor in experimental design is the quality and quantity of the input material that can be obtained. When intact single cells are obtained, it allows the analysis of reads stemming from the nucleus, cytoplasmic reads, and mitochondrial reads. The quantification of mitochondrial reads is often used for quality control as under stressful conditions, cells have been observed to up-regulate the mitochondrial reads as they become apoptotic [188][189]. A high number of mitochondrial reads is therefore considered an indicator for poor quality cells [189]. For example, in the PHH diversity project, cells were removed if their fraction of mitochondrial reads was above 10 % of all reads. However, the isolation of intact single cells is not always feasible. Human samples are usually archived fixed in formalin and paraffin-embedded (FFPE) and sometimes preserved frozen in biobanks. Because working on frozen samples limits the isolation of intact single cells at high yield, studying the respective health or disease condition of these samples is challenging using standard scRNA-seq pipelines [139] [140][141]. The work of this thesis shows that the snRNA-seq2 method, tailored for frozen archived liver samples and developed by Dr. Ioannis Deligiannis in the Martinez-Jimenez lab, allows the successful characterization of single nuclei at the transcriptomic level. Moreover, it leads to the detection of around 3,600 genes per single nucleus (Figure 2.1), which is higher number than other s*RNA-seq methods achieve [16][37][19][28]. In line with previous reports, the gene expression in the nuclear extracts correlated at a Pearson correlation of 0.62 to whole cell extracts, deeming the method suitable to infer cellular function from the single nuclei [143]. Moreover, in line with SMART-seq2 yielding more detected genes and less noise for lowly expressed genes, expression of transcription factors could be measured and compared between conditions from snRNA-seq2 [56]. This therefore is a promising method to study archived samples and dissect the cellular heterogeneity and gene regulation in health and disease states from archived human samples [117].

Regardless of the chosen s*RNA-seq method, investigating the impact of biological conditions (e.g. ploidy status, lipid accumulation, or multiple drug administration) requires to separate biological variation from technical noise. For instance, figure 2.1 showed that nuclei prepared on the same 384-well plate can differ substantially in the amount of transcripts per nucleus. Therefore, measures to adjust for technical differences need to be considered. Additionally, as the amount of cDNA obtained from a single cell or nucleus is usually too low for direct sequencing, molecules are commonly amplified in a polymerase chain reaction (PCR). However, a technical obstacle during the reaction is the generation of PCR duplicates, i.e. several reads stemming from the same initial mRNA molecule, which can bias differential expression analysis [57][124]. Two commonly used approaches exist to handle PCR duplicates. Experimentally, a careful validation of the number of PCR cycles needed depending on the cell type and the amount of cells is performed to minimize the impact of PCR duplicates. Moreover, their occurrence can be removed through the addition of unique molecular identifiers (UMIs) [190]. Essentially, each initial mRNA molecule gets labeled by a UMI, so that in the end only unique molecules are counted for each gene. In the absence of UMIs, PCR duplicates can be removed computationally by removing reads that have the identical start and stop site as this indicates they are coming from the same molecule [191]. However, despite successful removal of PCR duplicates, bias during PCR amplification can still impose issues for downstream analysis. For instance, longer genes have a higher chance of amplification than shorter ones [76]. This is particularly important when analyzing full-length transcript data and needs to be taken into account to remove unwanted technical variation. Moreover, ERCC spike-ins have been shown to be amplified differently from endogenous transcripts [84]. Initially developed for bulk experiments, the idea of incorporating ERCCs is to add the same amount of synthetic molecules to each sample (or single cell) and thereby be able to dissect technical noise from biological variation [81]. In the snRNA-seq data set, the amplification bias between ERCCs and endogeneous transcripts could be confirmed for some of the plates (Figure 2.1). Therefore, the relationship between ERCC and endogenous transcripts was used for the development of a new, conservative normalization technique (Methods).

5.1.1 Normalization

In general, normalization aims to remove differences between samples or cells stemming from technical sources [73][72]. A common and well-established normalization technique is called library size normalization, where each cell's total read count is divided by the average library size [77][78]. However, this approach has been shown to be insufficient for removing all the technical variation in single cell experiments [70][124]. Lun *et al.* have suggested that this is potentially due to different cell types featuring different amounts of mRNA levels. Therefore they have developed an approach in which the difference in mRNA amount between cell types is considered, called *scran*, which outperforms simple library size normalization techniques [70][124]. In the comparison of 2n against 4n hepatocytes in this thesis, however, conservation of differences in mRNA levels was deemed inappropriate for assessing differences that are not associated to the levels of transcript abundance alone. In particular, *scran* led to the detection of over 1,250 up-regulated genes in 4n, which is a comparatively large number and indicates that these results were partially driven by the difference in transcript abundance between 2n and 4n hepatocytes. The adjusted normalization approach (ERN) presented in chapter 2 of this thesis instead led to the detection of 241 up-regulated genes in 4n (Figure 2.2). This suggests that normalizing data with different ploidy status benefits from using the ERCCs as a resource to minimize the impact of genome content on differential expression. Overall, the ERN normalization approach diminishes the effect of mRNA amount between 2n and 4n hepatocytes and still enables the successful dissection of the different liver cell types.

5.1.2 Cell type identification

While the results of the Ploidy&Zonation project indicate that *scran* showed a better performance at separating cell types from each other, the ERN approach presented in this thesis also allowed recovering the expected proportions of cell types in the liver (Figure 2.2). Regarding cell type identification, the largest discrepancy between the two tested normalization techniques was found in the assignment of lymphocytes and endothelial cells (Figure 2.2). One potential reason for this observation is that *scran* preserves differences in mRNA content between cell types, and therefore, leads to more power to dissect those cell types by clustering [124]. Independently of the normalization technique, the ability to successfully dissect immune cell types could be a general limitation of the experimental approach. Denisenko *et al.* have shown that sequencing RNA from single nuclei leads to poorer detection of immune cells in comparison to whole cell extracts [125]. Thus, low capture efficiency of the lymphocytes in this experiment could result in their less reliable identification regardless of the normalization approach. In general, the appropriate normalization technique depends on the respective biological question and the subsequent experimental design. In the Ploidy&Zonation project, the ERN approach was developed and tailored to fit the task of identifying biologically relevant changes between 2n and 4n hepatocytes. The ERN approach requires ERCCs to be present in a data set and relies on the assumption that cells or nuclei with a higher fraction of ERCC reads tend to have a lower fraction of endogenous reads, and vice versa. Figure 2.1 D shows that this is the case in the Ploidy&Zonation project. Hence, in a setting where the biological question is centered in finding differences that can be confounded by genomic content, a normalization approach making use of external spike-ins can be advisable. However, in the absence of ERCCs, such an approach cannot be applied. At the time of this thesis, *10X Genomics* does not recommend the use of ERCCs [121]. As discussed above, the droplet-based *10X Genomics* method offers the opportunity to achieve a high throughput of single cells, which allows to better identify cellular subtypes [56]. This method was used for the PHH diversity project as the project aimed to dissect cellular heterogeneity in a seemingly homogeneous population of cells, shifting the focus from the difference in ploidy status to instead identify variation among a large amount of hepatocytes. For normalization of this data set, *scran* was used because a recent benchmarking study showed that it outperformed other common normalization techniques at removing technical variation while retaining biological variation [124]. Four subgroups of PHHs were identified in the PHH diversity project that were not confounded by library size. Instead, these subgroups were characterized by metabolic specialization, indicating that indeed *scran* was able to remove the technical effect of library size and retain biological variation. As the subgroups were not driven by library size it can be hypothesized that hepatocytes with different ploidy status were distributed evenly among subgroups. However, further studies would be needed to confirm this.

5.1.3 Batch correction

Correct cell type identification and downstream analysis rely on the removal of unwanted variation. While normalization is designed to mainly remove technical biases within a given batch, its power is limited when fluctuation in data quality is high between batches [124][126]. In the snRNA-seq2 data set, each plate used to prepare single nuclei for sequencing represents a batch. Similar sequencing depths were obtained for all plates in this data set. Hence, *ComBat* was used as batch correction method since it has been shown to achieve good results for small to medium batch effects [126]. *ComBat* uses a linear model assigning similar values of mean and variance for all genes before performing batch adjustments on the count matrix [127]. As shown in figure 2.3, *ComBat* successfully removed unwanted inter-batch variation for the snRNA-seq2 data set. On the other hand, the batch integration was more challenging for the PHH diversity project. The presence of similar cellular structure between donors for the PHH diversity data was promising for the successful combination of the batches (Figure 2.4). Nevertheless, the difference in sequencing depth for the two batches required a more powerful batch correction approach. Luecken *et al.* have performed a thorough comparison of different batch correction techniques, investigating their level of data set integration and conservation of biological variance [126]. Both, *Harmony*, and *scGen* were placed among the top ranking methods for batch integration. While *scGen* received an overall higher ranking than *Harmony*, the latter was shown to perform best on pancreatic data. In this thesis, comparison of *scGen* and *Harmony* showed that similar results could be obtained with both methods (Figure 2.7). However, *scGen* needs a cell type annotation to use for data integration [129]. The power of detecting cellular substructure in a seemingly homogeneous population of cells increases with the number of cells available [48]. Thus, defining PHH subgroups prior to integration is less powerful than identifying them on the combined batches, arguing favorably for an unsupervised integration method that needs no prior knowledge. The goal of the PHH diversity project was to identify and characterize cellular heterogeneity within a single cell type, the PHHs. Therefore, since the true subgroups cannot be known beforehand, *scGen* was deemed the less appropriate tool. Generally, obtaining similar results with two different integration approaches validated the presence of similar subgroups in PHHs. In summary, *Harmony* led to successful integration of the two batches while conserving the biological variation present in all four donors (Figure 2.6). Overall, for both projects presented in this thesis, batch correction allowed for a more reliable characterization of cell types and cellular subtypes.

5.2 Transcriptional heterogeneity

After the characterization of cellular groups of interest, larger-scale differences between these groups are usually measured by calculating differential expression. Hereby, the mean expression levels of each gene are compared between conditions, usually through a *t-test* [77][78]. However, more subtle differences between conditions can be assessed by looking at the deviations of genes from their respective means in the conditions [192][134]. This is called transcriptional heterogeneity and can be measured by calculating the coefficient of variation [45][138]. It has been shown that lowly expressed genes generally harbour a higher coefficient of variation, biasing the correct assessment of highly variable genes [77][131]. In this thesis, the relationship between low mean expression and high coefficient of variation was confirmed in the snRNA-seq2 data (Figure 2.8). Therefore, calling genes highly variable based on high coefficient of variation was shown to be biased towards lowly expressed genes (Figure 2.8). Canchola *et al.* have furthermore reported spurious values for the coefficient of variation when applying the traditional formula on log-transformed data and suggested an adapted formula for this use-case [138]. As shown for the snRNA-seq2 data, applying this formula for the calculation resulted in an unbiased assessment of highly variable genes (Figure 2.8). Furthermore, the presence of ERCCs allowed to calculate a threshold for separating technical from true biological variation. Overall, a reliable calculation for transcriptional variability is crucial to identify fluctuations of gene expression in single cells. This can become especially relevant when investigating the cells' response to a stimulus as higher transcriptional variability has been shown to indicate a less coordinated response to the stimulus [133][132].

While studying the response of dendritic cells to lipopolysaccharide, Shalek *et al.* have reported the existence of small functional differences within this seemingly homogeneous cell population. These differences were indicated by changes in the gene expression distribution, i.e., immune genes were found to show bimodal expression patterns, partially linked to known maturity states of dendritic cells [192]. Similarly, the PHH diversity project in this thesis showed that, in response to lipid accumulation, hepatocyte subgroups can shift gene expression towards less coordinated responses and potentially contribute to disease progression. For instance, subgroups I and II of PHH *in vitro* showed increased variability in response to lipid accumulation whereas subgroup III showed a decrease. This indicates that different subgroups of PHH have different capacity to handle and clear lipid accumulation, which may play a role in diseases such as non-alcoholic liver disease (NAFLD). *In vivo*, NAFLD can develop from intracellular lipid accumulation and has been shown to change the cell type composition of the liver [26][7].

Intracellular lipid accumulation in hepatocytes has additionally been shown to occur during healthy aging [92]. Aging is associated with an increase in transcriptional variability in several tissues and cell types, including heart, lung, and immune cells [193][132][133]. The data analyzed in the PHH diversity project comes from human donors covering an age range of 18 to 57 years. Thus, human variability and age-related differences could influence transcriptional variability between and across subgroups. While the sample size was unfortunately not sufficient to study the impact of age on this data set, further studies on age-associated cellular heterogeneity will contribute to our understanding of transcriptional variability in the liver.

Moreover, both aging and NAFLD have been associated to an increase in polyploid hepatocytes [4][60]. However, before studying the impact of aging or a disease on a tissue, it is important to know the cellular heterogeneity under physiological conditions. Hence, the Ploidy&Zonation project presented in this thesis aimed to dissect heterogeneity between 2n and 4n hepatocytes in young healthy mice. As shown in chapter 3 of this thesis, shifts in the gene expression distribution were observed for several key hepatic marker genes. For instance, transcription factors *Pparα*, *Mlxipl*, *Rxra*, *Nr1i2*, and *Nr1i3* changed in distribution between 2n and 4n hepatocytes (Figure 3.8). This indicates that gene regulatory programs slightly differ between 2n and 4n hepatocytes in healthy young mouse liver. With old and diseased individuals harbouring higher proportions of polyploid hepatocytes, these differences in gene regulatory programs might change in disease conditions [60][4][61]. Generally, the Ploidy&Zonation project showed that 4n hepatocytes feature less noisy gene expression in comparison to 2n hepatocytes, which is in line with an earlier study by Bahar and Halpern *et al.* [14].

5.3 Polyploidy in the liver

Increased polyploidization has been observed in NAFLD and is also associated to poor prognosis in hepatocellular carcinoma (HCC) [4][5]. To shed light on the role of polyploidy in the young healthy adult mouse liver, the transcriptomic profile was compared between 2n and 4n hepatocytes. Most of the key transcription factors that were analyzed showed changes in their distribution as described above. However, no significant difference in their expression level were measured between 2n and 4n hepatocytes in young healthy mice. Furthermore, isoform expression for these transcription factors was observed to be similar between 2n and 4n hepatocytes 3.5. These results indicate that in young healthy mice, the gene regulatory programs are overall similar between 2n and 4n hepatocytes with the only differences being measurable in the gene expression distribution. However, differential expression analysis on the snRNA-seq2 data revealed that 2n hepatocytes are rather involved in the metabolism of proteins and carbohydrates whereas gene expression in 4n hepatocytes is enriched in pathways related to lipid, sterol, and xenobiotic metabolism. Additionally, genes were identified that showed no significant changes in mean expression between 2n and 4n hepatocytes but which significantly changed in their expression distribution. For instance, genes related to glucose and lipid metabolism, such as *Gk* and *Apob*, respectively, changed their distribution between 2n and 4n hepatocytes. This again indicates the presence of small functional differences in sub-populations of cells, that can become relevant in response to external or internal factors [192].

A known internal factor is the existence of higher ploidy levels in the liver. The analysis of nuclei from 8n, and 16n cells has shown the presence of genes increasing in expression level with ploidy as well as genes that decrease with ploidy. Among the genes that increased from 4n to 16n were the solute carriers *Slc9a9* and *Slc44a2* (Figure 3.9). Slc9a9 has been shown to maintain cellular homeostasis through proton exchange [194][195]. Likewise, Slc44a2 orchestrates the uptake of choline into the cell and choline deficiency has been associated to NAFLD in humans [196]. At the same time, larger cells have been shown to produce higher amounts of mRNA and proteins [67][66]. The results from this analysis therefore indicate how polyploid cells increase their capability for certain molecular uptake to sustain to larger energy demand while maintaining healthy levels of metabolites. On the other hand, *Slc19a1* and *Slc27a2* were found to decrease with ploidy status. Slc19a1 is responsible for the uptake of cyclic dinucleotides whereas Slc27a2 (also known as Fatp2) enables uptake of fatty acids [197][198]. Depletion of Slc27a2 has been shown to increase the expression of genes regulated by transcription factor Pparα, such as genes involved in fatty acid metabolism [198]. Future studies are needed to shed light on the relationship between changed gene regulatory networks in polyploid cells and the association to chronic liver diseases such as NAFLD and HCC. In summary, these results suggests that, in young healthy mice, gene expression in polyploid hepatocytes is tightly coordinated and adapted to sustain cellular homeostasis and normal metabolic functionality.

5.4 Stem cell properties in hepatocytes

Polyploid hepatocytes have been shown to proliferate at a slower rate than diploid hepatocytes [199]. Despite their slow proliferation, recent reports have suggested their involvement in tissue regeneration after liver injury [200][201]. This raises the question to which extent polyploid hepatocyte feature stem cell-like properties. In line with studies from Chen *et al.* and Lin *et al.*, the results from the Ploidy&Zonation project in this thesis indicate that both, 2n and 4n hepatocytes, share stem cell properties [153][201]. In particular, co-expression of nine selected stem cell marker genes was found in similar proportions of nuclei in both, 2n and 4n hepatocytes. These results were also validated by Dr. Kelvin Yin through immunofluorescence staining of Lgr5 mRNA in 2n and 4n hepatocytes. In a recent study, Matsumoto *et al.* have performed a multiple color reporter assay and concluded that, upon liver injury, polyploid hepatocytes undergo reduction of ploidy level and subsequent re-polyploidization [202]. Hence, the data analyzed in the present thesis are in line with recent literature and strongly suggest a stem cell potential of polyploid hepatocytes. Apart from the regenerative potential of polyploid hepatocytes, divergent evidence has been reported for the presence of a stem cell niche in the liver. For example, Kuwahara *et al.* have performed a labeling approach leading to the suggestion of stem cell "hot-spots" in association to bile ducts [203]. Similarly, Wang *et al.* have reported the presence of a pericentral stem cell niche fueling regeneration through expression of Axin2 [152]. However, Sun *et al.* have reported evenly distributed expression of Axin2 along the liver lobule [154]. Likewise, regenerative marker gene *Tert* has also been shown to be expressed evenly across the liver lobule [153], whereas an enrichment of Lgr5 has been observed at the pericentral area [154]. In line with these reports, the immunofluorescence RNA staining for Lgr5 showed a higher proportion of Lgr5+ cells at the central vein. In summary, it can be postulated that both, 2n and 4n hepatocytes, share regenerative potential whereas some stem cell markers are potentially enriched at the pericentral region.

5.5 Zonation

The bi-directional expression gradient between the pericentral and the periportal area in the liver lobule is referred to as zonation [11][157]. Resulting from the gradual supply of oxygen and nutrients from periportal to pericentral, zonation leads to a division of labor in hepatocytes [6]. While periportal hepatocytes involve in gluconeogenesis and protein metabolism, pericentral hepatocytes are rather performing glycolysis and metabolism of xenobiotics [89]. To obtain an in-depth characterization of gene expression in hepatocytes along the liver lobule, Halpern *et al.* combined single-molecule imaging with scRNA-seq, identifying marker genes for nine zones from pericentral to periportal [16]. For instance, *Glul*, *Cyp27a1*, and *Cyp2e1* represent pericentral marker genes whereas *Alb*, *Pck1*, and *Gls2* are associated to the periportal area. Putting the transcriptome of single nuclei into the context of zonation, these marker genes were used to generate a vector of diffusion pseudospace, ordering the nuclei from the most pericentral to the most periportal expression profile. This is based on recent developments on the analysis of scRNA-seq data, in which cells are ordered based on similarities between their gene expression profiles [90][204][91]. Indeed, this analysis enabled the identification of pericentral- and periportal-associated nuclei and the bi-directional expression gradient of respective marker genes. The presence of a bi-directional expression gradient as seen in the nuclei transcriptomic data was experimentally validated by Dr. Kelvin Yin through measuring mRNA fluorescence intensities of PV marker Cyp2f2 and CV marker Cyp2e1 along the liver lobule. This additional experiment confirmed that the results obtained through snRNA-seq2 were also measurable on the intact tissue. Thus, the results from the Ploidy&Zonation project indicate that the analysis of the whole transcriptome in single nuclei allows an in-depth characterization of the changes in gene expression along a spatial trajectory.

Especially in the last ten years, insights have been gained into how hepatocytes are distributed along this trajectory with regards to their ploidy status. For instance, Morales-Navarette *et al.* have found that hepatocyte volume increases from periportal to pericentral, indicating higher ploidy levels near the central vein [15]. Likewise, Tanami *et al.* have used single-molecule imaging which led to the conclusion that 2n hepatocytes reside closer to the periportal vein while 4n hepatocytes show higher abundances at the mid-lobular and central vein areas [85]. In line with these studies, nuclei from the Ploidy&Zonation project assigned to PV were enriched in 2n nuclei whereas the proportion of 4n nuclei was higher in the nuclei assigned to CV. A recent study by Katsuda *et al.*, however, performed single cell quantitative PCR on FACS-sorted 2n, 4n, and 8n cells and reported a pericentral gene expression profile in 2n cells [156]. This could potentially be due to hepatobiliary cells infiltrating the 2n hepatocyte population as their gene expression profile is similar to hepatocytes [21][22]. Hence, these populations could be difficult to disentangle without having the whole transcriptome or spatial information available. As both, pericentral hepatocytes and hepatobiliary cells, are involved in bile synthesis, having hepatobiliary cells present in the population could therefore explain a pericentral gene expression profile within 2n cells [88]. In the snRNA-seq2 data set here, 2n hepatocytes featured a periportal gene expression profile in comparison to 4n hepatocytes. For example, performing gene ontology analysis revealed that the top 100 up-regulated genes in 2n hepatocytes were enriched in gluconeogenesis, which has been associated to the periportal region [157][6]. Furthermore, periportal marker genes *Pck1*, *Alb*, and *Cyp2f2* were significantly up-regulated in 2n hepatocytes. Additionally, marker genes for the periportal-associated protein metabolism, such as *Gls2* and *Hal*, were up-regulated in 2n hepatocytes. In contrast to that, 4n hepatocytes were enriched in genes related to xenobiotic metabolism, associated to the pericentral area [88][16][89]. For instance, cytochromes from the P450 family of enzymes, such as *Cyp2e1*, *Cyp1a2*, and *Cyp7a1* were significantly up-regulated in 4n hepatocytes. Together these results indicate that the differences between 2n and 4n hepatocytes in young healthy mouse livers are partially driven by their respective location along the liver lobule. In summary, 2n hepatocytes exhibit a periportal, whereas 4n hepatocytes exhibit a pericentral gene expression profile

5.6 Zonation-independent hepatocyte heterogeneity

It has been demonstrated that liver zonation influences gene expression patterns *in vivo*. To closely study this effect *in vitro*, several approaches have been developed to model zonation, mainly through generating an oxygen-gradient similar to the physiological condition *in vivo* [97]. For example, Scheidecker *et al.* have shown that modeling zonation in primary rat hepatocytes through biomimetic supply of oxygen induces drug metabolism in concordance to what is seen *in vivo* [116]. However, an inappropriate oxygen concentration can lead to undesirable gene expression profiles, associated to e.g. hepatocellular carcinoma and increased drug-induced injury [116]. The 2D culture of primary human hepatocytes (PHHs) represents the gold standard model for testing drug metabolism *in vitro* [97]. Without biomimetic supply of oxygen, no zonation patterns are present in PHHs in culture [97]. This was also shown in the transcriptomics data set of the PHH diversity project in this thesis. Calculating scores for the expression of CV, and PV markers, respectively, showed no enrichment of a CV or PV profile in any of the cells. However, as described in chapter 4 of this thesis, analysis of the individual donors revealed an underlying cellular heterogeneity that was consistently present across all four donors, independently of treatment condition and zonation. Thus, four subgroups sharing similar gene expression profiles across donors were defined. In line with studies from Hu *et al.* and Heslop *et al.*, one of the subgroups was shown to have de-differentiated along culture time resulting in a loss of the hepatocyte-like transcriptomic phenotype [114][115]. The other three identified subgroups showed distinct gene expression profiles suggesting their respective functional specialization. While subgroup I expressed genes related to bile acid and sterol metabolism, subgroup II expressed genes involved in carbohydrate and phase II metabolism. Finally, subgroup III expressed markers related to lipid and phase III metabolism. The analysis of two publicly available data sets from a total of 14 human livers confirmed the presence of these three metabolically active hepatocyte subgroups *in vivo* [19][18]. In one of the two publicly available data sets, MacParland *et al.* have reported the presence of six clusters of hepatocytes *in vivo* that were found to be partially influenced by zonation [18]. The comparison of this data set to the PHH data revealed that three of their clusters showed high transcriptomic similarity to *in vitro* subgroup II hepatocytes, whereas the other two subgroups from the PHH data were more difficult to annotate on this *in vivo* data set. This could potentially be due to inter-individual differences among the donors or due to the PHHs being cultured for 72 hours. Nevertheless, investigating the gene expression patterns of the PHH subgroups in the clusters of the publicly available *in vivo* data sets led to the successful identification of the three metabolically active PHH subgroups *in vivo*.

In the publicly available *in vivo* data sets, gene expression is affected by zonation [19][18]. This allowed to study the relationship between zonation and the subgroups' gene expression profiles *in vivo*. While CV, midzonal, and PV expression patterns were found in all three subgroups, subgroup III had the smallest proportion of cells exhibiting a periportal gene expression profile. This could indicate a smaller proportion of subgroup III hepatocytes near the portal vein *in vivo*. In general, xenobiotic metabolism predominantly takes place at the pericentral region in the liver lobule [6]. Hence, their specialization into phase III metabolism could suggest a rather midzonal to pericentral role of subgroup III hepatocytes. Overall, no subgroup was found to correspond to a specific zone indicating that also *in vivo* the subgroups are present independent of zonation. However, marker gene expression within the identified subgroups showed zonal patterning *in vivo*. For instance, cells within subgroup I in the Aizarani *et al.* data set showed distinct patterns for CV and PV marker genes and a split of the population into cells exhibiting a PV, and CV expression profile, respectively (Figure 4.4).

Data integration of the two publicly available *in vivo* data sets from 14 human donors increased the power of analysing the impact of zonation on gene expression within the metabolically active subgroups [19][18]. Moreover, trajectory inference from scRNA-seq data allowed to order cells by the similarities of their gene expression profiles [90][204]. Ordering the cells in each subgroup *in vivo* from the most PV to the most CV expression profile enabled to investigate the impact of zonation on each subgroup separately. In subgroup II, marker gene expression did not follow the expected zonation trajectory from PV to CV (Figure 3.11). This indicated that zonation was not a main driver of variation for subgroup II cells, and thus, that their gene expression profile was comparatively independent of zonation. However, the trajectory analysis revealed zonation-related labor-separation within subgroup I and III. For instance, in line with the known spatial separation of the pathway synthesizing bile acids from cholesterol, *CYP8B1* showed pericentral enrichment in subgroup I, whereas *HMGCS2* was periportally enriched [157][6]. Moreover, expression in subgroup III reflected the reports from Halpern *et al.* and Aizarani *et al.*, in which expression levels of *Apob* were highest in the midzonal regions while *Mlxipl* was pericentrally enriched [16]. In summary, the metabolic specialization into subgroups that was identified in PHHs *in vitro* was also present *in vivo* and independent of zonation. Furthermore, the in-depth study of hepatocytes from 14 human donors showed that specific marker genes within the subgroups showed spatial patterning *in vivo*.

5.7 Intracellular lipid accumulation

Recently, zonation has been shown to impact the cellular response to intracellular lipid accumulation [7]. In Park *et al.*, mice fed a high-fat diet have shown an increased lipid droplet accumulation in CV hepatocytes, whereas PV hepatocytes down-regulated their characteristic gene expression profiles upon lipid accumulation [7]. Apart from the zonation-related heterogeneity found among hepatocytes, this thesis has demonstrated the existence of hepatocyte subgroups characterized by zonation-independent metabolic heterogeneity, both in *in vivo* and *in vitro*. Therefore, studying the effect of intracellular lipid accumulation on these subgroups provides another layer of how steatosis affects hepatocytes. For instance, as described above, lipid accumulation differently impacts transcriptional variability in the PHH subgroups. Increased transcriptional variability has been linked to destabilization of biological programs during ageing and a less coordinated response of cells to a stimulus [133][130][132]. The PHH diversity project has shown that transcriptional variability is increased in subgroup I and II in response to intracellular lipid accumulation, indicating that these cells showed a less coordinated transcriptomic response to hepatic steatosis.

Moreover, studies by Anstee *et al.* and Pan *et al.* have shown that hepatic steatosis leads to increased expression of inflammatory pathways [205][167]. In particular, increased expression levels of chemokines, driving inflammatory responses, have been observed in NAFLD [167]. In the PHH diversity project, subgroup I showed increased expression of chemokines, suggesting this subgroup is more susceptible to inflammation and has greater potential to drive disease progression *in vivo*. Additionally, in subgroup II, intracellular lipid accumulation triggered the up-regulation of pathways involved in the clearance of neutral lipids and triglycerides. Triglycerides have been shown to accumulate within hepatocytes in NAFLD [206]. Adaptations in gene expression to clear the triglycerides from the cells through β-oxidation can result in increased levels of reactive oxygen species (ROS) [207]. While ROS levels were not directly measured from the PHHs, the gene expression profile of subgroup II cells indeed suggests oxidative stress. For example, the *LGALS* family of genes has been associated to oxidative stress in cancer and *LGALS1* was among the top five up-regulated genes in response to lipid accumulation in subgroup II [208]. Taken together, these results indicate that triglyceride accumulation associated to steatosis triggers their increased clearance and subsequent oxidative stress in subgroup II hepatocytes.

Subgroup III hepatocytes also up-regulated pathways related to triglyceride clearance in response to intracellular lipid accumulation. However, to a higher extent, they up-regulated genes enriched in pathways for lipid metabolism and the regulation of lipid metabolism. Across all treatment conditions, subgroup III cells were characterized by phase III and lipid metabolism, but the expression of genes related to lipid metabolism was further increased in the model of hepatic steatosis. Moreover, transcriptional variability was significantly decreased in this subgroup. These findings indicate that hepatocytes showing a transcriptomic specialization in lipid metabolism react in a more targeted fashion to hepatic steatosis. In summary, intracellular lipid accumulation triggers cellular responses specifically for each of the specialized hepatocyte subgroups. With the characterization of metabolically distinct hepatocytes that exist independently of zonation, this *in vitro* study offers a new aspect for future *in vivo* studies investigating the impact of steatosis at single cell resolution.

5.8 Drug metabolism in healthy and steatotic PHHs

NAFLD and its associated changes in lipid metabolism has been shown to directly impact the expression of genes in the cytochrome P450 superfamily [174][9]. Members of this superfamily orchestrate phase I metabolism of xenobiotic substances, in which oxidation, reduction, hydrolysis, and cyclization reactions occur [102]. Subsequent phase II drug metabolism is carried out by GST, SULT, UGT, and NAT families of enzymes [106]. Finally, the catabolites of drug metabolism are excreted from the cell by transmembrane transporter proteins [107]. Despite their tendency to undergo de-differentiation after 72h in culture, PHHs have been shown to express the majority of these drug-metabolising enzymes, respond to enzyme inducers, and produce a metabolic profile similar to *in vivo* [95][115]. Furthermore, among others, Bosilikovska *et al.* have shown their usability to study the activity of cytochromes upon stimulus [113][112][110]. Of particular interest here are the enzymes CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4 as they are involved in the metabolism of 70-80 % of all commonly used drugs [12][105].

In comparison to studying the effect of individual drugs, a cocktail approach allows to investigate the activity of several cytochromes simultaneously [110][112][113]. The Sanofi-Aventis cocktail used here consists of these selective substrates for each of the aforementioned five cytochromes: caffeine (CYP1A2), S-warfarin (CYP2C9), omeprazole (CYP2C19), metoprolol (CYP2D6), and midazolam (CYP3A4) [110]. While the original study reported no interactions in the metabolism of those compounds, a more recent study by Berger *et al.* has reported that CYP2B6, CYP2C9 and CYP3A4 contribute to the metabolism of metoprolol *in vitro* [110][175]. However, the authors conclude that the contribution is minor and does not compromise the appropriateness of metoprolol as phenotyping probe for CYP2D6 [175]. Additionally, Borkar *et al.* have shown that metoprolol can inhibit the expression of CYP2D6 and CYP3A4, but that this does not affect the metabolism of midazolam by CYP3A4 [176]. Furthermore, a time-dependent inhibitory effect of CYP2C19 expression has been reported for treatment with omeprazole [177]. Overall, the Sanofi-Aventis cocktail is considered suitable to measure liver metabolic capacity *in vivo*, study drug-drug interactions of new drugs in development, and phenotype cytochrome activity both *in vivo* and *in vitro* [110][209]. Exploring the gene expression profile of single hepatocytes in response to this cocktail enabled the detection of subtle effects that are potentially concealed in bulk studies. Pseudo-bulk analysis of the single cell data showed induced expression levels of CYP2C9, CYP3A4 and CYP1A2 upon cocktail treatment, although *CYP2C9* had a low $\log_2$-fold change towards DMSO. Expression levels of CYP2D6 and CYP2C19 were comparable to baseline DMSO-levels, which could be linked to the previously reported pharmacokinetic effects of metoprolol and omeprazole [176][177]. However, no down-regulation of any of the five targeted cytochromes was observable, suggesting no inhibitory effects of the compounds on the targeted cytochromes in the studied PHHs at the transcriptome level. Across all subgroups, treatment with the Sanofi-Aventis cocktail led to down-regulation of the alcohol-dehydrogenases ADH1B and ADH1C (Figure 4.9 C). These genes are main contributors to successful alcohol clearance in the liver [210]. Hence, their consistent down-regulation could lead to potential accumulation of alcohol, which represents an interaction between alcohol and drug metabolism that has previously been reported to lead to adverse drug reactions [211][212]. Moreover, at single cell resolution, differences in the expression level of the five targeted cytochromes were observable between the subgroups. Along with the loss of their characteristic hepatocyte-like expression, cells in subgroup IV showed the least expression levels of the five targeted cytochromes. This could become relevant when testing the efficacy and safety of new drugs in bulk studies. Despite PHHs overall retaining their drug-metabolic capacities *in vitro*, the proportion of cells losing their expression could partially conceal the true levels of cytochrome induction or inhibition across all cells in response to a drug when bulk studies are performed [95]. This is of concern as correct measures of cytochrome inhibition and induction in response to drugs are considered crucial for the assessment of potential drug-drug interactions and toxicity [111]. Additionally, different levels of cytochrome expression were also found between the three metabolically active subgroups. For instance, *CYP3A4* was 4-fold induced over DMSO-levels in subgroup I while no significant up-regulation was observed in subgroup III hepatocytes. Overall, these results highlight the impact of hepatocyte specialization for the accurate assessment of drug efficacy, toxicity, and potentially, drug-drug interactions.

Studies have reported that the expression of cytochrome P450 enzymes is furthermore impacted by aging and human ethnicity [213][214][215]. The donors from which the PHHs were acquired for the PHH diversity project were from different ethnic groups (three Caucasian and one Hispanic) and covered an age-range of 18 to 57 years. Thus, this human variability could also contribute to heterogeneity in the expression of cytochromes. While the sample size did not allow to investigate the effect of age or ethnic group in this data set, the overall expression levels of CYP2C9 were found to be overall higher than the expression levels of the other four cytochromes (Figure 4.9). This is in line with a study by Liu *et al.*, who reported higher basal expression levels of CYP2C9 in comparison to CYP1A2 and CYP2D6 for African Americans, Caucasians, and Hispanics [215].

Investigating the effect of the Sanofi-Aventis cocktail on the whole transcriptome of PHHs at single cell resolution further revealed how metabolic specialization impacts the drug-metabolism related response to a specific drug challenge. For instance, *PLIN2* and *OSGIN1* are involved in the coating and metabolism of lipids and were significantly up-regulated in response to the cocktail in subgroup III hepatocytes specialized in lipid and phase III metabolism [206]. This suggests that the identified PHH subgroups maintain their metabolic specialization in response to a stimulus. Performing gene ontology analysis on pathways related to the catabolism of xenobiotic compounds further increased the evidence for subgroup-specific responses to a drug challenge [178]. While some pathways were shared between subgroups, every subgroup up-regulated specific pathways associated to the metabolism of different xenobiotic compounds in response to the cocktail. This suggests that metabolic specialization of hepatocytes could eventually impact what xenobiotic compounds the cells are able to metabolize. Especially *in vivo* where gene expression within the subgroups is additionally influenced by zonation, individual hepatocytes might be more efficient at clearance of certain compounds whereas others might be more susceptible to contribute to drug-induced liver injury.

Furthermore, drug-induced liver injury and adverse drug reactions have been reported to occur more frequently in patients suffering from NAFLD [180][207]. Changes in the drug-related metabolism under steatosis include changes in the gene expression of cytochromes in phase I [174][9][180]. Comparing the expression of the five targeted cytochromes in response to the Sanofi-Aventis cocktail revealed overall lower expression levels in the steatosis model than under physiological condition across PHH subgroups. In particular, subgroup III hepatocytes significantly down-regulated four of the five targeted cytochromes towards their baseline DMSO levels when the cocktail was administered to cells experiencing intracellular lipid accumulation. In otherwise untreated cells, administration of the Sanofi-Aventis moderately induced expression of genes in the cytochrome P450 pathway, confirming the retained drug-metabolic capacity of PHHs *in vitro* [95]. However, while the cells still up-regulated genes related to stimulus responses, the presence of intracellular lipid accumulation led to a gene expression profile that was less specific to drug metabolism. Moreover, PHHs treated with both, FFA and the cocktail, specifically up-regulated genes enriched in pathways related to general cellular processes and response to lipid. These results suggest diminished capacity for drug clearance in PHHs under hepatic steatosis. Eventually, this could contribute to drug toxicity as reported in patients suffering from chronic liver diseases [216][217]. Furthermore, the changes in amount of differential expression between cocktail treatment alone and FFA+cocktail treatment was similar between subgroups. While this does not indicate which specific pathways are affected per subgroup, the low amount of drug-specific genes upon intracellular lipid accumulation indicate diminished drug-related metabolic efficiency across all subgroups.

The administration of certain drugs has been associated to trigger insulin resistance in some patients *in vivo* [218][219]. To study this, the effect of the cocktail-treatment was tested on steatotic and otherwise untreated PHHs *in vitro*. It was revealed that the combination of steatosis and drug administration triggered up-regulation of genes in the insulin resistance pathway. Neither the steatosis model, nor the admission of the Sanofi-Aventis cocktail alone led to an up-regulation of insulin resistance related genes in any of the subgroups. Furthermore, in comparison to cocktail-treatment alone, cocktail-treatment under modelled steatosis led to the down-regulation of cytochrome P450 superfamily genes. These results provide further evidence for less efficient drug clearance and the potential development of adverse drug-reactions in hepatocytes under steatosis [207][219]. Under DMSO, PHHs up-regulated cytochrome P450 genes to initiate drug clearance, whereas steatosis diminished the level of cytochrome expression in all subgroups. In particular, subgroup III hepatocytes down-regulated the expression of several cytochromes when the cocktail was administered in the presence of intracellular lipid accumulation. This suggests that lipid accumulation has an inhibitory effect on the cytochromes in subgroup III hepatocytes and thus, drug metabolism is particularly impaired in these cells under hepatic steatosis. Previously, variability in the incidence and progression of NAFLD has been observed between patients associated to genetic and environmental factors [220]. However, the results of this thesis suggest that the proportions of different hepatocyte subgroups in a patient could also contribute to this variability. Especially, different proportions of subgroups could influence which xenobiotic compounds the patient can efficiently clear, and the risk of developing adverse drug reactions under steatosis.

6 Conclusions and Outlook

Single cell RNA-sequencing has made it possible to characterize and study functional differences in a seemingly homogeneous population of cells [33][54][187]. However, the experimental design and data analysis influence the interpretation of the results [124]. For instance, conserving differences in the mRNA content between cell types has been confirmed to be advantageous for correct cell type identification [70][124]. On the other hand, when investigating the differences between cells featuring whole genome duplication, the results shown here indicate the benefit of adapting library size normalization to regard differences in genome content. Thus, not only the experimental design, but also the steps along the computational analysis need to be considered individually for each biological question, depending on the specific context.

Overall, the analysis of single cell transcriptomic data presented in this thesis has led to the identification and characterization of hepatic cellular heterogeneity *in vivo*, and *in vitro*. With the recent advances in the field of single cell genomics, it is now possible to study gene regulation by assessing chromatin openness [221][222] and DNA methylation [223]. Both -omics can also be measured together with gene expression within the same single cell [224][225]. In the context of polyploidyzation in the liver, this paves the way to study gene dosage compensation, i.e. mechanisms that allow cells to establish functional mRNA levels despite having more than two copies of a given gene. Exploring the regulatory landscape behind genes increasing or decreasing in their expression level with higher ploidy shown in this thesis could give an entry point for understanding gene dosage compensation and lead to a better understanding of ploidy contributing to tissue function. In line with previous studies, the snRNA-seq2 data here revealed that, in the young healthy adult mouse liver, 2n hepatocytes feature a rather periportal gene expression profile whereas 4n hepatocytes express pericentral pathways [15][85]. This suggests that polyploidyzation and zonation are connected. Here, gene expression in the context of zonation was studied using well-established zonation marker genes and diffusion maps, but recent technological advances now enable to explore transcriptomics directly in the spatial context of a tissue [226][227]. This technique is called spatial transcriptomics and offers a promising approach to further study zonation, especially in the context of chronic liver diseases [228]. For instance, in the present thesis, fibrosis in a mouse model has been shown to disrupt zonation patterns. While sample sizes were not sufficient to perform statistical testing, the fibrosis model showed a decrease of 2n hepatocytes in comparison to wild type animals. These findings indicate a complex interplay between ploidy level and zonation that is affected by chronic liver disease. Applying spatial transcriptomic techniques to measure the impact of chronic liver diseases on zonal expression patterns will hence improve the understanding of these diseases and yield potential treatment targets in the future.

Accumulation of lipids, referred to as hepatic steatosis, is associated to the development of chronic liver diseases, such as NAFLD. Research has shown that *in vivo*, hepatic steatosis alters the transcriptional landscape of hepatocytes in a zonation-dependent manner [27][7]. While the context of zonation is lost in the standard 2D culture of primary human hepatocytes *in vitro*, the results in this thesis indicate that zonation-independent metabolic specialization exists *in vitro* and *in vivo*. Therefore, susceptibility to the development of NAFLD and other chronic liver conditions may be impacted by the proportions of different subgroups of hepatocytes in humans. For instance, based on their metabolic profile, some hepatocytes may be more susceptible to contribute to inflammation in hepatic steatosis whereas others can help with lipid clearance but are at higher risk for developing adverse drug reactions and contribute to drug-induced liver injury. The impact of hepatocyte subgroups has therefore to be further studied carefully *in vivo*, especially in the context of drug toxicity and hepatic steatosis. Additionally, increased levels of bi-nucleated tetraploid hepatocytes have been shown to arise in NAFLD *in vivo* [4]. However, how ploidy status impacts hepatocyte function under hepatic steatosis also remains to be studied in the future.

In this thesis, cellular heterogeneity of hepatocytes is reflected in polyploidy, zonation, and metabolic specialization. Under physiological conditions, these factors are intertwined to sustain healthy hepatic functionality. Intrinsic and external perturbations such as lipid accumulation, senescence or drug administration can, however, disrupt metabolism by causing imbalance between these different layers of heterogeneity. Future studies applying new technologies therefore need to take into account how the polyploidy, zonation, and metabolic specialization shape the response to different stimuli, such as disease, polypharmacy, aging, and a combination thereof.

7 Methods

7.1 Background

In eukaryotic cells, DNA is stored in the nucleus, where genes get transcribed by RNA-polymerase II. The resulting messenger RNA (mRNA) molecules get subsequently capped at their 5’ end and poly-adenylated at their 3’ end before they get exported from the nucleus, undergo splicing, and are translated into proteins. Thus, measuring gene expression - and thereby inferring information on what proteins eventually end up defining the cells functionality - can be performed by extracting the mRNA from cells or nuclei. While nuclei do not contained mature, spliced mRNA, the nuclear transcriptome is still representative of the whole cell [143]. Usually, the mRNA molecules within a cell/nucleus (or within a bulk of cells) are captured by their poly-adenylated end [229][230]. Other methods, including capture by cap-sequence or rRNA-depletion without pre-selection, have been described but are considered less efficient and are therefore less commonly used [231] [230]. The captured mRNA molecules are then reverse transcribed to generate cDNA that is amplified and subsequently sequenced. To understand which genes are expressed to what extent in a single cell/nucleus at the time of collection, those reads have to be assigned to a single cell/nucleus and aligned to a reference genome and counted in a meaningful context. Assigning reads to individual cells/nuclei generally works by using a barcode to label each individual cell/nucleus. Through the tremendous work of the last decades, regions in the human as well as the mouse genome have been and are being annotated, leading to a publicly available resource making it possible to quantify reads in a meaningful context [232][233]. Moreover, algorithms have been developed to align reads to a reference genome taking into account splice junctions to better handle reads stemming from RNA-sequencing [118].

There are two major ways to label single cells or nuclei with barcodes. The first one relies on putting the cells/nuclei into the wells of plates containing the barcodes [234], while the second one makes use of liquid droplets to encapsulate the cells or nuclei [55].

7.2 Read alignment and building count matrices

For the single nucleus RNA-sequencing data set coming from adult mouse liver cells, a plate-based sequencing approach was used, meaning that single nuclei were sorted into 384-well plates through fluorescence activated cell sorting (FACS)[31]. Raw sequencing reads were aligned against a customized genome containing the mm10 (GRCm38, assembly version 93) genome, as well as the ERCC92 sequences [80]. This was performed through the use of STAR-2.7.1a with the parameters *–outFilterMultimapNmax 1 –outSAMtype BAM SortedByCoordinate*. During the amplification step in library preparation, molecules will be duplicated, leading to potential biases when analysing the number of counts per gene. Thus, PCR-duplicates need to be accounted for. Computationally, they can be removed by using the *picard* tool *MarkDuplicates*, which was performed here using version 2.20.2 with the parameter *REMOVE_DUPLICATES* set to *true*. In standard single cell RNA-sequencing analysis, reads would then be counted in the context of exons, i.e. generating a count matrix containing the information which exons are present to which extent in the single cells. This data set, however, has been produced using a method developed by Picelli *et al.*, leading to the generation of full-length transcripts [53]. Moreover, nuclei contain a larger proportion of unspliced transcripts than whole cells. Therefore, a more versatile approach for counting reads has been chosen to retain the information. In every nucleus, reads were counted per transcript and then aggregated per genes by using *htseq-count*, version 0.11.3 with *-m intersection-nonempty -f bam -r pos -s no –nonunique all -t transcript -i gene_id –additional-attr=gene_name*. Thus, this ensured that reads mapping to more than one gene were counted for the gene they mapped to by larger proportion, and that reads mapping in equal proportions to two genes were counted for both genes. This approach resulted in a raw count matrix of 2,496 single nuclei times 54,329 genes.

The data were generated by Dr. Ioannis Deligiannis using SMARTer chemistry, which yields full-length transcripts. Thus, count matrices for exons, introns and individual transcripts were built in addition to the count matrix based on genes per nucleus. To generate the intronic and exonic count matrix the R-based tool *featureCounts* was used [235]. This was mainly done to specifically count what number of overall reads were falling into introns and exons, respectively. Moreover, by counting the reads associated to individual transcripts per nucleus, the presence of alternative transcripts and their respective share per single nucleus can be identified. This was performed using the pseudo-alignment tool *kallisto* [236]. In the context of whole genes, this can be used to identify which transcripts are preferentially transcribed for a given gene. Disentangling this information in individual cell types and cell states can eventually lead to the detection of differences in which transcripts are preferentially transcribed in a given cell type or state. For instance, in the context of hepatocyte ploidy, this analysis can be used to identify differences in alternative transcript preferences between diploid (2n) and tetraploid (4n) hepatocytes.

The PHH diversity project aimed to characterize cellular heterogeneity of primary human hepatocytes (PHHs) in response to a five-drug cocktail. For this project, experiments for data generation were performed by Eva Sanchez-Quant and Dr. Ioannis Deligiannis using the 10X Genomics Single Cell 3' Reagent Kit for library preparation and the 10X Genomics Chromium platform for sequencing. In this case, whole cells were encapsulated into droplets and both, nuclear and cytoplasmic mRNA was captured. For each of the two batches, the raw reads were subsequently aligned to GRCh38 and counted in the context of exons using the software *cellranger*, version 4.0.0 by 10X Genomics with standard parameters. Different methods exist to account for PCR duplicates stemming from the cDNA amplification step in library preparation. Other than removing them computationally after alignment as described above, they can experimentally be controlled for through the use of unique molecular identifiers (UMIs)[190], a method where every molecule gets labelled by a barcode and only unique molecules get counted. This approach was used here, and thus, no computational identification method was needed. The two count matrices generated through *cellranger* were finally concatenated using the AnnData function *AnnData.concatenate()*, resulting in a combined count matrix of 63,527 cells times 19,971 genes.

In both main projects reported in this thesis, the raw count matrix was loaded into python and stored as an AnnData object, anndata version 0.7.1. The respective downstream analyses, described in detail in the following sections, were in parts inspired and adapted from Luecken *et al.* [44].

7.3 Quality control and filtering

In the Ploidy&Zonation project, reads were not only mapped against the mouse mm10 genome but also against the sequences of 92 spike-ins (ERCC92). To calculate the percentage of ERCC reads per single nucleus, the reported number of uniquely mapped reads per nucleus was taken from the *STAR* log file and ERCC reads were counted from the according *.bam* file. From that, the proportion of ERCC reads per uniquely mapped reads per nucleus were calculated and added to the nuclei metadata in R version 4.0.3 [237]. The complete table containing nuclei metadata was added to the count matrix in python version 3.7.6. In order to remove low quality nuclei, in the initial count matrix of 2,496 single nuclei times 54,329 genes, nuclei were kept if the percentage of ERCC reads was between than 5% and 90%, and if nuclei had more than 1,000 genes detected. The R tool *scPower* was used to perform a power analysis reporting the power to detect rare cell types in the population [48]. With the data coming from four biological replicates with an average of 624 single nuclei per replicate, the power to detect at least five cells from a cell type making up 2% of the population was 0.98. The detection of rare cell types was not an aim of this project and the power to detect at least five cells from a cell type making up only 1% of the population instead was only 0.31. Therefore, this analysis was used as a basis to remove genes that were present in less than 1% of the population (fewer than 25 nuclei) and had low coverage (fewer than 250 reads) to reduce noise stemming from lowly expressed genes. During the FACS sorting of nuclei into the wells of 384-well plates, some events might have occurred where either no nucleus had been sorted into a well, or two single nuclei ended up in one well. In these cases, the events had been noted down and data from the respective wells were removed from the analysis. Eventually, nuclei were kept if they had less than 7,000 genes detected, and a library size between 10,000 and 300,000 reads per nucleus. This filtering process resulted in a count matrix of 2,016 single nuclei times 19,340 genes. Additionally to the 2n and 4n nuclei, hepatocyte nuclei with eight, and sixteen genome copies were isolated and sequenced by Dr. Kelvin Yin. Since they were processed under another chemistry, yielding a higher number of genes per single nucleus, these nuclei with higher ploidies were kept if they had at least 1% ERCC reads, between 1,000 and 12,000 genes, and 5,000 to 700,000 reads covered per nucleus. Furthermore, 2n and 4n nuclei from another healthy wild type control mouse and a mouse treated with carbon tetrachloride (CCl_4)

to model liver fibrosis were isolated and sequenced. This fibrotic mouse model was established by Prof. Dr. Neil Henderson, and the respective samples were processed and sequenced by Dr. Ioannis Deligiannis. These samples were again processed with the chemistry providing a higher number of reads and genes per nucleus. Hence, the filtering was adapted to keep nuclei if they had between 30,000 and 1,000,000 reads. After normalisation, nuclei with more than 300,000 gene-length normalized counts were further removed. For this fibrosis data set, the hepatocytes were computationally separated from the non-parenchymal nuclei by an initial *Louvain* clustering at a resolution of 0.2. Finally, differential expression analysis and the pseudospatial ordering based on zonation markers was performed individually for the wild type nuclei and the nuclei from the CCl_4-treated animal.

For the human data set, cells were kept if they had a minimum of 1,000 reads and 500 genes. Genes were removed if they were present in less than 5 cells and had more than 5 million reads. In liquid droplet based methods, the capture of more than one single cell within on droplet can occur, which is referred to as doublets []. Therefore, computational methods have been developed to detect these doublets and subsequently remove them from downstream analyses. This is especially relevant for data sets composed of several cell types where a doublet can consist of a mixture of two cell types. However, as described above, hepatocytes are subject to polyploidization. Thus, it was expected that not only diploid, but also tetraploid cells were present in the study, with tetraploid cells naturally being of larger size than diploid cells [66]. Moreover, intracellular lipid accumulation results in the enlargement of cells rendering it unlikely that more than one single cell is captured in a liquid droplet during library preparation. Due to these properties, a lenient cutoff of 0.15 was selected for eliminating potential doublets, leading to the removal of 1.7% of cells using *scrublet* [238]. Finally, the proportion of mitochondrial reads per cell is an indicator for oxidative stress [188][189]. The majority of cells (around 96%) in this data set had a proportion of mitochondrial below 1%, indicating that only few cells were under oxidative stress. This is most likely due to the cells having been filtered for viability using magnetic beads. Due to the overall low number of mitochondrial reads, cells with more than 1% mitochondrial reads were removed, resulting in a filtered matrix of 49,378 cells times 16,256 genes.

7.4 Normalization

In the snRNA-seq2 data set, ERCC spike-ins were used as a tool to control for technical noise. When libraries containing both ERCCs and endogenous transcripts get sequenced to saturation, the proportional amounts of recovered synthetic spike-ins and endogenous transcripts depend on the number of endogenous starting material. During library preparation of this data set, two different dilutions were used when adding the spike-ins, 1:100,000, and 1:300,000, respectively. Thus, the ERCC size factor per nucleus was calculated separately for the two dilutions by taking the sum of ERCC reads per nucleus and dividing it by the mean ERCC reads across all nuclei in one dilution. Methods generating full-length transcripts can experience bias in which genes are sequenced more deeply due to their corresponding gene length. Hence, it is good practise to normalize transcript counts by gene length to adjust for this. Per gene, read counts were therefore first divided by the respective gene length in kilobases (kb). Since the proportional number of endogenous transcripts against ERCC reads depends on the input amount of endogenous transcripts, it was anticipated that tetraploid hepatocytes would on average have higher proportion of endogenous transcripts than diploid hepatocytes. To avoid differential expression between diploid and tetraploid hepatocytes to be purely driven by the higher number of counts in tetraploid nuclei, an adjusted normalization approach was developed, in which the ratios of endogenous to ERCC reads were considered per nucleus. In that approach, the scaling factor per nucleus was calculated by dividing the sum of endogenous transcript counts per nucleus by 10,000 times the previously calculated ERCC size factor. The endogenous transcript counts per nucleus were divided by this factor. Thereby, the counts in a nucleus with few endogenous transcript reads and many ERCC reads were divided by a smaller factor than counts in a nucleus with many endogenous reads and proportionally few ERCC reads. The formula used for this normalization was

$$x'_{ij} = \frac{\frac{x_{ij}}{L_j}}{\frac{\sum_j^N \frac{x_{ij}}{L_j}}{10000 \cdot sf_i}} \tag{7.1}$$

where x'_{ij} is the normalized count of gene j in cell i, L_j is the length of gene j and sf_i is the ERCC size factor of cell i. This normalization technique is referred to as ERN (ERCC ratio normalization) in the thesis.

A further filtering step was performed after normalization, in which nuclei with more than 50,000 gene length-normalized counts were removed. Out of the total 11 plates sequenced in this study, two were technical replicates of other plates. Since by containing the same nuclei as SNI-160 and SNI-116, these two technical replicate plates (SNI-234(R2) and SNI-635(R2)) did not contain additional information. Hence, they were subsequently removed from the downstream analyses. This resulted in a final normalized count matrix of 1,649 nuclei times 19,258 genes. This count matrix was then log-transformed by adding one to every value in the matrix to remove zeros and applying a natural logarithm.

When sequencing the PHHs for the PHH diversity project in this thesis, their respective genome content was unknown and ERCCs were not present. Therefore, normalization was performed using *scran*[70]. This normalization method first subsets the data into groups with similar library sizes by performing hierarchical clustering using a distance metric based on Spearman's correlation. Then, scaling factors are calculated for every cell within a group by randomly pooling subsets of cells in a group, summing their library sizes and comparing to the average library size of that group. This is performed iterative to finally de-convolute scaling factors for every single cell from the set of pool-wise size factors. Finally, normalization is performed between groups.

7.5 Batch correction

The human cell atlas aiming to identify cell types present in humans has shown that cell types are similar to each other across biological replicates [39]. However, performing experiments on different days, different machines, or eventually in different laboratories can lead to unwanted variation in the data, purely driven by technical artifacts. In order to gain biologically meaningful results, these effects need to be removed before inferring differences between the conditions of interest. Experimentally, unwanted technical variation can be kept low by processing all conditions of interest at the same time in one batch. However, this is not always feasible and computational methods have been developed to adjust for technical variation in the data. In the Ploidy&Zonation project, every plate usually contained a mixture of diploid and tetraploid nuclei stemming from either one or two mice to minimize batch effects. While parameters were mostly kept consistent, different ERCC dilutions were used and plates were processed at different laboratories, rendering batch effect correction between plates an important tool to remove unwanted technical variation caused by these circumstances. Therefore, the tool *combat* was used to with the plates as a covariate in the model [127]. Briefly, this method first builds a linear model to assign similar values of mean and variance for all genes before applying Bayes' Theorem to estimate the batch adjustments. These adjustments are then used to change the values within the count matrix, removing the unwanted variation.

Due to different sequencing depths being used for the two pairs of biological donors in the PHH diversity project of the thesis, batch effects were expected to be larger than for the snRNA-seq2 data set. Therefore, the integration of the two batches was an important and major focus of this project to disentangle the biological sources of cellular heterogeneity. In a first step, each of the four individual donors was analyzed separately to identify similar sub-types of hepatocytes between donors. For this individual donor analysis, the two batches were filtered and normalized separately. Briefly, for the first batch, cells were kept if they had between 1,000 and 250,000 reads per cell and at least 500 genes covered. Genes were kept if they were covered at least once in at least five cells. For the second batch, cells were kept if they had more than 3,000 reads and at least 500 genes covered, and genes were kept if they had fewer than 5,000,000 reads and were present in at least five cells. In both cases, cells with more then 1% mitochondrial reads were removed and *scran* was used for normalization. While in the first batch, pre-clustering was performed and *scran* was used with parameter *min_mean = 0.01*, *min_mean* was set to 0.001 for the second batch due to the differences in sequencing depth between the batches. After normalization, cells with more than 100,000 normalized counts were removed in the first batch. In the second batch, cells were removed if they had more than 20,000 normalized counts. To identify similar subgroups of hepatocytes between all four human donors, *Louvain* clustering - as described in greater detail in the section below - was performed in such manner that every cluster contained all treatment conditions. For all donors, 20 principle components (PCs) were calculated and the neighboring graph was constructed with 50 neighbors and 10 of the PCs. The following table describes the *Louvain* resolution that was used for each of the four donors and the number of resulting clusters.

Table 7.1: The *Louvain* resolution per human donor and resulting number of clusters

Donor identifier	Resolution	Number of clusters
HUM4152	0.75	7
HUM180812	0.25	4
HUM4190	0.4	5
HUM181641	0.25	4

Comparison of the resulting clusters between donors was done by calculating the top 1,000 genes per cluster before determining the percentage of their overlaps between clusters from different donors. Performing hierarchical clustering to identify groups of similar *Louvain* clusters resulted in the detection of four groups of *Louvain* clusters, of which three contained clusters of all four donors whereas one only contained clusters stemming from the first batch. Two additional *Louvain* clusters, stemming from one donor of the first, and a donor of the second batch, respectively, were not assigned to a group and treated as individual clusters. Cells from all donors where then pooled together computationally and annotated according to the identified shared clusters. Filtering and normalization was performed as described above, and *Harmony* and *scGen* were used as data integration tools to combine the two batches. Briefly, *scGen* predicts the impact of a condition on gene expression through a variable autoencoder (VAE) depending on the low-dimensional representation of the cells. Here, the condition of interest is the batch from which each cell is coming and the goal is to remove its impact on gene expression through the model by adapting the expression values. *scGen* is a supervised method relying on the accurate identification of cell sub-types before integration. This can limit the detection of cellular heterogeneity beyond the pre-identified sub-types. Therefore, and since the power to detect cellular sub-types increases with the number of cells studied, *Harmony* was also used as an unsupervised method for data integration. The aim was to clarify whether different integration approaches would lead to the same result, rendering the detection of sub-types robust and reliable. Due to the consistency between methods and the advantage given by an unsupervised method, *Harmony* was used on the combined, filtered and normalized data set [128]. *Louvain* clustering on the *Harmony* integrated data set led to the detection that some of the cells from the second batch that had been assigned to subgroup II were now clustering together with cells from subgroup IV. Therefore, they were labelled as subgroup IV and the sub-type annotation was transfered to the *scGen* integrated data set to investigate *Louvain* clustering with respect to those cells. Indeed, these cells were grouped together at the center of the UMAP based on the embedding after *scGen* integration. However, due to the supervised nature of the *scGen* integration approach, the annotation performed beforehand limited their correct assignment. Based on these analyses, the final decision was made to use *Harmony* as an unbiased data integration tool in this analysis. *Harmony* projects cells from different batches into a shared embedding, in which cells are grouped by their cell type specific gene expression rather than the variation stemming from batch effects [128].

7.6 Clustering and annotation

Sequencing the whole transcriptome in mice and humans results in several thousands of genes being detected. Accordingly, a total of 19,258 genes were present in the snRNA-seq2 data set after filtering. This yields a high dimensionality, in which not every factor adds relevant information. Hence, this dimensionality was reduced by performing a principal component analysis (PCA), where factors are ordered by the amount of variance they explain in the data. To that end, the first 50 principal components (PCs) were calculated. For the purpose of later grouping nuclei with similar gene expression profiles together during clustering, these 50 PCs were used to construct a neighborhood graph based on the 15 nearest neighbors per nucleus. The connectivities between nuclei were calculated by Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [239]. For the purpose of visualization, the nuclei were then embedded in a t-distributed stochastic neighborhood embedding (tSNE) with a perplexity of 30 and using the first 15 out of the 50 calculated PCs. In addition to that, UMAP and PCA embeddings were also used for visualization. To group together nuclei with similar gene expression, *Louvain* clustering was performed on the neighbourhood graph. In the Ploidy&Zonation project, an initial resolution of 0.2 was used to computationally separate the nuclei of non-parenchymal liver cells from the hepatocyte nuclei. Hepatocytes were then temporarily removed from the normalized expression matrix to further dissect non-parenchymal cell types. A more in depth *Louvain* clustering with a resolution of 1 was performed on the non-parenchymal nuclei, revealing the presence of nine clusters. To annotate these clusters as cell types, established marker genes were taken and adapted from Aizarani *et al.* and PanglaoDB [19][50]. First, the expression levels of these marker genes were visually investigated on the t-SNE and in violin plots to obtain a rough idea of cell type belonging. Subsequently, differential expression analysis was performed between clusters and the top differential expressed genes per cluster were investigated for overlaps with the list of known marker genes. Furthermore, nuclei with an expression level of greater than 50 for the marker Epcam were annotated as Epcam-positive epithelial cells. As an approach to separate nuclei coming from bi-nucleated tetraploid (2x2n) hepatocytes from nuclei from truly diploid hepatocytes, the neighborhood graph was re-calculated with 300 neighbors for the hepatocytes and *Louvain* clustering was performed at resolution 1 and nuclei were again embedded in t-SNE and UMAP. This yielded no substructure in the hepatocytes that would justify the annotation of 2x2n hepatocytes in the data.

For analyzing the four human donors individually in the PHH diversity project, the top 20 PCs were determined, of which the top 10 were used to calculate the neighborhood graph with 50 neighbors. These parameters were chosen to capture larger groups of hepatocytes based on the main drivers of variation in the data. After the data from the four donors had been combined and normalized using *scran* as described above, the top 50 PCs of the combined data set were computed and a neighboring graph was constructed with 15 neighbors in order to calculate a UMAP embedding and *Louvain* clusters. As mentioned before, hepatocytes are a large cell type which diminishes their encapsulation probability. Hence, for two of the samples from a donor in the second batch, the encapsulation of the cells in liquid droplets was partially unsuccessful during the 10X library preparation. To identify the cells stemming from these samples, an initial *Louvain* resolution of 0.5 was used with the goal to separate incorrectly encapsulated cells from the rest, and potentially rescue cells in the failed samples which were encapsulated correctly. It was observed that cells from one of these samples indeed clustered apart from the other cells, and two other clusters containing cells from the failed samples showed fractal structures not present in the samples in which cells were encapsulated correctly. Hence, these clusters - that mainly contained cells from the failed samples - were removed, which resulted in a final matrix of 38,232 cells times 16,256 genes. After that, the group labels from the individual donor analysis were added as annotation, again the top 50 PCs were calculated and a neighborhood graph was built with 15 neighbors. UMAP embedding was computed and *Louvain* clustering was performed with resolution of 1, leading to the detection that some of the cells from the second batch, that were initially annotated as subgroup II, were now clustering with subgroup IV cells from the first batch. Furthermore, one of the two donor-specific clusters that had not been assigned to one of the shared groups between donors showed separation into several clusters on the combined data set that were hence annotated as subgroup I, and II, depending on the annotation of their surrounding cells in the *Louvain* cluster. The other donor-specific cluster was assigned to subgroup IV. Several *Louvain* clusters were aggregated to assure that every subgroup of hepatocytes contained cells from all donors. To correctly aggregate the *Louvain* clusters and subsequently annotate the resulting hepatocyte subgroups, expression of marker genes was investigated in the *Louvain* clusters. Thus, known marker genes for specific metabolic pathways were taken from literature and can be found in the Supplementary Material of Sanchez-Quant *et al.* [164]. A limiting feature of primary human hepatocytes in culture is their loss of characteristic hepatocyte-like transcriptional signature along culture time [115]. Accordingly, one subgroup of hepatocytes was identified for which any given gene was expressed in a lower percentage of cells than for the other three subgroups. Initially, this subgroup had only been detected in the first batch and cells from the second batch were only annotated accordingly after integration. This observation was most likely due to the first batch being sequenced more deeply, allowing to better separate hepatocytes losing their characteristic expression from dropouts.

7.7 Alternative splicing

Kallisto was used to align the reads from the Ploidy&Zonation project to the mm10 reference genome with respect to different transcript isoforms of a gene being present in the nucleus. After subsetting the nuclei to the set of previously annotated nuclei, this process yielded a matrix of 2,496 nuclei times 118,489 transcripts. Furthermore, transcripts were removed if they were present in fewer than five nuclei and only hepatocytes were considered, resulting in 1,061 hepatocytes and 56,357 transcripts. Key hepatocyte-specific marker genes were selected to explore the expression of alternative transcripts between 2n and 4n hepatocytes for these genes (Figure 3.5).

7.8 Cell cycle analysis

During the FAC-sorting of single nuclei into the wells of 384-well plates performed by Dr. Ioannis Deligiannis for the Ploidy&Zonation project, the intercallating DNA staining dye Hoechst was used. This allowed to separate diploid from tetraploid hepatocyte nuclei by genome content based on flourescence intensity. However, it was observed that some presumably tetraploid nuclei were present in the cluster of non-parenchymal cells. Non-parenchymal cells remain diploid throughout the lifespan. Therefore, one potential explanation would be that nuclei during cell division were sorted as tetraploid due to the genome duplication during replication. Hence, cell cycle analysis was performed using *cyclone* to assign nuclei to cell cycle phases [144].

Cultured hepatocytes are not expected to divide. However, it is possible that hepatocyte were captured during cell division when taken from the donors. Therefore, cell cycle analysis was performed in the same manner for the PHH diversity project, highlighting that one of the identified subgroups could be described as being in division.

7.9 Differential expression analysis

To address differences between groups of interest in the Ploidy&Zonation project, differential expression analyses were performed on all genes present in the count matrix by using Welch's t-test, implemented in the *scanpy* function *rank_genes_groups*. For the non-parenchymal cells this was partially done to identify known cell type marker genes among the top differential expressed genes (DEGs) but also to identify differences between the cell types. Moreover, differential expression analysis was done between diploid and tetraploid hepatocytes. In all these analyses, genes were defined as significantly up-regulated if they had a Bonferroni-adjusted p-value below 0.05 and a $\log_2$ fold change above 0.5. Likewise, genes were defined as significantly down-regulated if they had a Bonferroni-adjusted p-value below 0.05 and a $\log_2$ fold change below -0.5. For the purpose of visualizing the DEGs in an MA-plot without over-plotting, only the genes with a mean expression between 0.1 and 100 were depicted. Furthermore, to generate informative heatmaps, the R package *ComplexHeatmap* was used with 40 randomly selected nuclei per cell type and their top five up-regulated genes. For cell types, of which fewer than 40 nuclei were present, all nuclei were taken. Additionally to the differential expression analysis, changes in the distribution of genes were analyzed between 2n and 4n hepatocytes. Thus, a Kolmogorov-Smirnov test was performed, based on which genes were defined as significantly changing their expression distribution between the two ploidy states if they had a p-value below 0.05 and a test statisitc greater than 0.15.

For the PHH diversity project, differential expression analysis was mainly performed to investigate the impact of a given treatment condition towards DMSO levels in the functional subgroups of hepatocytes. In the same fashion as for the snRNA-seq2 data set, Welch's t-tests were performed for every gene between groups of interest. When comparing the expression profiles of Cocktail- and DMSO-treated cells, genes were defined as significantly up-regulated if they had a Bonferroni-adjusted p-value below 0.05 and a $\log_2$ fold change above 1. Similar, genes with adjusted p-value below 0.05 and $\log_2$ fold change below -1 were dedicated significantly down-regulated. In the comparison of FFA-treated cells to DMSO, an average of 3.5 times fewer genes with positive $\log_2$ fold change were detected than in the comparison of Cocktail to DMSO. Hence, the cutoff of $\log_2$ fold change to call genes significantly up- or down-regulated was adjusted to characterize the subtler impact of intracellular lipid accumulation. Accordingly, genes were defined as significantly up- or down-regulated if they had an adjusted p-value below 0.05 and a $\log_2$ fold change higher than 0.75 or below -0.75, respectively. Moreover, to explore the drivers of functional specialization between subgroups of PHHs, the top 500 DEGs per subgroup were calculated considering only DMSO-treated cells and used to predict the underlying transcription factor networks through the online tool *ChEA3* [240]. Out of the top 25 predicted transcription factors per subgroup, five were used to visualize their differential expression between subgroups.

7.10 Gene ontology and gene set enrichment analysis

Based on data bases assigning genes to functional groups of pathways (ontologies), the lists of DEGs between groups can be used to find enrichment into pathways, helping to further characterize biological differences between groups of interest. Therefore, enrichment analysis was done using a Fisher's exact test as implemented in the python package *gprofiler*. When characterizing the differences between 2n and 4n hepatocytes, as well as the subgroup-specific impacts of intracellular lipid accumulation, this analysis was done with focus on the gene ontology "biological processes (BP)". In both cases, a top number of pathways were depicted in a dot plot with the pathway names on the y-axis, the gene ratio on the x-axis, dot color referring to significance level and the dot size depicting gene overlap.

When comparing Cocktail- to DMSO-treated cells per subgroup, the focus lied on identifying the subgroup-specific differences in drug metabolism. Hence, only the subgroup-specific significantly up-regulated genes upon Cocktail treatment were chosen for gene ontology analysis. Furthermore, the online tool *ShinyGO* was used as it allows accessing a data base containing pathways known to be involved in the metabolism of specific drugs [241][178][179]. Hence, upon Cocktail treatment, drug metabolism pathways specific for each of the functional subgroups could be identified. For visualization purposes, respective pathways were depicted on the x-axis, significance level on the y-axis and dot size referred to gene overlap.

In addition, the effect of intracellular lipid accumulation on gene expression of drug metabolism related genes, gene set enrichment analysis (GSEA) was performed on genes significantly up-regulated upon Cocktail treatment but not in the FFA+Cocktail treatment condition [242][243]. This showed higher expression of these genes in Cocktail than in FFA+Cocktail.

7.11 Co-expression analysis of markers genes

Another approach to gain insights into the functional commitment of single cells is to explore co-expression of key marker genes. For example, studies have suggested pericentral hepatocytes to serve as a stem cell niche in liver regeneration [244] while others show even distribution of liver stem cell markers across the liver lobule [153][154][155]. Hence, as an indicator of regenerative potential, the co-expression of liver stem cell marker genes was analyzed in all hepatocytes from the Ploidy&Zonation project, and separately for diploid and tetraploid hepatocytes. Namely, the marker genes *Prom1, Icam1, Sox9, Afp, Epcam, Axin2, Itga6, Tert, Notch2, Tbx3,* and *Lgr5* were investigated. To address whether these genes were co-expressed in a nucleus, gene expression was first assigned to binary values based on whether a gene was expressed or not. Thereby, expression levels of the genes were neglected and only the presence or absence of a gene in a nucleus was considered. Then, nuclei not expressing any of the eleven stem cell markers were excluded, yielding a binary matrix of 364 nuclei times 11 stem cell markers. This matrix was used to calculate the pairwise Jaccard distances between marker genes and between nuclei that were further used to calculate linkage for hierarchical clustering. The formula to calculate pairwise Jaccard distances can be described as follows:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{7.2}$$

where X is the binary expression vector of a gene j_1 across all nuclei, and Y is the binary expression vector of gene j_2 across all nuclei.

This co-expression analysis was also performed on the data from the PHH diversity project to investigate co-expression of genes involved in the three phases of drug metabolism within the identified subgroups. Furthermore, in this data set, co-expression of marker genes specific to different metabolic pathways was investigated with respect to differential expression levels through the calculation of scores. Briefly, scores for functional groups of marker genes were calculated for every treatment condition per subgroup in the following way: in every group of marker genes, the normalized expression of a given gene was summed up over all cells per subgroup in a treatment condition and weighted by whether it was significantly up-regulated in this subgroup under this treatment condition (multiplied by 1.2 if it was and by 0.8 if it was not). Then the mean weighted expression of the individual genes in a marker gene group was determined and divided by the number of genes present in that group.

7.12 Transcriptional variability

Transcriptional variability can be measured by calculating the coefficient of variation of a given gene. This is usually defined as the standard deviation divided by the mean. However, genes with low expression values have a higher standard deviation, and therefore, higher transcriptional variability. Furthermore, calculating the coefficient of variation in the classic way on log-transformed, normalized expression data may lead to inaccurate values [138]. Hence, to reduce the impact of lowly expressed genes, in the Ploidy&Zonation data set, genes that had a mean log-transformed, normalized expression below 0.25 were removed, leading to the remaining of 2,102 genes in the matrix. Likewise, in the PHH diversity data set, this filtering led to 3,434 remaining genes. To correctly calculate the coefficient of variation (CV) on the normalised, log-transformed and batch-corrected matrix the formula described in Canchola *et al.* was used [138]:

$$CV = \sqrt{e^{\sigma^2} - 1} \tag{7.3}$$

where σ^2 is the variation of gene j in the group of interest.

To calculate whether significant differences in transcriptional variability were present between groups of interest, Mann-Whitney-U tests were performed. This was done to either infer differences between 2n and 4n hepatocytes or the impact of treatment condition on transcriptional variability in subgroups of primary human hepatocytes. Additionally, for the Ploidy&Zonation data set, highly variable genes in 2n and 4n hepatocytes were defined as genes with a coefficient of variation greater than 1.7. To identify this specific threshold, the ERCC reads were used as a measure of baseline, non-biological variability. In brief, the ERCC reads were normalized in the same way as the endogenous transcripts, but with 1,000 instead of 10,000 as a scaling factor. The ERCC reads were then log-transformed and the coefficient of variation of every ERCC was calculated, yielding 1.66 as the median plus one standard deviation coefficient of variation for the ERCC reads. Hence, genes with greater coefficients of variation were defined as highly variable.

7.13 Pseudotemporal ordering

Transcriptome data can be used to construct a trajectory of cells or nuclei, ordering them based on their gene expression values. In differentiating cells, this can be used to computationally infer changes between cell states along the constructed pseudo-time. As hepatocytes in culture are known to lose their characteristic hepatocyte-like expression along culture time, this could potentially be used to re-construct the process of expression loss in primary human hepatocytes in the PHH diversity project. Hence, diffusion pseudotime was calculated for the cultured hepatocytes based on key hepatic transcription factors. However, the ordering of cells based on their expression profiles did not correspond to the levels of gene expression.

Apart from constructing pseudo-time, this type of analysis can also be used to infer spatial relationships between cells, particularly in the presence of a known expression gradient as is observed for liver zonation. Thereby, hepatocytes can be assigned to different areas of the liver lobule based on their expression profiles. For the Ploidy&Zonation project, zonation markers were obtained from literature, specifically from Halpern *et al.* and Aizarani *et al.* [16][19]. Since only endothelial cells and hepatocytes are subject to zonation and the focus of the Ploidy&Zonation project lied on investigating the relationship between liver zonation and polyploidization in hepatocytes, non-parenchymal cells were removed. This resulted in a matrix of 1,061 hepatocyte nuclei times 1,742 zonation marker genes. For this matrix, principle components, the neighborhood graph using 15 nearest neighbors, as well as t-SNE and UMAP embeddings were re-calculated. Additionally to PCA, diffusion components were calculated as a means of dimensionality reduction with respect to ordering the nuclei based on their expression profiles [90]. *Louvain* clustering was then performed with resolution of 0.7. According to the expression levels of zonation marker genes, clusters 0 and 2 were aggregated into a pericentral cluster, and cluster 1 was annotated as periportal cluster. This allowed to calculate the percentage of 2n and 4n hepatocytes assigned to pericentral (CV), and periportal (PV), respectively, yielding a 1.3-time enrichment of 4n nuclei in the pericentral cluster. For visualization purposes, the tetraploid nuclei were sub-sampled to the number of diploid hepatocyte nuclei. Differential expression analysis was performed between the CV and the PV cluster and the top 30 DEGs per cluster were used for visualization along the vector of pseudospace. Additionally, the pseudospace vector was divided into ten bins and the mean expression of representative zonation marker genes was calculated per bin to depict in a line plot.

In the same manner, spatial relationships between hepatocytes were inferred in the publicly available human liver data [19][18] to explore the relationship between zonation and the formation of hepatocyte subgroups. As described further below, the hepatocyte subgroups characterized in the PHH diversity *in vitro* data set were identified in two publicly available human *in vivo* data sets. The study from Aizarani *et al.* reported 35 zones along the pericentral-periportal axis, into which zonation marker genes were grouped. For the purpose of assigning the hepatocytes from this study to three major zones, pericentral, mid, and periportal, the number of zones was reduced by grouping the zonation markers into three groups through binning. Then, the hepatocytes were scored for the expression of markers of each of the three zones using the *scanpy* function *sc.tl.score_genes*. Based on the respective ranges of scores, appropriate thresholds were chosen to assign the hepatocytes to zones. Specifically, the cells from

Aizarani et al. were assigned to the pericentral zone if they had a CV score greater than 0.45 and a PV score below 1. Furthermore, cells with a PV score of at least 0.65 and a CV score less or equal than 0.45 were assigned to the periportal zone. The remaining cells were assigned to the mid-zone. To explore the contributions of the subgroups to each of the zones, the percentages of cells in CV, mid, and PV were calculated per subgroup. Additionally, zonation within subgroups *in vivo* and *in vitro* was investigated through marker gene expression. To that end, zonation marker genes in every subgroup were depicted on UMAPs, individually for the respective subgroup. After integrating the two *in vivo* data sets, the same analysis was performed on the combined data. Adjustments were made to the thresholds to assign the hepatocytes from both data sets to the three zones. Hence, cells with a CV score greater than 0.4 and a PV score below 1 were assigned pericentral; cells with a PV score of at least 0.8 and a CV score less or equal to 0.4 were assigned to periportal; the remaining cells were assigned to mid-zone. Calculating the percentage of subgroups per assigned zone further confirmed that all subgroups were present in the three zones with subgroup I being the most split into CV and PV and subgroup II and III being more present in the mid and CV zone. Since zonation had been observed to influence marker gene expression within the subgroups, the subgroups were analyzed separately to investigate whether diffusion pseudospace would refer to zonation. However, exploring this on the integrated data set showed that the first diffusion component referred to number of genes per cell, thus meaning that the cells were ordered according to their coverage instead of biological properties. The number of genes per cell was therefore regressed out using *sc.pp.regress_out*. After that, diffusion pseudospace was calculated per subgroup on the zonation marker genes. To keep the pseudospatial order with respect to zonation consistent between subgroups, the root cell was chosen as the cell with the highest PV score within a subgroup. The vector of pseudospace was then divided into five bins and the mean expression of two representative zonated marker genes per bin were depicted in a line plot for each subgroup.

7.14 Comparison to other data sets

The snRNA-seq2 data in the Ploidy&Zonation project was based on the development of a new laboratory protocol for the isolation of nucleic acids from single nuclei featuring the addition of a new lysis buffer (LB2), work done by Dr. Ioannis

Deligiannis. Hence, to highlight the advantages of this new protocol, comparisons with other approaches were performed. Firstly, to address the benefit of the new lysis buffer, one plate was processed without the additional lysis buffer. Preprocessing and filtering for reads from that plate were performed exactly as for the other plates. The number of genes per nucleus produced by either protocol were compared, showing that the addition of LB2 resulted in a greater number of genes per nucleus. Additionally, the total number of protein coding genes yielded by both protocols was compared. Despite yielding fewer protein coding genes and a lower number of genes per nucleus, the plate where LB2 had not been added was processed for any potential downstream analyses. Hence, nuclei with fewer than 500 genes and 5,000 counts were removed. Furthermore, the results from the snRNA-seq2 approach were compared to publicly available data sets. Hence, data were either downloaded from GEO (accession numbers GSE84498, GSE124395, GSE1483395) or from https://doi.org/10.6084/m9.figshare.5829687.v7 and https://doi.org/10.6084/m9.figshare.5968960.v215. In the case of raw, unfiltered data, genes were removed if they were not expressed in any nucleus or cell and nuclei or cells were filtered out if they had no genes expressed. For the snRNA-seq2 data and these publicly available data sets, the numbers of genes per nucleus or cell was visualized as violin plots. Only protein coding genes were used to compare the total number of detected genes between data sets. The overlaps of protein coding genes detected by the different methods was depicted in Venn diagrams (https://github.com/LankyCyril/pyvenn/blob/master/pyvenn-demo.ipynb). Finally, to assess whether nuclei appropriately represent the gene expression of whole single cells, gene expression was correlated between the snRNA-seq2 data and the Tabula muris smart-seq2 data. For both data sets, only hepatocytes were considered and the respective raw count matrix on exon reads was taken. Pearson correlation was then calculated between the $\log_2$ mean gene expression of genes present in both data sets.

In addition to investigating the benefits of LB2 and comparing to publicly available data sets, the data were further compared to 10X experiments. First, 10X experiments that had been previously performed by Dr. Celia Martinez-Jimenez were analyzed by aligning the reads to mm10 using 10X Genomics Cell Ranger v4.0.12. Genes were removed that were not present in any nucleus and nuclei were removed that had no genes expressed. Again, the number of genes per nucleus and the number of overall detected protein coding genes were compared between methods (10X or plate-based, with and without LB2). The eventual goal was to address whether similar levels of cellular heterogeneity within hepatocytes could be detected through performing a 10X experiment at similar sequencing depth as in the plate-based approach. Hence, an additional 10X experiment on FAC-sorted 4n nuclei treated with an additional lysis buffer was performed by Dr. Ioannis Deligiannis. This experiment aimed to obtain 550,000 reads per nucleus or more, in concordance to the snRNA-seq2 data. The resulting reads were also aligned to mm10 using 10X Genomics Cell Ranger v4.0.012 and nuclei without any genes detected as well as genes not sequenced in any nuclei were removed. This additional 10X experiment achieved a mean library size of roughly 1,000,000 reads per single nucleus, yielding a median of 2,776 genes per nucleus.

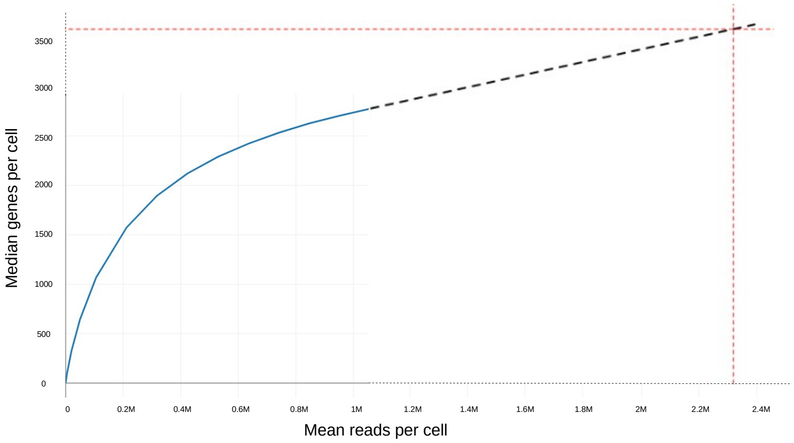

Figure 7.1: Blue line plot showing the saturation curve obtained from running cellranger on a deeply sequenced 10X experiment. Dashed black line shows extrapolation from the end point of the curve. Dashed red lines indicate the hypothetical sequencing depth needed to achieve 3,600 genes per nucleus.

Based on the sequencing saturation curve provided by the Cell Ranger output, extrapolation was done assuming linear continuation from the end of the curve to address the question what library size would be needed to reach a median of around 3,600 genes per nucleus as observed in the snRNA-seq2 data set. This extrapolation was done by taking two data points near the end of the saturation curve to calculate a slope, from which it was then estimated at what sequencing depth a median of 3,600 genes per nucleus would be reached. In this scenario, an average sequencing depth of 2.3 million reads per nucleus would have been required to reach a median of 3,600 genes per nucleus. Additionally to the newly generated 10X data, the data from the first batch of the PHH diversity project was used for comparison in the same manner. This particular first batch had an average sequencing depth of around 700,000 reads per cell, yielding a median of 2,906 genes per cell. Again, data points at the end of the saturation curve were considered for the calculation of a slope to extrapolate at what sequencing depth 3,600 genes per cell would be reached. In that data set, a hypothetical sequencing depth of 2.7 million reads per cell would have been needed to reach the same amount of genes per cell as observed in the snRNA-seq2 data.

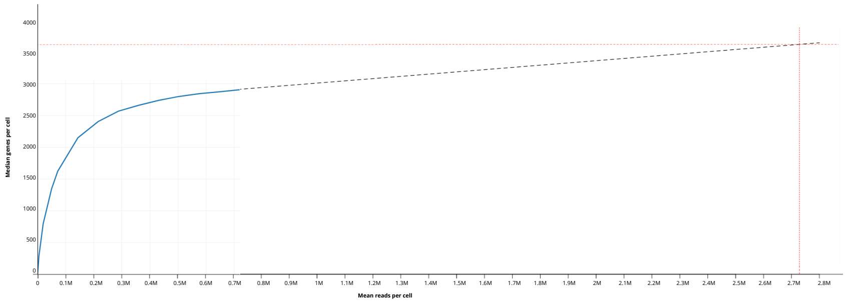

Figure 7.2: Blue line plot showing the saturation curve obtained from running cellranger on the first batch of the precision toxicology data set on PHHs performed in a 10X experiment. Dashed black line shows extrapolation from the end point of the curve. Dashed red lines indicate the hypothetical sequencing depth needed to achieve 3,600 genes per nucleus.

While for the snRNA-seq2 data set comparisons to other data sets were done with the purpose of highlighting the advantages of the new method, in the PHH diversity project, the data was compared to publicly available data to address to what extent *in vivo* zonation is related to the identified subgroups. For this purpose, data from nine human livers were obtained from GEO (Accession number GSE124395)[19]. In a first step, genes that were not present in any cell and cells that did not express any genes were removed. Moreover, only cells with between 100 and 6,000 genes and 800 and 30,000 reads were kept. Finally, genes were removed if they were sequenced in fewer than 10 cells, resulting in a count matrix of 11,059 cells times 19,416 genes. To keep the data sets comparable, normalization was done using *scran* with parameter *min.mean=0.05* and counts were log-transformed, after which cells with more than 20,000 normalized counts were removed. Dimensionality reduction and *Louvain* clustering was then performed on the remaining 11,043 cells using *scanpy* functions. An initial *Louvain* resolution of 0.08 was chosen to computationally separate the different liver cell types present *in vivo* from each other. To identify the hepatocyte cluster, the expression of key hepatocyte marker genes, such as *ALB, HNF4α*, and *TTR* was investigated. Based on this, the count matrix was subset to only contain hepatocytes, and *Louvain* clustering was performed at resolution of 0.2, resulting in 6 clusters. The marker genes used to identify hepatocyte subgroups *in vitro* were then explored on this *in vivo* data set. For instance, in comparison to the other clusters, cluster 0 showed high expression levels of *CYP8B1, HSD17B4, HSD3B7, CYP27A1, NR1H4,* and *HMGCS2*, hence justifying its annotation as subgroup I, specializing in bile and sterol metabolism. Likewise, clusters 1 and 4 highly expressed markers such as *SULT2A1, LDHA,* and *GAMT* that had been observed in subgroup II *in vitro*. Finally, clusters 2, 3, and 5 were defined by expressing phase III marker genes, such as *ABCC2* and *ABCC3* as well as lipid metabolism marker genes, such as *PLIN5*.

To investigate the relationship between the annotated subgroups and liver zonation *in vivo*, zonation markers were obtained from Aizarani *et al.*[19]. In order to keep the results easy to interpret, the genes defining each of the 35 zones reported in the original study were grouped into three zones, representing pericentral (CV), midzonal, and periportal (PV) marker genes. The cells were then scored for the expression of the marker genes for each of these three zones using the *scanpy* function *sc.tl.score_genes*. Cells with a CV score greater than 0.45 and a PV score below 1 were assigned pericentral. Cells with a PV score greater or equal to 0.65 and a CV score smaller or equal to 0.45 were assigned periportal. The remaining cells were assigned mid-zone. To address the relationship between the subgroups and zonation, the percentage of cells in each of the three zones was calculated per subgroup and depicted in a bar plot. Furthermore, the influence of zonation within each of the subgroups was analyzed by investigating the expression of zonated marker genes in every subgroup. The same analysis was performed on the *in vitro* data set yielding similar CV and PV scores across all cells and no zonation patterns of marker genes in the subgroups, confirming that zonation is not conserved in 2D culture while subgroup specialization remains. Additionally to the data set from Aizarani *et al.*, data from five human livers were downloaded from GEO (Accession number GSE115469)[18] to further confirm the presence of hepatocyte subgroups that are independent of zonation *in vivo*. As this data set generated by MacParland *et al.* had already been pre-processed and normalized, additional filtering included only the removal of genes that were sequenced in fewer than three cells. For better comparability of the three data sets, the normalized counts were log-transformed. The original study's annotation was used to computationally isolate the hepatocytes. *Louvain* clustering was performed at a resolution of 0.5 to separate subgroups of hepatocytes, resulting in six clusters showing high similarity to the clusters reported by the original study. Again, the expression of the subgroup-defining marker genes was investigated in these clusters leading to the successful identification of the subgroups in this *in vivo* data set. Expression of zonation marker genes was also investigated in this data set, again showing that the subgroups were independent of zonation while zonation markers showed distinct patterns within the subgroups. To increase the power for the identification of similar hepatocyte subgroups *in vitro* and *in vivo*, the two *in vivo* data sets were integrated using *scGen* [129]. The concordance of hepatocyte subgroups between *in vitro* and *in vivo* was quantified by calculating the top 10 DEGs for every *in vitro* subgroup and correlating their scaled mean expression to their scaled mean expression in the integrated *in vivo* data. Furthermore, combining the *in vivo* data sets also increased the power for the accurate assessment of the relationship between zonation and subgroup specialization as described above.

Acknowledgments

First of all, I want to thank my primary supervisor Dr. Celia Martinez-Jimenez for giving me the opportunity work on such exciting projects to unravel cellular heterogeneity in the liver and for allowing to shape the projects through my own ideas. I would also like to thank my second supervisor, Prof. Dr. Maria Colomé-Tatché for all her advice and feedback during the last years. Moreover, thank you to the members of my thesis advisory committee, Dr. Antonio Scialdone and Prof. Dr. Wolfgang Enard. Your input was highly valuable for the progression of this thesis.

My gratitude also goes to the administration team of the Helmholtz pioneer campus for providing solutions for any organizational issues. Additionally, I would like to thank the graduate school HELENA and the epigenetics research school for all the support! It was a pleasure to take part in interesting courses, and to be part of an amazing team of students from an international background that are involved in different aspects of epigenetic research.

Furthermore, thank you to all the co-authors of the publications resulting from this thesis. Especially to Dr. Catalina Vallejos for hosting me in the lovely city of Edinburgh and sharing her knowledge on assessing transcriptional variability from scRNA-seq data. Thank you also to Dr. Kelvin Yin for performing validation experiments for the findings from the snRNA-seq2 study making this study more powerful. Special thank you to Anna Danese for always being there for me to answer any question, be it related to data analysis or not. Of course, I hereby want to express my gratitude to all current and former members of both, the Martinez-Jimenez, and the Colomé-Tatché lab. Thank you for all the scientific and non-scientific discussions, for arranging virtual home-office coffee breaks to keep exchange ongoing during pandemic times, and for generally creating a nurturing working atmosphere! I especially want to thank Dr. Rizqah Kamies for her constant availability to proofread any manuscript and correct its grammar and style as well as for all the fruitful discussions we had. Moreover, I would like to particularly thank Eva Sanchez-Quant for all her work, her enthusiasm and initiative, and of course, for the cute Christmas decorations every year! It was tremendously fun to work with you and I think we were an amazing team, perfectly complementing each other's skill set and personality.

Last but not least, I would like to thank my family and friends, and especially my husband for their constant supply of love, occasional snacks and the reminder of a life beyond the screen.

Bibliography

[1] Elijah Trefts, Maureen Gannon, and David H. Wasserman. **The liver**. *Current biology : CB* 27:21 (Nov. 2017), R1147–R1151. ISSN: 0960-9822. DOI: 10.1016/j.cub.2017.09.019. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5897118/ (visited on 05/03/2021) (see pages iii, v, 1, 2, 11).

[2] Germain Margall-Ducos, Séverine Celton-Morizur, Dominique Couton, Olivier Brégerie, and Chantal Desdouets. **Liver tetraploidization is controlled by a new process of incomplete cytokinesis**. *Journal of Cell Science* 120:20 (Oct. 2007), 3633–3639. ISSN: 0021-9533. DOI: 10.1242/jcs.016907. URL: https://doi.org/10.1242/jcs.016907 (visited on 02/09/2022) (see pages iii, v, 7).

[3] Shuyuan Zhang, Kejin Zhou, Xin Luo, Lin Li, Ho-Chou Tu, Alfica Sehgal, Liem H. Nguyen, Yu Zhang, Purva Gopal, Branden D. Tarlow, Daniel J. Siegwart, and Hao Zhu. **The Polyploid State Plays a Tumor-Suppressive Role in the Liver**. en. *Developmental Cell* 44:4 (Feb. 2018), 447–459.e5. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2018.01.010. URL: https://www.sciencedirect.com/science/article/pii/S1534580718300108 (visited on 02/09/2022) (see pages iii, v, 7, 46).

[4] Séverine Celton-Morizur, Grégory Merlen, Dominique Couton, Germain Margall-Ducos, and Chantal Desdouets. **The insulin/Akt pathway controls a specific cell division program that leads to generation of binucleated tetraploid liver cells in rodents**. eng. *The Journal of Clinical Investigation* 119:7 (July 2009), 1880–1887. ISSN: 1558-8238. DOI: 10.1172/jci38677 (see pages iii, v, 1, 7, 90, 91, 104).

[5] Myriam Bou-Nader, Stefano Caruso, Romain Donne, Séverine Celton-Morizur, Julien Calderaro, Géraldine Gentric, Mathilde Cadoux, Antoine L'Hermitte, Christophe Klein, Thomas Guilbert, Miguel Albuquerque, Gabrielle Couchy, Valérie Paradis, Jean-Pierre Couty, Jessica Zucman-Rossi, and Chantal Desdouets. **Polyploidy spectrum: a new marker in HCC classification**. eng. *Gut* 69:2 (Feb. 2020), 355–364. ISSN: 1468-3288. DOI: 10.1136/gutjnl-2018-318021 (see pages iii, v, 7, 46, 91).

[6] Shani Ben-Moshe, Yonatan Shapira, Andreas E. Moor, Rita Manco, Tamar Veg, Keren Bahar Halpern, and Shalev Itzkovitz. **Spatial sorting enables comprehensive characterization of liver zonation**. en. *Nature Metabolism* 1:9 (Sept. 2019). Number: 9 Publisher: Nature Publishing Group, 899–911. ISSN: 2522-5812.

DOI: 10.1038/s42255-019-0109-9. URL: https://www.nature.com/articles/s42255-019-0109-9 (visited on 02/04/2022) (see pages iii, v, 2, 3, 11, 12, 55, 93, 94, 96).

[7] Sung Rye Park, Chun-Seok Cho, Jingyue Xi, Hyun Min Kang, and Jun Hee Lee. **Holistic characterization of single-hepatocyte transcriptome responses to high-fat diet**. *American Journal of Physiology - Endocrinology and Metabolism* 320:2 (Feb. 2021), E244–E258. ISSN: 0193-1849. DOI: 10.1152/ajpendo.00391.2020. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8260362/ (visited on 02/11/2022) (see pages iv, vi, 15, 71, 90, 97, 104).

[8] Hannele Yki-Järvinen, Panu K. Luukkonen, Leanne Hodson, and J. Bernadette Moore. **Dietary carbohydrates and fats in nonalcoholic fatty liver disease**. eng. *Nature Reviews. Gastroenterology & Hepatology* 18:11 (Nov. 2021), 770–786. ISSN: 1759-5053. DOI: 10.1038/s41575-021-00472-y (see pages iv, vi, 15).

[9] Craig D. Fisher, Andrew J. Lickteig, Lisa M. Augustine, James Ranger-Moore, Jonathan P. Jackson, Stephen S. Ferguson, and Nathan J. Cherrington. **Hepatic Cytochrome P450 Enzyme Alterations in Humans with Progressive Stages of Nonalcoholic Fatty Liver Disease**. *Drug Metabolism and Disposition* 37:10 (Oct. 2009), 2087–2094. ISSN: 0090-9556. DOI: 10.1124/dmd.109.027466. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2769034/ (visited on 09/26/2022) (see pages iv, vi, 74, 77, 98, 101).

[10] Irene Kyrmizi, Pantelis Hatzis, Nitsa Katrakili, Francois Tronche, Frank J. Gonzalez, and Iannis Talianidis. **Plasticity and expanding complexity of the hepatic transcription factor network during liver development**. en. *Genes & Development* 20:16 (Aug. 2006). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 2293–2305. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.390906. URL: http://genesdev.cshlp.org/content/20/16/2293 (visited on 05/03/2021) (see pages 1, 43).

[11] Janie L. Baratta, Anthony Ngo, Bryan Lopez, Natasha Kasabwalla, Kenneth J. Longmuir, and Richard T. Robertson. **Cellular Organization of Normal Mouse Liver: A Histological, Quantitative Immunocytochemical, and Fine Structural Analysis**. *Histochemistry and cell biology* 131:6 (June 2009), 713–726. ISSN: 0948-6143. DOI: 10.1007/s00418-009-0577-1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2761764/ (visited on 05/03/2021) (see pages 1, 43, 93).

[12] Cristina Rodríguez-Antona, M. Teresa Donato, Eugenia Pareja, Maria-José Gómez-Lechón, and José V. Castell. **Cytochrome P-450 mRNA Expression in Human Liver and Its Relationship with Enzyme Activity**. en. *Archives of Biochemistry and Biophysics* 393:2 (Sept. 2001), 308–315. ISSN: 0003-9861. DOI:

10.1006/abbi.2001.2499. URL: https://www.sciencedirect.com/science/article/pii/S0003986101924993 (visited on 02/11/2022) (see pages 1, 16, 99).

[13] Séverine Celton-Morizur and Chantal Desdouets. “Polyploidization of liver cells.” en. In: *Polyploidization and Cancer*. Ed. by Randy Y. C. Poon. Advances in Experimental Medicine and Biology. New York, NY: Springer, 2010, 123–135. ISBN: 978-1-4419-6199-0. DOI: 10.1007/978-1-4419-6199-0_8. URL: https://doi.org/10.1007/978-1-4419-6199-0_8 (visited on 02/04/2022) (see pages 1, 8, 50).

[14] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. **Bursty Gene Expression in the Intact Mammalian Liver**. en. *Molecular Cell* 58:1 (Apr. 2015), 147–156. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.01.027. URL: https://www.sciencedirect.com/science/article/pii/S1097276515000507 (visited on 02/09/2022) (see pages 1, 46, 47, 53, 90).

[15] Hernán Morales-Navarrete, Fabián Segovia-Miranda, Piotr Klukowski, Kirstin Meyer, Hidenori Nonaka, Giovanni Marsico, Mikhail Chernykh, Alexander Kalaidzidis, Marino Zerial, and Yannis Kalaidzidis. **A versatile pipeline for the multi-scale digital reconstruction and quantitative analysis of 3D tissue architecture**. *eLife* 4 (Dec. 2015). Ed. by Fiona M Watt. Publisher: eLife Sciences Publications, Ltd, e11214. ISSN: 2050-084X. DOI: 10.7554/eLife.11214. URL: https://doi.org/10.7554/eLife.11214 (visited on 02/04/2022) (see pages 1, 94, 103).

[16] Keren Bahar Halpern, Rom Shenhav, Orit Matcovitch-Natan, Beáta Tóth, Doron Lemze, Matan Golan, Efi E. Massasa, Shaked Baydatch, Shanie Landen, Andreas E. Moor, Alexander Brandis, Amir Giladi, Avigail Stokar-Avihail, Eyal David, Ido Amit, and Shalev Itzkovitz. **Single-cell spatial reconstruction reveals global division of labour in the mammalian liver**. en. *Nature* 542:7641 (Feb. 2017). Number: 7641 Publisher: Nature Publishing Group, 352–356. ISSN: 1476-4687. DOI: 10.1038/nature21065. URL: https://www.nature.com/articles/nature21065 (visited on 02/04/2022) (see pages 1, 5, 12–14, 17, 18, 53, 84, 93, 94, 96, 124).

[17] Keren Bahar Halpern, Rom Shenhav, Hassan Massalha, Beata Toth, Adi Egozi, Efi E. Massasa, Chiara Medgalia, Eyal David, Amir Giladi, Andreas E. Moor, Ziv Porat, Ido Amit, and Shalev Itzkovitz. **Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells**. en. *Nature Biotechnology* 36:10 (Nov. 2018). Number: 10 Publisher: Nature Publishing Group, 962–970. ISSN: 1546-1696. DOI: 10.1038/nbt.4231. URL: https://www.nature.com/articles/nbt.4231 (visited on 02/04/2022) (see page 1).

[18] Sonya A. MacParland, Jeff C. Liu, Xue-Zhong Ma, Brendan T. Innes, Agata M. Bartczak, Blair K. Gage, Justin Manuel, Nicholas Khuu, Juan Echeverri, Ivan Linares, Rahul Gupta, Michael L. Cheng, Lewis Y. Liu, Damra Camat, Sai W. Chung, Rebecca K. Seliga, Zigong Shao, Elizabeth Lee, Shinichiro Ogawa, Mina Ogawa, Michael D. Wilson, Jason E. Fish, Markus Selzner, Anand Ghanekar, David Grant, Paul Greig, Gonzalo Sapisochin, Nazia Selzner, Neil Winegarden, Oyedele Adeyi, Gordon Keller, Gary D. Bader, and Ian D. McGilvray. **Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations**. en. *Nature Communications* 9:1 (Oct. 2018). Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Liver;RNA sequencing Subject_term_id: liver;rna-sequencing, 4383. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06318-7. URL: https://www.nature.com/articles/s41467-018-06318-7 (visited on 09/20/2021) (see pages 1, 2, 13, 66, 67, 95, 96, 124, 130).

[19] Nadim Aizarani, Antonio Saviano, Sagar, Laurent Mailly, Sarah Durand, Josip S. Herman, Patrick Pessaux, Thomas F. Baumert, and Dominic Grün. **A human liver cell atlas reveals heterogeneity and epithelial progenitors**. en. *Nature* 572:7768 (Aug. 2019). Number: 7768 Publisher: Nature Publishing Group, 199–204. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1373-2. URL: https://www.nature.com/articles/s41586-019-1373-2 (visited on 02/04/2022) (see pages 1, 14, 15, 25, 64, 65, 84, 95, 96, 116, 124, 129).

[20] Jesus M. Banales, Robert C. Huebert, Tom Karlsen, Mario Strazzabosco, Nicholas F. LaRusso, and Gregory J. Gores. **Cholangiocyte pathobiology**. *Nature reviews. Gastroenterology & hepatology* 16:5 (May 2019), 269–281. ISSN: 1759-5045. DOI: 10.1038/s41575-019-0125-y. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6563606/ (visited on 12/02/2022) (see page 1).

[21] Peter S Vestentoft, Peter Jelnes, Branden M Hopkinson, Ben Vainer, Kjeld Møllgård, Bjørn Quistorff, and Hanne C Bisgaard. **Three-dimensional reconstructions of intrahepatic bile duct tubulogenesis in human liver**. *BMC Developmental Biology* 11 (Sept. 2011), 56. ISSN: 1471-213X. DOI: 10.1186/1471-213X-11-56. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3192761/ (visited on 05/09/2022) (see pages 1, 94).

[22] Joan Font-Burgada, Shabnam Shalapour, Suvasini Ramaswamy, Brian Hsueh, David Rossell, Atsushi Umemura, Koji Taniguchi, Hayato Nakagawa, Mark A. Valasek, Li Ye, Janel L. Kopp, Maike Sander, Hannah Carter, Karl Deisseroth, Inder M. Verma, and Michael Karin. **Hybrid Periportal Hepatocytes Regenerate the Injured Liver without Giving Rise to Cancer**. en. *Cell* 162:4 (Aug. 2015), 766–779. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.07.026. URL:

https://www.sciencedirect.com/science/article/pii/S0092867415009034 (visited on 02/09/2022) (see pages 2, 94).

[23] Chen Ding, Yanyan Li, Feifei Guo, Ying Jiang, Wantao Ying, Dong Li, Dong Yang, Xia Xia, Wanlin Liu, Yan Zhao, Yangzhige He, Xianyu Li, Wei Sun, Qiongming Liu, Lei Song, Bei Zhen, Pumin Zhang, Xiaohong Qian, Jun Qin, and Fuchu He. **A Cell-type-resolved Liver Proteome**. *Molecular & Cellular Proteomics : MCP* 15:10 (Oct. 2016), 3190–3202. ISSN: 1535-9476. DOI: 10.1074/mcp.M116.060145. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5054343/ (visited on 12/02/2022) (see page 2).

[24] Anh Thu Nguyen-Lefebvre and Anatolij Horuzsko. **Kupffer Cell Metabolism and Function**. *Journal of enzymology and metabolism* 1:1 (2015). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4771376/ (visited on 05/31/2021) (see page 2).

[25] Peter J. Murray, Judith E. Allen, Subhra K. Biswas, Edward A. Fisher, Derek W. Gilroy, Sergij Goerdt, Siamon Gordon, John A. Hamilton, Lionel B. Ivashkiv, Toby Lawrence, Massimo Locati, Alberto Mantovani, Fernando O. Martinez, Jean-Louis Mege, David M. Mosser, Gioacchino Natoli, Jeroen P. Saeij, Joachim L. Schultze, Kari Ann Shirey, Antonio Sica, Jill Suttles, Irina Udalova, Jo A. van Ginderachter, Stefanie N. Vogel, and Thomas A. Wynn. **Macrophage Activation and Polarization: Nomenclature and Experimental Guidelines**. en. *Immunity* 41:1 (July 2014), 14–20. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2014.06.008. URL: https://www.sciencedirect.com/science/article/pii/S1074761314002283 (visited on 11/21/2022) (see page 2).

[26] Qi Su, Sun Y. Kim, Funmi Adewale, Ye Zhou, Christina Aldler, Min Ni, Yi Wei, Michael E. Burczynski, Gurinder S. Atwal, Mark W. Sleeman, Andrew J. Murphy, Yurong Xin, and Xiping Cheng. **Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver**. en. *iScience* 24:11 (Nov. 2021), 103233. ISSN: 2589-0042. DOI: 10.1016/j.isci.2021.103233. URL: https://www.sciencedirect.com/science/article/pii/S2589004221012013 (visited on 02/04/2022) (see pages 2, 15, 71, 90).

[27] Xuelian Xiong, Henry Kuang, Sahar Ansari, Tongyu Liu, Jianke Gong, Shuai Wang, Xu-Yun Zhao, Yewei Ji, Chuan Li, Liang Guo, Linkang Zhou, Zhimin Chen, Paola Leon-Mimila, Meng Ting Chung, Katsuo Kurabayashi, Judy Opp, Francisco Campos-Pérez, Hugo Villamil-Ramírez, Samuel Canizales-Quinteros, Robert Lyons, Carey N. Lumeng, Beiyan Zhou, Ling Qi, Adriana Huertas-Vazquez, Aldons J. Lusis, X. Z. Shawn Xu, Siming Li, Yonghao Yu, Jun Z. Li, and Jiandie D. Lin. **Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-Cell Secretome Gene Analysis**. en. *Molecular Cell* 75:3 (Aug. 2019), 644–660.e5. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2019.07.028. URL:

https://www.sciencedirect.com/science/article/pii/S1097276519305830 (visited on 02/07/2022) (see pages 2, 15, 71, 104).

[28] Rance Nault, Kelly A. Fader, Sudin Bhattacharya, and Tim R. Zacharewski. **Single-Nuclei RNA Sequencing Assessment of the Hepatic Effects of 2,3,7,8-Tetrachlorodibenzo-p-dioxin**. en. *Cellular and Molecular Gastroenterology and Hepatology* 11:1 (Jan. 2021), 147–159. ISSN: 2352-345X. DOI: 10.1016/j.jcmgh.2020.07.012. URL: https://www.sciencedirect.com/science/article/pii/S2352345X20301181 (visited on 05/16/2022) (see pages 2, 17, 40, 84).

[29] Scott L. Friedman. **Hepatic Stellate Cells: Protean, Multifunctional, and Enigmatic Cells of the Liver**. *Physiological reviews* 88:1 (Jan. 2008), 125–172. ISSN: 0031-9333. DOI: 10.1152/physrev.00013.2007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2888531/ (visited on 05/17/2022) (see page 2).

[30] Gaoxiong Wang, Youshi Zheng, Yingchao Wang, Zhixiong Cai, Naishun Liao, Jingfeng Liu, and Wenmin Zhang. **Co-culture system of hepatocytes and endothelial cells: two in vitro approaches for enhancing liver-specific functions of hepatocytes**. *Cytotechnology* 70:4 (Aug. 2018), 1279–1290. ISSN: 0920-9069. DOI: 10.1007/s10616-018-0219-3. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6081931/ (visited on 12/02/2022) (see page 3).

[31] W. A. Bonner, H. R. Hulett, R. G. Sweet, and L. A. Herzenberg. **Fluorescence Activated Cell Sorting**. *Review of Scientific Instruments* 43:3 (Mar. 1972), 404–409. ISSN: 0034-6748. DOI: 10.1063/1.1685647. URL: https://aip.scitation.org/doi/abs/10.1063/1.1685647 (visited on 01/04/2019) (see pages 3, 108).

[32] Berend Snijder and Lucas Pelkmans. **Origins of regulated cell-to-cell variability**. en. *Nature Reviews Molecular Cell Biology* 12:2 (Feb. 2011). Number: 2 Publisher: Nature Publishing Group, 119–125. ISSN: 1471-0080. DOI: 10.1038/nrm3044. URL: https://www.nature.com/articles/nrm3044 (visited on 10/18/2022) (see page 3).

[33] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. **mRNA-Seq whole-transcriptome analysis of a single cell**. eng. *Nature Methods* 6:5 (May 2009), 377–382. ISSN: 1548-7105. DOI: 10.1038/nmeth.1315 (see pages 3, 4, 83, 103).

[34] Antonis Koussounadis, Simon P. Langdon, In Hwa Um, David J. Harrison, and V. Anne Smith. **Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system**. en. *Scientific Reports* 5:1 (June 2015). Number: 1 Publisher: Nature Publishing Group, 10775. ISSN: 2045-2322. DOI: 10.1038/srep10775. URL: https://www.nature.com/articles/srep10775 (visited on 04/21/2022) (see pages 4, 36).

[35] Raquel de Sousa Abreu, Luiz O. Penalva, Edward M. Marcotte, and Christine Vogel. **Global signatures of protein and mRNA expression levels**. *Molecular bioSystems* 5:12 (Dec. 2009), 1512–1526. ISSN: 1742-206X. DOI: 10.1039/b908315d. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4089977/ (visited on 04/21/2022) (see page 4).

[36] Christine Vogel and Edward M. Marcotte. **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses**. en. *Nature Reviews Genetics* 13:4 (Apr. 2012). Number: 4 Publisher: Nature Publishing Group, 227–232. ISSN: 1471-0064. DOI: 10.1038/nrg3185. URL: https://www.nature.com/articles/nrg3185 (visited on 04/21/2022) (see page 4).

[37] **Single-cell RNA counting at allele and isoform resolution using Smart-seq3 | Nature Biotechnology**. In: URL: https://www.nature.com/articles/s41587-020-0497-0 (visited on 10/18/2022) (see pages 4, 5, 25, 41, 84).

[38] Nicola J. Allen, Mariko L. Bennett, Lynette C. Foo, Gordon X. Wang, Chandrani Chakraborty, Stephen J. Smith, and Ben A. Barres. **Astrocyte glypicans 4 and 6 promote formation of excitatory synapses via GluA1 AMPA receptors**. eng. *Nature* 486:7403 (May 2012), 410–414. ISSN: 1476-4687. DOI: 10.1038/nature11059 (see page 4).

[39] Aviv Regev, Sarah Teichmann, Orit Rozenblatt-Rosen, Michael Stubbington, Kristin Ardlie, Ido Amit, Paola Arlotta, Gary Bader, Christophe Benoist, Moshe Biton, Bernd Bodenmiller, Benoit Bruneau, Peter Campbell, Mary Carmichael, Piero Carninci, Leslie Castelo-Soccio, Menna Clatworthy, Hans Clevers, Christian Conrad, Roland Eils, Jeremy Freeman, Lars Fugger, Berthold Goettgens, Daniel Graham, Anna Greka, Nir Hacohen, Muzlifah Haniffa, Ingo Helbig, Robert Heuckeroth, Sekar Kathiresan, Seung Kim, Allon Klein, Bartha Knoppers, Arnold Kriegstein, Eric Lander, Jane Lee, Ed Lein, Sten Linnarsson, Evan Macosko, Sonya MacParland, Robert Majovski, Partha Majumder, John Marioni, Ian McGilvray, Miriam Merad, Musa Mhlanga, Shalin Naik, Martijn Nawijn, Garry Nolan, Benedict Paten, Dana Pe'er, Anthony Philippakis, Chris Ponting, Steve Quake, Jayaraj Rajagopal, Nikolaus Rajewsky, Wolf Reik, Jennifer Rood, Kourosh Saeb-Parsy, Herbert Schiller, Steve Scott, Alex Shalek, Ehud Shapiro, Jay Shin, Kenneth Skeldon, Michael Stratton, Jenna Streicher, Henk Stunnenberg, Kai Tan, Deanne Taylor, Adrian Thorogood, Ludovic Vallier, Alexander van Oudenaarden, Fiona Watt, Wilko Weicher, Jonathan Weissman, Andrew Wells, Barbara Wold, Ramnik Xavier, Xiaowei Zhuang, and Human Cell Atlas Organizing Committee. **The Human Cell Atlas White Paper**. *arXiv:1810.05192 [q-bio]* (Oct. 2018). arXiv: 1810.05192. URL: http://arxiv.org/abs/1810.05192 (visited on 05/06/2022) (see pages 4, 6, 83, 113).

[40] Monika Litviňuková, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L. Worth, Eric L. Lindberg, Masatoshi Kanda, Krzysztof Polanski, Matthias Heinig, Michael Lee, Emily R. Nadelmann, Kenny Roberts, Liz Tuck, Eirini S. Fasouli, Daniel M. DeLaughter, Barbara McDonough, Hiroko Wakimoto, Joshua M. Gorham, Sara Samari, Krishnaa T. Mahbubani, Kourosh Saeb-Parsy, Giannino Patone, Joseph J. Boyle, Hongbo Zhang, Hao Zhang, Anissa Viveiros, Gavin Y. Oudit, Omer Ali Bayraktar, J. G. Seidman, Christine E. Seidman, Michela Noseda, Norbert Hubner, and Sarah A. Teichmann. **Cells of the adult human heart**. en. *Nature* 588:7838 (Dec. 2020). Number: 7838 Publisher: Nature Publishing Group, 466–472. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2797-4. URL: https://www.nature.com/articles/s41586-020-2797-4 (visited on 02/11/2022) (see pages 4, 6, 83).

[41] Herbert B. Schiller, Daniel T. Montoro, Lukas M. Simon, Emma L. Rawlins, Kerstin B. Meyer, Maximilian Strunz, Felipe A. Vieira Braga, Wim Timens, Gerard H. Koppelman, G. R. Scott Budinger, Janette K. Burgess, Avinash Waghray, Maarten van den Berge, Fabian J. Theis, Aviv Regev, Naftali Kaminski, Jayaraj Rajagopal, Sarah A. Teichmann, Alexander V. Misharin, and Martijn C. Nawijn. **The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease**. *American Journal of Respiratory Cell and Molecular Biology* 61:1 (July 2019), 31–41. ISSN: 1044-1549. DOI: 10.1165/rcmb.2018-0416TR. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6604220/ (visited on 11/21/2022) (see page 4).

[42] L. Sikkema, D. Strobl, L. Zappia, E. Madissoon, N. S. Markov, L. Zaragosi, M. Ansari, M. Arguel, L. Apperloo, C. Bécavin, M. Berg, E. Chichelnitskiy, M. Chung, A. Collin, A. C. A. Gay, B. Hooshiar Kashani, M. Jain, T. Kapellos, T. M. Kole, C. Mayr, M. von Papen, L. Peter, C. Ramírez-Suástegui, J. Schniering, C. Taylor, T. Walzthoeni, C. Xu, L. T. Bui, C. de Donno, L. Dony, M. Guo, A. J. Gutierrez, L. Heumos, N. Huang, I. Ibarra, N. Jackson, P. Kadur Lakshminarasimha Murthy, M. Lotfollahi, T. Tabib, C. Talavera-Lopez, K. Travaglini, A. Wilbrey-Clark, K. B. Worlock, M. Yoshida, Lung Biological Network Consortium, T. Desai, O. Eickelberg, C. Falk, N. Kaminski, M. Krasnow, R. Lafyatis, M. Nikolíc, J. Powell, J. Rajagopal, O. Rozenblatt-Rosen, M. A. Seibold, D. Sheppard, D. Shepherd, S. A. Teichmann, A. Tsankov, J. Whitsett, Y. Xu, N. E. Banovich, P. Barbry, T. E. Duong, K. B. Meyer, J. A. Kropski, D. Pe'er, H. B. Schiller, P. R. Tata, J. L. Schultze, A. V. Misharin, M. C. Nawijn, M. D. Luecken, and F. Theis. **An integrated cell atlas of the human lung in health and disease**. en. In: Pages: 2022.03.10.483747 Section: New Results. bioRxiv, Mar. 2022. DOI: 10.1101/2022.03.10.483747. URL: https://www.biorxiv.org/content/10.1101/2022.03.10.483747v1 (visited on 11/21/2022) (see page 4).

[43] THE TABULA SAPIENS CONSORTIUM. **The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans**. *Science* 376:6594 (May 2022). Publisher: American Association for the Advancement of Science, eabl4896. DOI: 10.1126/science.abl4896. URL: https://www.science.org/doi/10.1126/science.abl4896 (visited on 05/16/2022) (see pages 4, 5).

[44] **Current best practices in single-cell RNA-seq analysis: a tutorial**. *Molecular Systems Biology* 15:6 (June 2019). Publisher: John Wiley & Sons, Ltd, e8746. ISSN: 1744-4292. DOI: 10.15252/msb.20188746. URL: https://www.embopress.org/doi/full/10.15252/msb.20188746 (visited on 02/04/2022) (see pages 4, 8, 109).

[45] Leonard Kaufman and Peter J. Rousseeuw. **Finding Groups in Data: An Introduction to Cluster Analysis**. en. Google-Books-ID: YeFQHiikNo0C. John Wiley & Sons, Sept. 2009. ISBN: 978-0-470-31748-8 (see pages 5, 89).

[46] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. **NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set**. *Journal of Statistical Software* 61 (Oct. 2014), 1–36. DOI: 10.18637/jss.v061.i06 (see page 5).

[47] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. **powsimR: power analysis for bulk and single cell RNA-seq experiments**. *Bioinformatics* 33:21 (Nov. 2017), 3486–3488. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx435. URL: https://doi.org/10.1093/bioinformatics/btx435 (visited on 05/20/2022) (see pages 5, 34, 69).

[48] Katharina T. Schmid, Barbara Höllbacher, Cristiana Cruceanu, Anika Böttcher, Heiko Lickert, Elisabeth B. Binder, Fabian J. Theis, and Matthias Heinig. **scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies**. en. *Nature Communications* 12:1 (Nov. 2021). Number: 1 Publisher: Nature Publishing Group, 6625. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26779-7. URL: https://www.nature.com/articles/s41467-021-26779-7 (visited on 05/20/2022) (see pages 5, 34, 69, 88, 110).

[49] Hongyu Guo and Jun Li. **scSorter: assigning cells to known cell types according to marker genes**. *Genome Biology* 22:1 (Feb. 2021), 69. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02281-7. URL: https://doi.org/10.1186/s13059-021-02281-7 (visited on 05/27/2022) (see page 5).

[50] Oscar Franzén, Li-Ming Gan, and Johan L M Björkegren. **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data**. *Database* 2019 (Jan. 2019), baz046. ISSN: 1758-0463. DOI: 10.1093/database/baz046. URL: https://doi.org/10.1093/database/baz046 (visited on 05/27/2022) (see pages 5, 42, 116).

[51] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V. Misharin, and Fabian J. Theis. **Mapping single-cell data to reference atlases by transfer learning**. en. *Nature Biotechnology* 40:1 (Jan. 2022). Number: 1 Publisher: Nature Publishing Group, 121–130. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01001-7. URL: https://www.nature.com/articles/s41587-021-01001-7 (visited on 11/21/2022) (see page 5).

[52] Simone Picelli, Åsa K. Björklund, Omid R. Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. **Smart-seq2 for sensitive full-length transcriptome profiling in single cells**. en. *Nature Methods* 10:11 (Nov. 2013). Number: 11 Publisher: Nature Publishing Group, 1096–1098. ISSN: 1548-7105. DOI: 10.1038/nmeth.2639. URL: https://www.nature.com/articles/nmeth.2639 (visited on 10/18/2022) (see page 6).

[53] Simone Picelli, Omid R. Faridani, Åsa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. **Full-length RNA-seq from single cells using Smart-seq2**. en. *Nature Protocols* 9:1 (Jan. 2014). Number: 1 Publisher: Nature Publishing Group, 171–181. ISSN: 1750-2799. DOI: 10.1038/nprot.2014.006. URL: https://www.nature.com/articles/nprot.2014.006 (visited on 09/06/2022) (see pages 6, 108).

[54] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. **Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types**. *Science (New York, N.Y.)* 343:6172 (Feb. 2014), 776–779. ISSN: 0036-8075. DOI: 10.1126/science.1247651. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4412462/ (visited on 05/20/2022) (see pages 6, 36, 103).

[55] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets**. eng. *Cell* 161:5 (May 2015), 1202–1214. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.05.002 (see pages 6, 108).

[56] Xiliang Wang, Yao He, Qiming Zhang, Xianwen Ren, and Zemin Zhang. **Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2**. en. *Genomics, Proteomics & Bioinformatics*. Single-cell Omics Analysis 19:2 (Apr. 2021), 253–266. ISSN: 1672-0229. DOI: 10.1016/j.gpb.2020.02.005. URL: https://www.sciencedirect.com/science/article/pii/S1672022921000486 (visited on 07/27/2022) (see pages 6, 24, 25, 40, 83–85, 87).

[57] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. **Comparative Analysis of Single-Cell RNA Sequencing Methods**. en. *Molecular Cell* 65:4 (Feb. 2017), 631–643.e4. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2017.01.023. URL: https://www.sciencedirect.com/science/article/pii/S1097276517300497 (visited on 07/27/2022) (see pages 6–8, 25, 40, 44, 83, 85).

[58] Peter See, Josephine Lum, Jinmiao Chen, and Florent Ginhoux. **A Single-Cell Sequencing Guide for Immunologists**. *Frontiers in Immunology* 9 (2018). ISSN: 1664-3224. URL: https://www.frontiersin.org/articles/10.3389/fimmu.2018.02425 (visited on 10/18/2022) (see page 6).

[59] Charles J. Epstein. **Cell Size, Nuclear Content, and the Development of Polyploidy in the Mammalian Liver**. en. *Proceedings of the National Academy of Sciences* 57:2 (Feb. 1967). Publisher: National Academy of Sciences Section: Biological Sciences: Zoology, 327–334. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.57.2.327. URL: https://www.pnas.org/content/57/2/327 (visited on 02/04/2022) (see page 7).

[60] B. N. Kudryavtsev, M. V. Kudryavtseva, G. A. Sakuta, and G. I. Stein. **Human hepatocyte polyploidization kinetics in the course of life cycle**. en. *Virchows Archiv B* 64:1 (Dec. 1993), 387. ISSN: 0340-6075. DOI: 10.1007/BF02915139. URL: https://doi.org/10.1007/BF02915139 (visited on 02/09/2022) (see pages 7, 90).

[61] Min-Jun Wang, Fei Chen, Joseph T. Y. Lau, and Yi-Ping Hu. **Hepatocyte polyploidization and its association with pathophysiological processes**. en. *Cell Death & Disease* 8:5 (May 2017). Number: 5 Publisher: Nature Publishing Group, e2805–e2805. ISSN: 2041-4889. DOI: 10.1038/cddis.2017.167. URL: https://www.nature.com/articles/cddis2017167 (visited on 02/09/2022) (see pages 7, 90).

[62] M. Winkelmann, P. Pfitzer, and W. Schneider. **Significance of polyploidy in megakaryocytes and other cells in health and tumor disease**. en. *Klinische Wochenschrift* 65:23 (Dec. 1987), 1115–1131. ISSN: 1432-1440. DOI: 10.1007/BF01734832. URL: https://doi.org/10.1007/BF01734832 (visited on 11/22/2022) (see page 7).

[63] Hyun O. Lee, Jean M. Davidson, and Robert J. Duronio. **Endoreplication: polyploidy with purpose**. en. *Genes & Development* 23:21 (Nov. 2009). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 2461–2477. ISSN:

0890-9369, 1549-5477. DOI: 10.1101/gad.1829209. URL: http://genesdev.cshlp.org/content/23/21/2461 (visited on 05/27/2022) (see page 7).

[64] Séverine Celton-Morizur, Grégory Merlen, Dominique Couton, and Chantal Desdouets. **Polyploidy and liver proliferation: Central role of insulin signaling**. en. *Cell Cycle* 9:3 (Feb. 2010), 460–466. ISSN: 1538-4101, 1551-4005. DOI: 10.4161/cc.9.3.10542. URL: http://www.tandfonline.com/doi/abs/10.4161/cc.9.3.10542 (visited on 05/27/2022) (see page 7).

[65] Shusil K. Pandit, Bart Westendorp, and Alain de Bruin. **Physiological significance of polyploidization in mammalian cells**. en. *Trends in Cell Biology* 23:11 (Nov. 2013), 556–566. ISSN: 0962-8924. DOI: 10.1016/j.tcb.2013.06.002. URL: https://www.sciencedirect.com/science/article/pii/S0962892413000962 (visited on 05/27/2022) (see pages 7, 46).

[66] Kurt M. Schmoller and Jan M. Skotheim. **The Biosynthetic Basis of Cell Size Control**. en. *Trends in Cell Biology*. Special Issue: Quantitative Cell Biology 25:12 (Dec. 2015), 793–802. ISSN: 0962-8924. DOI: 10.1016/j.tcb.2015.10.006. URL: https://www.sciencedirect.com/science/article/pii/S0962892415001932 (visited on 02/09/2022) (see pages 7, 26, 46, 91, 111).

[67] Olivia Padovan-Merhar, Gautham P. Nair, Andrew G. Biaesch, Andreas Mayer, Steven Scarfone, Shawn W. Foley, Angela R. Wu, L. Stirling Churchman, Abhyudai Singh, and Arjun Raj. **Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms**. en. *Molecular Cell* 58:2 (Apr. 2015), 339–352. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.03.005. URL: https://www.sciencedirect.com/science/article/pii/S1097276515001707 (visited on 02/09/2022) (see pages 7, 11, 26, 29, 46, 50, 91).

[68] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. **Missing data and technical variability in single-cell RNA-sequencing experiments**. *Biostatistics* 19:4 (Oct. 2018), 562–578. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxx053. URL: https://doi.org/10.1093/biostatistics/kxx053 (visited on 11/22/2022) (see page 8).

[69] Christoph Hafemeister and Rahul Satija. **Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression**. *Genome Biology* 20:1 (Dec. 2019), 296. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1874-1. URL: https://doi.org/10.1186/s13059-019-1874-1 (visited on 04/19/2021) (see pages 8, 9).

[70] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. **Pooling across cells to normalize single-cell RNA sequencing data with many zero counts**. *Genome Biology* 17:1 (Apr. 2016), 75. ISSN: 1474-760X. DOI: 10.1186/s13059-016-

0947-7. URL: https://doi.org/10.1186/s13059-016-0947-7 (visited on 04/19/2021) (see pages 8, 10, 86, 103, 113).

[71] Aaron T. L. Lun and John C. Marioni. **Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data**. *Biostatistics* 18:3 (July 2017), 451–464. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxw055. URL: https://doi.org/10.1093/biostatistics/kxw055 (visited on 02/04/2022) (see pages 8, 30).

[72] Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. **Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions**. *Briefings in Bioinformatics* 19:5 (Feb. 2017), 776–792. ISSN: 1467-5463. DOI: 10.1093/bib/bbx008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6171491/ (visited on 05/25/2022) (see pages 9–11, 26, 86).

[73] Michael I. Love, Wolfgang Huber, and Simon Anders. **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biology* 15:12 (Dec. 2014), 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8 (visited on 05/25/2022) (see pages 9, 86).

[74] Simon Anders and Wolfgang Huber. **Differential expression analysis for sequence count data**. *Genome Biology* 11:10 (Oct. 2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106. URL: https://doi.org/10.1186/gb-2010-11-10-r106 (visited on 05/27/2022) (see page 9).

[75] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. **A general and flexible method for signal extraction from single-cell RNA-seq data**. en. *Nature Communications* 9:1 (Jan. 2018). Number: 1 Publisher: Nature Publishing Group, 284. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02554-5. URL: https://www.nature.com/articles/s41467-017-02554-5 (visited on 05/30/2022) (see page 9).

[76] Belinda Phipson, Luke Zappia, and Alicia Oshlack. **Gene length and detection bias in single cell RNA sequencing protocols**. *F1000Research* 6 (Apr. 2017), 595. ISSN: 2046-1402. DOI: 10.12688/f1000research.11290.1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5428526/ (visited on 09/29/2022) (see pages 9, 85).

[77] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. **Spatial reconstruction of single-cell gene expression data**. en. *Nature Biotechnology* 33:5 (May 2015). Number: 5 Publisher: Nature Publishing Group, 495–502. ISSN: 1546-1696. DOI: 10.1038/nbt.3192. URL: https://www.nature.com/articles/nbt.3192 (visited on 05/30/2022) (see pages 10, 14, 37, 38, 53, 86, 89).

[78] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. **SCANPY: large-scale single-cell gene expression data analysis**. *Genome Biology* 19:1 (Feb. 2018), 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: https://doi.org/10.1186/s13059-017-1382-0 (visited on 01/09/2019) (see pages 10, 38, 86, 89).

[79] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. **Integrating single-cell transcriptomic data across different conditions, technologies, and species**. eng. *Nature Biotechnology* 36:5 (2018), 411–420. ISSN: 1546-1696. DOI: 10.1038/nbt.4096 (see page 10).

[80] Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P Conley, Rosalie Elespuru, Michael Fero, Carole Foy, James Fuscoe, Xiaolian Gao, David Lee Gerhold, Patrick Gilles, Federico Goodsaid, Xu Guo, Joe Hackett, Richard D Hockett, Pranvera Ikonomi, Rafael A Irizarry, Ernest S Kawasaki, Tamma Kaysser-Kranich, Kathleen Kerr, Gretchen Kiser, Walter H Koch, Kathy Y Lee, Chunmei Liu, Z Lewis Liu, Anne Lucas, Chitra F Manohar, Garry Miyada, Zora Modrusan, Helen Parkes, Raj K Puri, Laura Reid, Thomas B Ryder, Marc Salit, Raymond R Samaha, Uwe Scherf, Timothy J Sendera, Robert A Setterquist, Leming Shi, Richard Shippy, Jesus V Soriano, Elizabeth A Wagar, Janet A Warrington, Mickey Williams, Frederike Wilmer, Mike Wilson, Paul K Wolber, Xiaoning Wu, Renata Zadro, and The External RNA Controls Consortium. **The External RNA Controls Consortium: a progress report**. en. *Nature Methods* 2:10 (Oct. 2005). Number: 10 Publisher: Nature Publishing Group, 731–734. ISSN: 1548-7105. DOI: 10.1038/nmeth1005-731. URL: https://www.nature.com/articles/nmeth1005-731 (visited on 05/27/2022) (see pages 10, 108).

[81] Hangnoh Lee, P. Scott Pine, Jennifer McDaniel, Marc Salit, and Brian Oliver. **External RNA Controls Consortium Beta Version Update**. en. *Journal of Genomics* 4 (2016), 19–22. ISSN: 1839-9940. DOI: 10.7150/jgen.16082. URL: http://www.jgenomics.com/v04p0019.htm (visited on 02/09/2022) (see pages 10, 22, 26, 85).

[82] Aaron T.L. Lun, Fernando J. Calero-Nieto, Liora Haim-Vilmovsky, Berthold Göttgens, and John C. Marioni. **Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data**. *Genome Research* 27:11 (Nov. 2017), 1795–1806. ISSN: 1088-9051. DOI: 10.1101/gr.222877.117. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5668938/ (visited on 12/06/2022) (see pages 10, 22, 26).

[83] Davide Risso, John Ngai, Terence P. Speed, and Sandrine Dudoit. **Normalization of RNA-seq data using factor analysis of control genes or samples**. *Nature biotechnology* 32:9 (Sept. 2014), 896–902. ISSN: 1087-0156. DOI: 10.1038/nbt.2931.

URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4404308/ (visited on 05/30/2022) (see pages 10, 26).

[84] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J. Miragaia, Charlotte Labalette, Iain C. Macaulay, Ana Cvejic, and Sarah A. Teichmann. **Power analysis of single-cell RNA-sequencing experiments**. en. *Nature Methods* 14:4 (Apr. 2017). Number: 4 Publisher: Nature Publishing Group, 381–387. ISSN: 1548-7105. DOI: 10.1038/nmeth.4220. URL: https://www.nature.com/articles/nmeth.4220 (visited on 09/28/2022) (see pages 10, 11, 26, 85).

[85] Sivan Tanami, Shani Ben-Moshe, Anat Elkayam, Avi Mayo, Keren Bahar Halpern, and Shalev Itzkovitz. **Dynamic zonation of liver polyploidy**. en. *Cell and Tissue Research* 368:2 (May 2017), 405–410. ISSN: 1432-0878. DOI: 10.1007/s00441-016-2427-5. URL: https://doi.org/10.1007/s00441-016-2427-5 (visited on 02/09/2022) (see pages 11, 53, 94, 103).

[86] Cyril Torre, Christine Perret, and Sabine Colnot. "Chapter 5 - Molecular Determinants of Liver Zonation." en. In: *Progress in Molecular Biology and Translational Science*. Ed. by Klaus H. Kaestner. Vol. 97. Development, Differentiation and Disease of the Para-Alimentary Tract. Academic Press, Jan. 2010, 127–150. DOI: 10.1016/B978-0-12-385233-5.00005-2. URL: https://www.sciencedirect.com/science/article/pii/B9780123852335000052 (visited on 05/30/2022) (see page 12).

[87] Dieter Sasse, Norbert Katz, and Kurt Jungermann. **Functional heterogeneity of rat liver parenchyma and of isolated hepatocytes**. en. *FEBS Letters* 57:1 (1975). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/0014-5793%2875%2980157-8, 83–88. ISSN: 1873-3468. DOI: 10.1016/0014-5793(75)80157-8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1016/0014-5793%2875%2980157-8 (visited on 05/30/2022) (see page 12).

[88] Albert Braeuning, Carina Ittrich, Christoph Köhle, Stephan Hailfinger, Michael Bonin, Albrecht Buchmann, and Michael Schwarz. **Differential gene expression in periportal and perivenous mouse hepatocytes**. en. *The FEBS Journal* 273:22 (2006). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1742-4658.2006.05503.x, 5051–5061. ISSN: 1742-4658. DOI: 10.1111/j.1742-4658.2006.05503.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-4658.2006.05503.x (visited on 05/31/2022) (see pages 13, 17, 94).

[89] Nikolaus Berndt, Erik Kolbe, Robert Gajowski, Johannes Eckstein, Fritzi Ott, David Meierhofer, Hermann-Georg Holzhütter, and Madlen Matz-Soja. **Functional Consequences of Metabolic Zonation in Murine Livers: Insights for an Old Story**. en. *Hepatology* 73:2 (2021), 795–810. ISSN: 1527-3350. DOI: 10.1002/hep.31274. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hep.31274 (visited on 05/31/2022) (see pages 13, 17, 93, 94).

[90] Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. **Diffusion pseudotime robustly reconstructs lineage branching**. eng. *Nature Methods* 13:10 (2016), 845–848. ISSN: 1548-7105. DOI: 10.1038/nmeth.3971 (see pages 14, 53, 93, 96, 124).

[91] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. **A comparison of single-cell trajectory inference methods**. en. *Nature Biotechnology* 37:5 (May 2019). Number: 5 Publisher: Nature Publishing Group, 547–554. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0071-9. URL: https://www.nature.com/articles/s41587-019-0071-9 (visited on 02/09/2022) (see pages 14, 93).

[92] Mikolaj Ogrodnik, Satomi Miwa, Tamar Tchkonia, Dina Tiniakos, Caroline L. Wilson, Albert Lahat, Christoper P. Day, Alastair Burt, Allyson Palmer, Quentin M. Anstee, Sushma Nagaraja Grellscheid, Jan H. J. Hoeijmakers, Sander Barnhoorn, Derek A. Mann, Thomas G. Bird, Wilbert P. Vermeij, James L. Kirkland, João F. Passos, Thomas von Zglinicki, and Diana Jurk. **Cellular senescence drives age-dependent hepatic steatosis**. en. *Nature Communications* 8:1 (June 2017). Number: 1 Publisher: Nature Publishing Group, 15691. ISSN: 2041-1723. DOI: 10.1038/ncomms15691. URL: https://www.nature.com/articles/ncomms15691 (visited on 02/11/2022) (see pages 15, 90).

[93] Timothy Hardy, Fiona Oakley, Quentin M. Anstee, and Christopher P. Day. **Nonalcoholic Fatty Liver Disease: Pathogenesis and Disease Spectrum**. *Annual Review of Pathology: Mechanisms of Disease* 11:1 (2016), 451–496. DOI: 10.1146/annurev-pathol-012615-044224. URL: https://doi.org/10.1146/annurev-pathol-012615-044224 (visited on 10/18/2022) (see page 15).

[94] Pierre-Antoine Soret, Julie Magusto, Chantal Housset, and Jérémie Gautheron. **In Vitro and In Vivo Models of Non-Alcoholic Fatty Liver Disease: A Critical Appraisal**. *Journal of Clinical Medicine* 10:1 (Dec. 2020), 36. ISSN: 2077-0383. DOI: 10.3390/jcm10010036. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7794936/ (visited on 10/18/2022) (see pages 15, 16, 18).

[95] Jos V Castell, Ramiro Jover, Celia P Martnez-Jimnez, and Mara Jos Gmez-Lechn. **Hepatocyte cell lines: their use, scope and limitations in drug metabolism studies**. *Expert Opinion on Drug Metabolism & Toxicology* 2:2 (Apr. 2006). Publisher: Taylor & Francis _eprint: https://doi.org/10.1517/17425255.2.2.183, 183–212. ISSN: 1742-5255. DOI: 10.1517/17425255.2.2.183. URL: https://doi.org/10.1517/17425255.2.2.183 (visited on 02/03/2022) (see pages 15, 17–19, 62, 64, 74, 98, 100, 101).

[96] María José Gómez-Lechón, María Teresa Donato, Alicia Martínez-Romero, Nuria Jiménez, José Vicente Castell, and José-Enrique O'Connor. **A human hepatocellular in vitro model to investigate steatosis**. en. *Chemico-Biological Interactions* 165:2 (Jan. 2007), 106–116. ISSN: 0009-2797. DOI: 10.1016/j.cbi.2006.11.004.

URL: https://www.sciencedirect.com/science/article/pii/S000927970600336X (visited on 12/06/2022) (see pages 16, 18, 71).

[97] Valerie Y. Soldatow, Edward L. LeCluyse, Linda G. Griffith, and Ivan Rusyn. **In vitro models for liver toxicity testing**. *Toxicology research* 2:1 (Jan. 2013), 23–39. ISSN: 2045-452X. DOI: 10.1039/C2TX20051A. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3593300/ (visited on 02/11/2022) (see pages 16, 17, 74, 95).

[98] María José Gómez-Lechón, José V. Castell, and María Teresa Donato. "The Use of Hepatocytes to Investigate Drug Toxicity." en. In: *Hepatocytes: Methods and Protocols*. Ed. by Patrick Maurel. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2010, 389–415. ISBN: 978-1-60761-688-7. DOI: 10.1007/978-1-60761-688-7_21. URL: https://doi.org/10.1007/978-1-60761-688-7_21 (visited on 02/11/2022) (see pages 16, 18, 19).

[99] Celia P. Martínez-Jiménez, M. José Gómez-Lechón, José V. Castell, and Ramiro Jover. **Transcriptional Regulation of the Human Hepatic CYP3A4: Identification of a New Distal Enhancer Region Responsive to CCAAT/Enhancer-Binding Protein 03b2 Isoforms (Liver Activating Protein and Liver Inhibitory Protein)**. en. *Molecular Pharmacology* 67:6 (June 2005). Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Article, 2088–2101. ISSN: 0026-895X, 1521-0111. DOI: 10.1124/mol.104.008169. URL: https://molpharm.aspetjournals.org/content/67/6/2088 (visited on 02/04/2022) (see page 16).

[100] Celia P. Martínez-Jiménez, José V. Castell, M. José Gómez-Lechón, and Ramiro Jover. **Transcriptional Activation of CYP2C9, CYP1A1, and CYP1A2 by Hepatocyte Nuclear Factor 403b1 Requires Coactivators Peroxisomal Proliferator Activated Receptor-03b3 Coactivator 103b1 and Steroid Receptor Coactivator 1**. en. *Molecular Pharmacology* 70:5 (Nov. 2006). Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Article, 1681–1692. ISSN: 0026-895X, 1521-0111. DOI: 10.1124/mol.106.025403. URL: https://molpharm.aspetjournals.org/content/70/5/1681 (visited on 02/03/2022) (see pages 16, 44).

[101] Celia P. Martínez-Jiménez, M. José Gómez-Lechón, José V. Castell, and Ramiro Jover. **Underexpressed Coactivators PGC103b1 AND SRC1 Impair Hepatocyte Nuclear Factor 403b1 Function and Promote Dedifferentiation in Human Hepatoma Cells ***. English. *Journal of Biological Chemistry* 281:40 (Oct. 2006). Publisher: Elsevier, 29840–29849. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M604046200. URL: https://www.jbc.org/article/S0021-9258(19)33891-8/abstract (visited on 02/03/2022) (see page 16).

[102] Yoko Yagishita, Akira Uruno, and Masayuki Yamamoto. "Chapter 27 - NRF2-Mediated Gene Regulation and Glucose Homeostasis." en. In: *Molecular Nutrition and Diabetes*. Ed. by Didac Mauricio. San Diego: Academic Press, Jan. 2016, 331–348. ISBN: 978-0-12-801585-8. DOI: 10.1016/B978-0-12-801585-8.00027-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128015858000270 (visited on 05/13/2022) (see pages 16, 98).

[103] David F. V. Lewis. **57 varieties: the human cytochromes P450**. eng. *Pharmacogenomics* 5:3 (Apr. 2004), 305–318. ISSN: 1462-2416. DOI: 10.1517/phgs.5.3.305.29827 (see page 16).

[104] Leroy Hood and Lee Rowen. **The Human Genome Project: big science transforms biology and medicine**. *Genome Medicine* 5:9 (Sept. 2013), 79. ISSN: 1756-994X. DOI: 10.1186/gm483. URL: https://doi.org/10.1186/gm483 (visited on 05/13/2022) (see page 16).

[105] Ulrich M. Zanger and Matthias Schwab. **Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation**. en. *Pharmacology & Therapeutics* 138:1 (Apr. 2013), 103–141. ISSN: 0163-7258. DOI: 10.1016/j.pharmthera.2012.12.007. URL: https://www.sciencedirect.com/science/article/pii/S0163725813000065 (visited on 02/11/2022) (see pages 16, 17, 99).

[106] Petra Jancova, Pavel Anzenbacher, and Eva Anzenbacherova. **Phase II drug metabolizing enzymes**. eng. *Biomedical Papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia* 154:2 (June 2010), 103–116. ISSN: 1213-8118. DOI: 10.5507/bp.2010.017 (see pages 16, 98).

[107] Curtis J. Omiecinski, John P. Vanden Heuvel, Gary H. Perdew, and Jeffrey M. Peters. **Xenobiotic metabolism, disposition, and regulation by receptors: from biochemical phenomenon to predictors of major toxicities**. eng. *Toxicological Sciences: An Official Journal of the Society of Toxicology* 120 Suppl 1 (Mar. 2011), S49–75. ISSN: 1096-0929. DOI: 10.1093/toxsci/kfq338 (see pages 16, 98).

[108] Maurice Arnaud. **Metabolism of caffeine and other components of coffee**. *Caffeine, Coffee and Health* (Jan. 1993), 43–95 (see page 16).

[109] Jan Grzegorzewski, Florian Bartsch, Adrian Köller, and Matthias König. **Pharmacokinetics of Caffeine: A Systematic Analysis of Reported Data for Application in Metabolic Phenotyping and Liver Function Testing**. *Frontiers in Pharmacology* 12 (2022). ISSN: 1663-9812. URL: https://www.frontiersin.org/article/10.3389/fphar.2021.752826 (visited on 05/17/2022) (see page 16).

[110] Sandrine Turpault, William Brian, Robert Van Horn, Alix Santoni, Franck Poitiers, Yves Donazzolo, and Xavier Boulenc. **Pharmacokinetic assessment of a five-probe cocktail for CYPs 1A2, 2C9, 2C19, 2D6 and 3A**. eng. *British Journal of Clinical Pharmacology* 68:6 (Dec. 2009), 928–935. ISSN: 1365-2125. DOI: 10.1111/j.1365-2125.2009.03548.x (see pages 17, 74, 98, 99).

[111] Jukka Hakkola, Janne Hukkanen, Miia Turpeinen, and Olavi Pelkonen. **Inhibition and induction of CYP enzymes in humans: an update**. *Archives of Toxicology* 94:11 (2020), 3671–3722. ISSN: 0340-5761. DOI: 10.1007/s00204-020-02936-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7603454/ (visited on 10/12/2022) (see pages 17, 100).

[112] Benjamin Berger, Massimiliano Donzelli, Swarna Maseneni, Franziska Boess, Adrian Roth, Stephan Krähenbühl, and Manuel Haschke. **Comparison Of Liver Cell Models Using The Basel Phenotyping Cocktail**. *Frontiers in Pharmacology* 7 (2016). ISSN: 1663-9812. URL: https://www.frontiersin.org/article/10.3389/fphar.2016.00443 (visited on 02/11/2022) (see pages 17, 74, 98, 99).

[113] M Bosilkovska, C F Samer, J Déglon, M Rebsamen, C Staub, P Dayer, B Walder, J A Desmeules, and Y Daali. **Geneva Cocktail for Cytochrome P450 and P-Glycoprotein Activity Assessment Using Dried Blood Spots**. *Clinical Pharmacology and Therapeutics* 96:3 (Sept. 2014), 349–359. ISSN: 0009-9236. DOI: 10.1038/clpt.2014.83. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151019/ (visited on 02/11/2022) (see pages 17, 98, 99).

[114] Chenxia Hu and Lanjuan Li. **In vitro culture of isolated primary hepatocytes and stem cell-derived hepatocyte-like cells for liver regeneration**. en. *Protein & Cell* 6:8 (Aug. 2015), 562–574. ISSN: 1674-8018. DOI: 10.1007/s13238-015-0180-2. URL: https://doi.org/10.1007/s13238-015-0180-2 (visited on 02/11/2022) (see pages 19, 34, 62, 95).

[115] James A. Heslop, Cliff Rowe, Joanne Walsh, Rowena Sison-Young, Roz Jenkins, Laleh Kamalian, Richard Kia, David Hay, Robert P. Jones, Hassan Z. Malik, Stephen Fenwick, Amy E. Chadwick, John Mills, Neil R. Kitteringham, Chris E. P. Goldring, and B. Kevin Park. **Mechanistic evaluation of primary human hepatocyte culture using global proteomic analysis reveals a selective dedifferentiation profile**. en. *Archives of Toxicology* 91:1 (Jan. 2017), 439–452. ISSN: 1432-0738. DOI: 10.1007/s00204-016-1694-y. URL: https://doi.org/10.1007/s00204-016-1694-y (visited on 02/11/2022) (see pages 19, 34, 62, 95, 98, 118).

[116] Benedikt Scheidecker, Marie Shinohara, Masahiro Sugimoto, Mathieu Danoy, Masaki Nishikawa, and Yasuyuki Sakai. **Induction of in vitro Metabolic Zonation in Primary Hepatocytes Requires Both Near-Physiological Oxygen Concentration and Flux**. *Frontiers in Bioengineering and Biotechnol-*

ogy 8 (2020). ISSN: 2296-4185. URL: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00524 (visited on 10/11/2022) (see pages 19, 95).

[117] M. L. Richter, I. K. Deligiannis, K. Yin, A. Danese, E. Lleshi, P. Coupland, C. A. Vallejos, K. P. Matchett, N. C. Henderson, M. Colome-Tatche, and C. P. Martinez-Jimenez. **Single-nucleus RNA-seq2 reveals functional crosstalk between liver zonation and ploidy**. en. *Nature Communications* 12:1 (July 2021). Number: 1 Publisher: Nature Publishing Group, 4264. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24543-5. URL: https://www.nature.com/articles/s41467-021-24543-5 (visited on 02/04/2022) (see pages 21, 22, 39–41, 46, 48, 50, 52, 54, 56, 58, 59, 85).

[118] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. **STAR: ultrafast universal RNA-seq aligner**. eng. *Bioinformatics (Oxford, England)* 29:1 (Jan. 2013), 15–21. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts635 (see pages 22, 107).

[119] Giacomo Baruzzo, Katharina E. Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A. FitzGerald, and Gregory R. Grant. **Simulation-based comprehensive benchmarking of RNA-seq aligners**. en. *Nature Methods* 14:2 (Feb. 2017). Number: 2 Publisher: Nature Publishing Group, 135–139. ISSN: 1548-7105. DOI: 10.1038/nmeth.4106. URL: https://www.nature.com/articles/nmeth.4106 (visited on 07/25/2022) (see page 22).

[120] Yang Liao, Gordon K Smyth, and Wei Shi. **The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads**. *Nucleic Acids Research* 47:8 (May 2019), e47. ISSN: 0305-1048. DOI: 10.1093/nar/gkz114. URL: https://doi.org/10.1093/nar/gkz114 (visited on 09/28/2022) (see page 23).

[121] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. **Massively parallel digital transcriptional profiling of single cells**. eng. *Nature Communications* 8 (2017), 14049. ISSN: 2041-1723. DOI: 10.1038/ncomms14049 (see pages 23, 87).

[122] E. Sanchez-Quant, M. L. Richter, M. Colomé-Tatché, and C. P. Martinez-Jimenez. **Single-cell metabolic profiling reveals subgroups of primary human hepatocytes showing heterogeneous responses to drug challenge**. en. In: Pages: 2022.06.08.495252 Section: New Results. bioRxiv, June 2022. DOI:

10.1101/2022.06.08.495252. URL: https://www.biorxiv.org/content/10.1101/2022.06.08.495252v1 (visited on 09/28/2022) (see pages 25, 66, 68).

[123] Morva Mansoury, Maya Hamed, Rashid Karmustaji, Fatima Al Hannan, and Stephen T. Safrany. **The edge effect: A global problem. The trouble with culturing cells in 96-well plates**. en. *Biochemistry and Biophysics Reports* 26 (July 2021), 100987. ISSN: 2405-5808. DOI: 10.1016/j.bbrep.2021.100987. URL: https://www.sciencedirect.com/science/article/pii/S2405580821000819 (visited on 12/07/2022) (see page 26).

[124] Beate Vieth, Swati Parekh, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann. **A systematic evaluation of single cell RNA-seq analysis pipelines**. en. *Nature Communications* 10:1 (Oct. 2019). Number: 1 Publisher: Nature Publishing Group, 4667. ISSN: 2041-1723. DOI: 10.1038/s41467-019-12266-7. URL: https://www.nature.com/articles/s41467-019-12266-7 (visited on 04/19/2021) (see pages 28, 29, 85–88, 103).

[125] Elena Denisenko, Belinda B. Guo, Matthew Jones, Rui Hou, Leanne de Kock, Timo Lassmann, Daniel Poppe, Olivier Clément, Rebecca K. Simmons, Ryan Lister, and Alistair R. R. Forrest. **Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows**. *Genome Biology* 21:1 (June 2020), 130. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02048-6. URL: https://doi.org/10.1186/s13059-020-02048-6 (visited on 02/04/2022) (see pages 28, 87).

[126] Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. **Benchmarking atlas-level data integration in single-cell genomics**. en. *Nature Methods* 19:1 (Jan. 2022). Number: 1 Publisher: Nature Publishing Group, 41–50. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01336-8. URL: https://www.nature.com/articles/s41592-021-01336-8 (visited on 09/30/2022) (see pages 29–31, 33, 88).

[127] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. **Adjusting batch effects in microarray expression data using empirical Bayes methods**. *Biostatistics* 8:1 (Jan. 2007), 118–127. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037. URL: https://doi.org/10.1093/biostatistics/kxj037 (visited on 02/09/2022) (see pages 30, 88, 114).

[128] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. **Fast, sensitive and accurate integration of single-cell data with Harmony**. en. *Nature Methods* 16:12 (Dec. 2019). Number: 12 Publisher: Nature Publishing Group, 1289–1296. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0619-0. URL:

https://www.nature.com/articles/s41592-019-0619-0 (visited on 03/16/2022) (see pages 33, 115, 116).

[129] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. **scGen predicts single-cell perturbation responses**. en. *Nature Methods* 16:8 (Aug. 2019). Number: 8 Publisher: Nature Publishing Group, 715–721. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0494-8. URL: https://www.nature.com/articles/s41592-019-0494-8 (visited on 03/16/2022) (see pages 35, 88, 130).

[130] Nils Eling, Michael D. Morgan, and John C. Marioni. **Challenges in measuring and understanding biological noise**. en. *Nature Reviews Genetics* 20:9 (Sept. 2019). Number: 9 Publisher: Nature Publishing Group, 536–548. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0130-6. URL: https://www.nature.com/articles/s41576-019-0130-6 (visited on 02/04/2022) (see pages 36, 37, 97).

[131] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. **Validation of noise models for single-cell transcriptomics**. en. *Nature Methods* 11:6 (June 2014). Number: 6 Publisher: Nature Publishing Group, 637–640. ISSN: 1548-7105. DOI: 10.1038/nmeth.2930. URL: https://www.nature.com/articles/nmeth.2930 (visited on 02/09/2022) (see pages 37, 89).

[132] Ilias Angelidis, Lukas M. Simon, Isis E. Fernandez, Maximilian Strunz, Christoph H. Mayr, Flavia R. Greiffo, George Tsitsiridis, Meshal Ansari, Elisabeth Graf, Tim-Matthias Strom, Monica Nagendran, Tushar Desai, Oliver Eickelberg, Matthias Mann, Fabian J. Theis, and Herbert B. Schiller. **An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics**. en. *Nature Communications* 10:1 (Feb. 2019). Number: 1 Publisher: Nature Publishing Group, 963. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08831-9. URL: https://www.nature.com/articles/s41467-019-08831-9 (visited on 09/23/2022) (see pages 37, 71, 89, 90, 97).

[133] Celia Pilar Martinez-Jimenez, Nils Eling, Hung-Chang Chen, Catalina A. Vallejos, Aleksandra A. Kolodziejczyk, Frances Connor, Lovorka Stojic, Timothy F. Rayner, Michael J. T. Stubbington, Sarah A. Teichmann, Maike de la Roche, John C. Marioni, and Duncan T. Odom. **Aging increases cell-to-cell transcriptional variability upon immune stimulation**. EN. *Science* (Mar. 2017). Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.aah4115. URL: https://www.science.org/doi/abs/10.1126/science.aah4115 (visited on 02/04/2022) (see pages 37, 89, 90, 97).

[134] Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. **BASiCS: Bayesian Analysis of Single-Cell Sequencing Data**. en. *PLOS Computational Biology* 11:6 (June 2015). Publisher: Public Library of Science, e1004333. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004333. URL: https://journals.plos.org/ploscompbiol/

article?id=10.1371/journal.pcbi.1004333 (visited on 04/19/2021) (see pages 37, 47, 89).

[135] Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Jason C.H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C. Marioni, and Sarah A. Teichmann. **Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation**. *Cell Stem Cell* 17:4 (Oct. 2015), 471–485. ISSN: 1934-5909. DOI: 10.1016/j.stem.2015.09.011. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4595712/ (visited on 12/19/2022) (see page 37).

[136] Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, and Catalina A. Vallejos. **Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data**. *Cell Systems* 7:3 (Sept. 2018), 284–294.e12. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.06.011. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6167088/ (visited on 12/19/2022) (see page 37).

[137] Dominic Grün. **Revealing dynamics of gene expression variability in cell state space**. eng. *Nature Methods* 17:1 (Jan. 2020), 45–49. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0632-3 (see pages 37, 38).

[138] **Correct use of percent coefficient of variation (%CV) formula for log-transformed data**. English. *MOJ Proteomics & Bioinformatics* Volume 6:Issue 4 (Nov. 2017). Publisher: MedCrave Publishing. ISSN: 2374-6920. DOI: 10.15406/mojpb.2017.06.00200. URL: https://medcraveonline.com/MOJPB/MOJPB-06-00200.pdf (visited on 02/09/2022) (see pages 38, 46, 89, 122, 123).

[139] Suguna Rani Krishnaswami, Rashel V. Grindberg, Mark Novotny, Pratap Venepally, Benjamin Lacar, Kunal Bhutani, Sara B. Linker, Son Pham, Jennifer A. Erwin, Jeremy A. Miller, Rebecca Hodge, James K. McCarthy, Martijn Kelder, Jamison McCorrison, Brian D. Aevermann, Francisco Diez Fuertes, Richard H. Scheuermann, Jun Lee, Ed S. Lein, Nicholas Schork, Michael J. McConnell, Fred H. Gage, and Roger S. Lasken. **Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons**. en. *Nature Protocols* 11:3 (Mar. 2016). Number: 3 Publisher: Nature Publishing Group, 499–524. ISSN: 1750-2799. DOI: 10.1038/nprot.2016.015. URL: https://www.nature.com/articles/nprot.2016.015 (visited on 10/14/2022) (see pages 40, 84).

[140] Susanne C. van den Brink, Fanny Sage, Ábel Vértesy, Bastiaan Spanjaard, Josi Peterson-Maduro, Chloé S. Baron, Catherine Robin, and Alexander van Oudenaarden. **Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations**. en. *Nature Methods* 14:10 (Oct. 2017). Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Adult stem

cells;Gene expression analysis Subject_term_id: adult-stem-cells;gene-expression-analysis, 935–936. ISSN: 1548-7105. DOI: 10.1038/nmeth.4437. URL: https://www.nature.com/articles/nmeth.4437 (visited on 06/23/2021) (see pages 40, 84).

[141] Jorge L. Del-Aguila, Zeran Li, Umber Dube, Kathie A. Mihindukulasuriya, John P. Budde, Maria Victoria Fernandez, Laura Ibanez, Joseph Bradley, Fengxian Wang, Kristy Bergmann, Richard Davenport, John C. Morris, David M. Holtzman, Richard J. Perrin, Bruno A. Benitez, Joseph Dougherty, Carlos Cruchaga, and Oscar Harari. **A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain**. *Alzheimer's Research & Therapy* 11 (Aug. 2019), 71. ISSN: 1758-9193. DOI: 10.1186/s13195-019-0524-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6689177/ (visited on 10/14/2022) (see pages 40, 84).

[142] Rashel V. Grindberg, Joyclyn L. Yee-Greenbaum, Michael J. McConnell, Mark Novotny, Andy L. O'Shaughnessy, Georgina M. Lambert, Marcos J. Araúzo-Bravo, Jun Lee, Max Fishman, Gillian E. Robbins, Xiaoying Lin, Pratap Venepally, Jonathan H. Badger, David W. Galbraith, Fred H. Gage, and Roger S. Lasken. **RNA-sequencing from single nuclei**. en. *Proceedings of the National Academy of Sciences* 110:49 (Dec. 2013). Publisher: National Academy of Sciences Section: Biological Sciences, 19802–19807. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1319700110. URL: https://www.pnas.org/content/110/49/19802 (visited on 02/04/2022) (see page 40).

[143] Blue B. Lake, Simone Codeluppi, Yun C. Yung, Derek Gao, Jerold Chun, Peter V. Kharchenko, Sten Linnarsson, and Kun Zhang. **A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA**. en. *Scientific Reports* 7:1 (July 2017). Number: 1 Publisher: Nature Publishing Group, 6031. ISSN: 2045-2322. DOI: 10.1038/s41598-017-04426-w. URL: https://www.nature.com/articles/s41598-017-04426-w (visited on 02/04/2022) (see pages 40, 41, 84, 107).

[144] Antonio Scialdone, Kedar N. Natarajan, Luis R. Saraiva, Valentina Proserpio, Sarah A. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. **Computational assignment of cell-cycle stage from single-cell transcriptome data**. en. *Methods*. Inferring Gene Regulatory Interactions from Quantitative High-Throughput Measurements 85 (Sept. 2015), 54–61. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2015.06.021. URL: https://www.sciencedirect.com/science/article/pii/S1046202315300098 (visited on 02/04/2022) (see pages 43, 64, 119).

[145] Stewart Sell. **Heterogeneity and plasticity of hepatocyte lineage cells**. *Hepatology (Baltimore, Md.)* 33 (Apr. 2001), 738–50. DOI: 10.1053/jhep.2001.21900 (see pages 43, 64).

[146] Janel L. Kopp, Markus Grompe, and Maike Sander. **Stem cells versus plasticity in liver and pancreas regeneration**. en. *Nature Cell Biology* 18:3 (Mar. 2016), 238–245. ISSN: 1465-7392, 1476-4679. DOI: 10.1038/ncb3309. URL: https://www.nature.com/articles/ncb3309 (visited on 09/21/2022) (see pages 43, 64).

[147] Agnes Klochendler, Noa Weinberg-Corem, Maya Moran, Avital Swisa, Nathalie Pochet, Virginia Savova, Jonas Vikeså, Yves Van de Peer, Michael Brandeis, Aviv Regev, Finn Cilius Nielsen, Yuval Dor, and Amir Eden. **A Transgenic Mouse Marking Live Replicating Cells Reveals In00a0Vivo Transcriptional Program of Proliferation**. en. *Developmental Cell* 23:4 (Oct. 2012), 681–690. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2012.08.009. URL: https://www.sciencedirect.com/science/article/pii/S1534580712003772 (visited on 02/07/2023) (see page 43).

[148] Alexandra Pokhilko, Adam E. Handel, Fabiola Curion, Viola Volpato, Emma S. Whiteley, Sunniva Bøstrand, Sarah E. Newey, Colin J. Akerman, Caleb Webber, Michael B. Clark, Rory Bowden, and M. Zameel Cader. **Targeted single-cell RNA sequencing of transcription factors enhances the identification of cell types and trajectories**. *Genome Research* 31:6 (June 2021), 1069–1081. ISSN: 1088-9051. DOI: 10.1101/gr.273961.120. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8168586/ (visited on 02/10/2023) (see page 44).

[149] Fadila Benhamed, Pierre-Damien Denechaud, Maud Lemoine, Céline Robichon, Marthe Moldes, Justine Bertrand-Michel, Vlad Ratziu, Lawrence Serfaty, Chantal Housset, Jacqueline Capeau, Jean Girard, Hervé Guillou, and Catherine Postic. **The lipogenic transcription factor ChREBP dissociates hepatic steatosis from insulin resistance in mice and humans**. en. *The Journal of Clinical Investigation* 122:6 (June 2012). Publisher: American Society for Clinical Investigation, 2176–2194. ISSN: 0021-9738. DOI: 10.1172/JCI41636. URL: https://www.jci.org/articles/view/41636 (visited on 02/07/2022) (see page 44).

[150] Paula Ortega-Prieto and Catherine Postic. **Carbohydrate Sensing Through the Transcription Factor ChREBP**. *Frontiers in Genetics* 10 (2019). ISSN: 1664-8021. URL: https://www.frontiersin.org/articles/10.3389/fgene.2019.00472 (visited on 02/10/2023) (see page 44).

[151] Cesar A. Vargas-Garcia, Khem Raj Ghusinga, and Abhyudai Singh. **Cell size control and gene expression homeostasis in single-cells**. en. *Current Opinion in Systems Biology*. • Regulatory and metabolic networks • Special Section: Single cell and noise 8 (Apr. 2018), 109–116. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2018.01.002. URL: https://www.sciencedirect.com/science/article/pii/S2452310017301592 (visited on 02/09/2022) (see page 46).

[152] **Self-renewing diploid Axin2+ cells fuel homeostatic renewal of the liver | Nature**. In: URL: https://www.nature.com/articles/nature14863 (visited on 09/01/2022) (see pages 51, 92).

[153] Shengda Lin, Elisabete M. Nascimento, Chandresh R. Gajera, Lu Chen, Patrick Neuhöfer, Alina Garbuzov, Sui Wang, and Steven E. Artandi. **Distributed hepatocytes expressing telomerase repopulate the liver in homeostasis and injury**. en. *Nature* 556:7700 (Apr. 2018). Number: 7700 Publisher: Nature Publishing Group, 244–248. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0004-7. URL: https://www.nature.com/articles/s41586-018-0004-7 (visited on 02/09/2022) (see pages 51, 92, 121).

[154] Tianliang Sun, Monika Pikiolek, Vanessa Orsini, Sebastian Bergling, Sjoerd Holwerda, Lapo Morelli, Philipp S. Hoppe, Lara Planas-Paz, Yi Yang, Heinz Ruffner, Tewis Bouwmeester, Felix Lohmann, Luigi M. Terracciano, Guglielmo Roma, Feng Cong, and Jan S. Tchorz. **AXIN2+ Pericentral Hepatocytes Have Limited Contributions to Liver Homeostasis and Regeneration**. en. *Cell Stem Cell* 26:1 (Jan. 2020), 97–107.e6. ISSN: 1934-5909. DOI: 10.1016/j.stem.2019.10.011. URL: https://www.sciencedirect.com/science/article/pii/S1934590919304321 (visited on 02/09/2022) (see pages 51, 92, 93, 121).

[155] **Broad Distribution of Hepatocyte Proliferation in Liver Homeostasis and Regeneration - PMC**. In: URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8009755/ (visited on 09/01/2022) (see pages 51, 53, 121).

[156] Takeshi Katsuda, Kazunori Hosaka, Juntaro Matsuzaki, Wataru Usuba, Marta Prieto-Vila, Tomoko Yamaguchi, Atsunori Tsuchiya, Shuji Terai, and Takahiro Ochiya. **Transcriptomic Dissection of Hepatocyte Heterogeneity: Linking Ploidy, Zonation, and Stem/Progenitor Cell Characteristics**. en. *Cellular and Molecular Gastroenterology and Hepatology* 9:1 (Jan. 2020), 161–183. ISSN: 2352-345X. DOI: 10.1016/j.jcmgh.2019.08.011. URL: http://www.sciencedirect.com/science/article/pii/S2352345X19301171 (visited on 09/17/2020) (see pages 51, 94).

[157] Andreas E Moor and Shalev Itzkovitz. **Spatial transcriptomics: paving the way for tissue-level systems biology**. en. *Current Opinion in Biotechnology*. Systems biology • Nanobiotechnology 46 (Aug. 2017), 126–133. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2017.02.004. URL: https://www.sciencedirect.com/science/article/pii/S0958166916302397 (visited on 02/04/2022) (see pages 53, 93, 94, 96).

[158] J P Iredale, R C Benyon, J Pickering, M McCullen, M Northrop, S Pawley, C Hovell, and M J Arthur. **Mechanisms of spontaneous resolution of rat liver fibrosis. Hepatic stellate cell apoptosis and reduced hepatic expression of metalloproteinase inhibitors.** *Journal of Clinical Investigation* 102:3 (Aug.

1998), 538–549. ISSN: 0021-9738. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC508915/ (visited on 02/14/2023) (see page 57).

[159] Neil C. Henderson, Alison C. Mackinnon, Sarah L. Farnworth, Francoise Poirier, Francesco P. Russo, John P. Iredale, Christopher Haslett, Kenneth J. Simpson, and Tariq Sethi. **Galectin-3 regulates myofibroblast activation and hepatic fibrosis**. *Proceedings of the National Academy of Sciences of the United States of America* 103:13 (Mar. 2006), 5060–5065. ISSN: 0027-8424. DOI: 10.1073/pnas.0511167103. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1458794/ (visited on 02/14/2023) (see page 57).

[160] Neil C. Henderson, Thomas D. Arnold, Yoshio Katamura, Marilyn M. Giacomini, Juan D. Rodriguez, Joseph H. McCarty, Antonella Pellicoro, Elisabeth Raschperger, Christer Betsholtz, Peter G. Ruminski, David W. Griggs, Michael J. Prinsen, Jacquelyn J. Maher, John P. Iredale, Adam Lacy-Hulbert, Ralf H. Adams, and Dean Sheppard. **Targeting of 03b1v integrin identifies a core molecular pathway that regulates fibrosis in several organs**. en. *Nature Medicine* 19:12 (Dec. 2013). Number: 12 Publisher: Nature Publishing Group, 1617–1624. ISSN: 1546-170X. DOI: 10.1038/nm.3282. URL: https://www.nature.com/articles/nm.3282 (visited on 02/09/2022) (see page 57).

[161] Ross Dobie, John R. Wilson-Kanamori, Beth E.P. Henderson, James R. Smith, Kylie P. Matchett, Jordan R. Portman, Karolina Wallenborg, Simone Picelli, Anna Zagorska, Swetha V. Pendem, Thomas E. Hudson, Minnie M. Wu, Grant R. Budas, David G. Breckenridge, Ewen M. Harrison, Damian J. Mole, Stephen J. Wigmore, Prakash Ramachandran, Chris P. Ponting, Sarah A. Teichmann, John C. Marioni, and Neil C. Henderson. **Single-Cell Transcriptomics Uncovers Zonation of Function in the Mesenchyme during Liver Fibrosis**. *Cell Reports* 29:7 (Nov. 2019), 1832–1847.e8. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2019.10.024. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6856722/ (visited on 02/14/2023) (see page 57).

[162] Ahmed Ghallab, Maiju Myllys, Christian H. Holland, Ayham Zaza, Walaa Murad, Reham Hassan, Yasser A. Ahmed, Tahany Abbas, Eman A. Abdelrahim, Kai Markus Schneider, Madlen Matz-Soja, Jörg Reinders, Rolf Gebhardt, Marie-Luise Berres, Maximilian Hatting, Dirk Drasdo, Julio Saez-Rodriguez, Christian Trautwein, and Jan G. Hengstler. **Influence of Liver Fibrosis on Lobular Zonation**. en. *Cells* 8:12 (Dec. 2019). Publisher: Multidisciplinary Digital Publishing Institute (MDPI). DOI: 10.3390/cells8121556. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6953125/ (visited on 02/14/2023) (see page 57).

[163] Minoru Tanaka and Atsushi Miyajima. **Liver regeneration and fibrosis after inflammation**. *Inflammation and Regeneration* 36 (Oct. 2016), 19. ISSN: 1880-9693. DOI: 10.1186/s41232-016-0025-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5725806/ (visited on 02/14/2023) (see page 57).

[164] Eva Sanchez-Quant, Maria Lucia Richter, Maria Colomé-Tatché, and Celia Pilar Martinez-Jimenez. **Single-cell metabolic profiling reveals subgroups of primary human hepatocytes with heterogeneous responses to drug challenge**. *Genome Biology* 24:1 (Oct. 2023), 234. ISSN: 1474-760X. DOI: 10.1186/s13059-023-03075-9. URL: https://doi.org/10.1186/s13059-023-03075-9 (visited on 03/01/2024) (see pages 61, 63, 64, 72, 75, 78–80, 118).

[165] Ana S. Serras, Joana S. Rodrigues, Madalena Cipriano, Armanda V. Rodrigues, Nuno G. Oliveira, and Joana P. Miranda. **A Critical Perspective on 3D Liver Models for Drug Metabolism and Toxicology Studies**. *Frontiers in Cell and Developmental Biology* 9 (2021). ISSN: 2296-634X. URL: https://www.frontiersin.org/article/10.3389/fcell.2021.626805 (visited on 02/11/2022) (see page 62).

[166] Jonathan C. Cohen, Jay D. Horton, and Helen H. Hobbs. **Human Fatty Liver Disease: Old Questions and New Insights**. *Science (New York, N.Y.)* 332:6037 (June 2011), 1519–1523. ISSN: 0036-8075. DOI: 10.1126/science.1204265. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3229276/ (visited on 09/23/2022) (see pages 71, 73).

[167] Xiongfeng Pan, Atipatsa Chiwanda Kaminga, Aizhong Liu, Shi Wu Wen, Jihua Chen, and Jiayou Luo. **Chemokines in Non-alcoholic Fatty Liver Disease: A Systematic Review and Network Meta-Analysis**. *Frontiers in Immunology* 11 (Sept. 2020), 1802. ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.01802. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7530185/ (visited on 09/23/2022) (see pages 71, 97).

[168] Ting-Fang Kuo, Hideki Tatsukawa, Tomokazu Matsuura, Keisuke Nagatsuma, Shigehisa Hirose, and Soichi Kojima. **Free fatty acids induce transglutaminase 2-dependent apoptosis in hepatocytes via ER stress-stimulated PERK pathways**. en. *Journal of Cellular Physiology* 227:3 (2012). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcp.22833, 1130–1137. ISSN: 1097-4652. DOI: 10.1002/jcp.22833. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcp.22833 (visited on 02/21/2023) (see page 71).

[169] GuQi Wang, Herbert L. Bonkovsky, Andrew de Lemos, and Frank J. Burczynski. **Recent insights into the biological functions of liver fatty acid binding protein 1**. *Journal of Lipid Research* 56:12 (Dec. 2015), 2238–2247. ISSN: 0022-2275. DOI: 10.1194/jlr.R056705. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4655993/ (visited on 02/21/2023) (see page 73).

[170] T Aoyama, M Souri, S Ushikubo, T Kamijo, S Yamaguchi, R I Kelley, W J Rhead, K Uetake, K Tanaka, and T Hashimoto. **Purification of human very-long-chain acyl-coenzyme A dehydrogenase and characterization of its deficiency in seven patients**. *Journal of Clinical Investigation* 95:6 (June 1995), 2465–2473.

ISSN: 0021-9738. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC295925/ (visited on 02/21/2023) (see page 73).

[171] Dan Liu, Peng Zhang, Junjie Zhou, Rufang Liao, Yan Che, Mao-Mao Gao, Jiaqi Sun, Jingjing Cai, Xu Cheng, Yongping Huang, Guopeng Chen, Hongyu Nie, Yan-Xiao Ji, Xiao-Jing Zhang, Zan Huang, Haibo Xu, Zhi-Gang She, and Hongliang Li. **TNFAIP3 Interacting Protein 3 Overexpression Suppresses Nonalcoholic Steatohepatitis by Blocking TAK1 Activation**. en. *Cell Metabolism* 31:4 (Apr. 2020), 726–740.e8. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2020.03.007. URL: https://www.sciencedirect.com/science/article/pii/S1550413120301224 (visited on 02/21/2023) (see page 73).

[172] Jung-Hwan Baek, Da-Hyun Kim, Jaegyeong Lee, Seok-Jun Kim, and Kyung-Hee Chun. **Galectin-1 accelerates high-fat diet-induced obesity by activation of peroxisome proliferator-activated receptor gamma (PPAR03b3) in mice**. *Cell Death & Disease* 12:1 (Jan. 2021), 66. ISSN: 2041-4889. DOI: 10.1038/s41419-020-03367-z. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7801586/ (visited on 02/21/2023) (see page 73).

[173] Neha J. Pagidipati, Michael Pencina, and Allan D. Sniderman. **The Enigma of Glucose and Lipid Metabolism**. *JAMA Cardiology* 1:2 (May 2016), 145–146. ISSN: 2380-6583. DOI: 10.1001/jamacardio.2016.0183. URL: https://doi.org/10.1001/jamacardio.2016.0183 (visited on 09/26/2022) (see page 73).

[174] M. Teresa Donato, Agustín Lahoz, Nuria Jiménez, Gabriela Pérez, Alfonso Serralta, José Mir, José V. Castell, and M. José Gómez-Lechón. **Potential Impact of Steatosis on Cytochrome P450 Enzymes of Human Hepatocytes Isolated from Fatty Liver Grafts**. en. *Drug Metabolism and Disposition* 34:9 (Sept. 2006). Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Article, 1556–1562. ISSN: 0090-9556, 1521-009X. DOI: 10.1124/dmd.106.009670. URL: https://dmd.aspetjournals.org/content/34/9/1556 (visited on 09/26/2022) (see pages 74, 77, 98, 101).

[175] Benjamin Berger, Fabio Bachmann, Urs Duthaler, Stephan Krähenbühl, and Manuel Haschke. **Cytochrome P450 Enzymes Involved in Metoprolol Metabolism and Use of Metoprolol as a CYP2D6 Phenotyping Probe Drug**. *Frontiers in Pharmacology* 9 (July 2018), 774. ISSN: 1663-9812. DOI: 10.3389/fphar.2018.00774. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6066528/ (visited on 02/24/2023) (see pages 74, 99).

[176] Roshan M. Borkar, Murali Mohan Bhandi, Ajay P. Dubey, V. Ganga Reddy, Prashanth Komirishetty, Prajwal P. Nandekar, Abhay T. Sangamwar, Ahmed Kamal, Sanjay K. Banerjee, and R. Srinivas. **An evaluation of the CYP2D6 and CYP3A4 inhibition potential of metoprolol metabolites and their contribution to drug–drug and drug–herb interaction by LC-ESI/MS/MS**.

en. *Biomedical Chromatography* 30:10 (2016), 1556–1572. ISSN: 1099-0801. DOI: 10.1002/bmc.3721. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bmc.3721 (visited on 02/24/2023) (see pages 76, 99).

[177] Yoshiyuki Shirasaka, Jennifer E. Sager, Justin D. Lutz, Connie Davis, and Nina Isoherranen. **Inhibition of CYP2C19 and CYP3A4 by Omeprazole Metabolites and Their Contribution to Drug-Drug Interactions**. *Drug Metabolism and Disposition* 41:7 (July 2013), 1414–1424. ISSN: 0090-9556. DOI: 10.1124/dmd.113.051722. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3684819/ (visited on 02/24/2023) (see pages 76, 99).

[178] Allan Peter Davis, Thomas C. Wiegers, Phoebe M. Roberts, Benjamin L. King, Jean M. Lay, Kelley Lennon-Hopkins, Daniela Sciaky, Robin Johnson, Heather Keating, Nigel Greene, Robert Hernandez, Kevin J. McConnell, Ahmed E. Enayetallah, and Carolyn J. Mattingly. **A CTD–Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug–disease and drug–phenotype interactions**. *Database: The Journal of Biological Databases and Curation* 2013 (Nov. 2013), bat080. ISSN: 1758-0463. DOI: 10.1093/database/bat080. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776/ (visited on 09/23/2022) (see pages 77, 101, 121).

[179] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. **Comparative Toxicogenomics Database (CTD): update 2021**. *Nucleic Acids Research* 49:D1 (Jan. 2021), D1138–D1143. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa891. URL: https://doi.org/10.1093/nar/gkaa891 (visited on 09/23/2022) (see pages 77, 121).

[180] J. Aubert, K. Begriche, L. Knockaert, M. A. Robin, and B. Fromenty. **Increased expression of cytochrome P450 2E1 in nonalcoholic fatty liver disease: Mechanisms and pathophysiological role**. en. *Clinics and Research in Hepatology and Gastroenterology* 35:10 (Oct. 2011), 630–637. ISSN: 2210-7401. DOI: 10.1016/j.clinre.2011.04.015. URL: https://www.sciencedirect.com/science/article/pii/S2210740111001732 (visited on 09/26/2022) (see pages 77, 81, 101).

[181] David N. Juurlink, Muhammad Mamdani, Alexander Kopp, Andreas Laupacis, and Donald A. Redelmeier. **Drug-Drug Interactions Among Elderly Patients Hospitalized for Drug Toxicity**. *JAMA* 289:13 (Apr. 2003), 1652–1658. ISSN: 0098-7484. DOI: 10.1001/jama.289.13.1652. URL: https://doi.org/10.1001/jama.289.13.1652 (visited on 09/26/2022) (see page 77).

[182] Andreas Benesic, Kowcee Jalal, and Alexander L Gerbes. **Drug-Drug Combinations Can Enhance Toxicity as Shown by Monocyte-Derived Hepatocyte-like Cells From Patients With Idiosyncratic Drug-Induced Liver00a0Injury**. *Toxicological Sciences* 171:2 (Oct. 2019), 296–302. ISSN: 1096-6080. DOI: 10.1093/

toxsci/kfz156. URL: https://doi.org/10.1093/toxsci/kfz156 (visited on 09/26/2022) (see page 77).

[183] Kristina M. Utzschneider and Steven E. Kahn. **The Role of Insulin Resistance in Nonalcoholic Fatty Liver Disease**. *The Journal of Clinical Endocrinology & Metabolism* 91:12 (Dec. 2006), 4753–4761. ISSN: 0021-972X. DOI: 10.1210/jc.2006-0587. URL: https://doi.org/10.1210/jc.2006-0587 (visited on 02/23/2023) (see page 81).

[184] gianni. **Fatty liver and drugs: the two sides of the same coin**. en. In: Mar. 2017. URL: https://www.europeanreview.org/article/12435 (visited on 02/23/2023) (see page 81).

[185] Xu Li, Pujun Gao, and Junqi Niu. **Metabolic Comorbidities and Risk of Development and Severity of Drug-Induced Liver Injury**. *BioMed Research International* 2019 (Aug. 2019), 8764093. ISSN: 2314-6133. DOI: 10.1155/2019/8764093. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6720367/ (visited on 02/23/2023) (see page 81).

[186] Palrasu Manikandan and Siddavaram Nagini. **Cytochrome P450 Structure, Function and Clinical Significance: A Review**. *Current Drug Targets* 19:1 (Jan. 2018), 38–54. DOI: 10.2174/1389450118666170125144557 (see page 81).

[187] Qiao Rui Xing, Nadia Omega Cipta, Kiyofumi Hamashima, Yih-Cherng Liou, Cheng Gee Koh, and Yuin-Han Loh. **Unraveling Heterogeneity in Transcriptome and Its Regulation Through Single-Cell Multi-Omics Technologies**. English. *Frontiers in Genetics* 11 (2020). Publisher: Frontiers. ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00662. URL: https://www.frontiersin.org/articles/10.3389/fgene.2020.00662/full (visited on 08/10/2020) (see pages 83, 103).

[188] Quan Zhao, Jianghui Wang, Ilya V. Levichkin, Stan Stasinopoulos, Michael T. Ryan, and Nicholas J. Hoogenraad. **A mitochondrial specific stress response in mammalian cells**. en. *The EMBO Journal* 21:17 (Sept. 2002). Publisher: European Molecular Biology Organization, 4411. DOI: 10.1093/emboj/cdf445. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC126185/ (visited on 10/14/2022) (see pages 84, 111).

[189] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C. Marioni, and Sarah A. Teichmann. **Classification of low quality cells from single-cell RNA-seq data**. *Genome Biology* 17 (Feb. 2016), 29. ISSN: 1474-7596. DOI: 10.1186/s13059-016-0888-1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4758103/ (visited on 10/14/2022) (see pages 84, 111).

[190] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. **Quantitative single-cell RNA-seq with unique molecular identifiers**. en. *Nature Methods* 11:2 (Feb. 2014). Number: 2 Publisher: Nature Publishing Group, 163–166. ISSN: 1548-7105. DOI: 10.1038/nmeth.2772. URL: https://www.nature.com/articles/nmeth.2772 (visited on 09/06/2022) (see pages 85, 109).

[191] Broad Institute. **Picard Tools**. In: http://broadinstitute.github.io/picard/. URL: http://broadinstitute.github.io/picard/ (see page 85).

[192] Alex K. Shalek, Rahul Satija, Xian Adiconis, Rona S. Gertner, Jellert T. Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J. Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z. Levin, Hongkun Park, and Aviv Regev. **Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells**. en. *Nature* 498:7453 (June 2013). Number: 7453 Publisher: Nature Publishing Group, 236–240. ISSN: 1476-4687. DOI: 10.1038/nature12172. URL: https://www.nature.com/articles/nature12172 (visited on 02/09/2022) (see pages 89, 91).

[193] Rumana Bahar, Claudia H. Hartmann, Karl A. Rodriguez, Ashley D. Denny, Rita A. Busuttil, Martijn E. T. Dollé, R. Brent Calder, Gary B. Chisholm, Brad H. Pollock, Christoph A. Klein, and Jan Vijg. **Increased cell-to-cell variation in gene expression in ageing mouse heart**. en. *Nature* 441:7096 (June 2006). Number: 7096 Publisher: Nature Publishing Group, 1011–1014. ISSN: 1476-4687. DOI: 10.1038/nature04844. URL: https://www.nature.com/articles/nature04844 (visited on 07/07/2023) (see page 90).

[194] John Orlowski and Sergio Grinstein. **Diversity of the mammalian sodium proton exchanger SLC9 gene family**. en. *Pflügers Archiv* 447:5 (Feb. 2004), 549–565. ISSN: 1432-2013. DOI: 10.1007/s00424-003-1110-3. URL: https://doi.org/10.1007/s00424-003-1110-3 (visited on 10/05/2022) (see page 91).

[195] Masami Ueda, Tomohiro Iguchi, Takaaki Masuda, Hisateru Komatsu, Sho Nambara, Shotaro Sakimura, Hidenari Hirata, Ryutaro Uchi, Hidetoshi Eguchi, Shuhei Ito, Keishi Sugimachi, Tsunekazu Mizushima, Yuichiro Doki, Masaki Mori, and Koshi Mimori. **Up-regulation of SLC9A9 Promotes Cancer Progression and Is Involved in Poor Prognosis in Colorectal Cancer**. en. *Anticancer Research* 37:5 (May 2017). Publisher: International Institute of Anticancer Research Section: Experimental Studies, 2255–2263. ISSN: 0250-7005, 1791-7530. URL: https://ar.iiarjournals.org/content/37/5/2255 (visited on 10/05/2022) (see page 91).

[196] Leslie M Fischer, Kerry Ann daCosta, Lester Kwock, Paul W Stewart, Tsui-Shan Lu, Sally P Stabler, Robert H Allen, and Steven H Zeisel. **Sex and menopausal status influence human dietary requirements for the nutrient choline**. *The American journal of clinical nutrition* 85:5 (May 2007), 1275–1285. ISSN: 0002-9165. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2435503/ (visited on 10/05/2022) (see page 91).

[197] Rutger D. Luteijn, Shivam A. Zaver, Benjamin G. Gowen, Stacia K. Wyman, Nick E. Garelis, Liberty Onia, Sarah M. McWhirter, George E. Katibah, Jacob E. Corn, Joshua J. Woodward, and David H. Raulet. **SLC19A1 transports immunoreactive cyclic dinucleotides**. en. *Nature* 573:7774 (Sept. 2019). Number: 7774 Publisher: Nature Publishing Group, 434–438. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1553-0. URL: https://www.nature.com/articles/s41586-019-1553-0 (visited on 10/05/2022) (see page 92).

[198] Vincent M. Perez, Jeffrey Gabell, Mark Behrens, Nishikant Wase, Concetta C. DiRusso, and Paul N. Black. **Deletion of fatty acid transport protein 2 (FATP2) in the mouse liver changes the metabolic landscape by increasing the expression of PPAR03b1-regulated genes**. en. *Journal of Biological Chemistry* 295:17 (Apr. 2020), 5737–5750. ISSN: 0021-9258. DOI: 10.1074/jbc.RA120.012730. URL: https://www.sciencedirect.com/science/article/pii/S0021925817503002 (visited on 10/05/2022) (see page 92).

[199] Patrick D. Wilkinson, Frances Alencastro, Evan R. Delgado, Madeleine P. Leek, Matthew P. Weirich, P. Anthony Otero, Nairita Roy, Whitney K. Brown, Michael Oertel, and Andrew W. Duncan. **Polyploid Hepatocytes Facilitate Adaptation and Regeneration to Chronic Liver Injury**. en. *The American Journal of Pathology* 189:6 (June 2019), 1241–1255. ISSN: 0002-9440. DOI: 10.1016/j.ajpath.2019.02.008. URL: https://www.sciencedirect.com/science/article/pii/S0002944018309702 (visited on 02/09/2022) (see page 92).

[200] Patrick D. Wilkinson, Evan R. Delgado, Frances Alencastro, Madeleine P. Leek, Nairita Roy, Matthew P. Weirich, Elizabeth C. Stahl, P. Anthony Otero, Maelee I. Chen, Whitney K. Brown, and Andrew W. Duncan. **The polyploid state restricts hepatocyte proliferation and liver regeneration**. *Hepatology (Baltimore, Md.)* 69:3 (Mar. 2019), 1242–1258. ISSN: 0270-9139. DOI: 10.1002/hep.30286. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6532408/ (visited on 10/05/2022) (see page 92).

[201] Feng Chen, Robert J. Jimenez, Khushbu Sharma, Hubert Y. Luu, Bernadette Y. Hsu, Ajay Ravindranathan, Bradley A. Stohr, and Holger Willenbring. **Broad Distribution of Hepatocyte Proliferation in Liver Homeostasis and Regeneration**. *Cell stem cell* 26:1 (Jan. 2020), 27–33.e4. ISSN: 1934-5909. DOI: 10.1016/j.stem.2019.11.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8009755/ (visited on 10/05/2022) (see page 92).

[202] Tomonori Matsumoto, Leslie Wakefield, Branden David Tarlow, and Markus Grompe. **In00a0Vivo Lineage Tracing of Polyploid Hepatocytes Reveals Extensive Proliferation during Liver Regeneration**. en. *Cell Stem Cell* 26:1 (Jan. 2020), 34–47.e3. ISSN: 1934-5909. DOI: 10.1016/j.stem.2019.11.014. URL: https://www.sciencedirect.com/science/article/pii/S1934590919304692 (visited on 02/09/2022) (see page 92).

[203] Reiichiro Kuwahara, Alexander V. Kofman, Charles S. Landis, E. Scott Swenson, Els Barendswaard, and Neil D. Theise. **The Hepatic Stem Cell Niche: Identification by Label-Retaining Cell Assay**. *Hepatology (Baltimore, Md.)* 47:6 (June 2008), 1994–2002. ISSN: 0270-9139. DOI: 10.1002/hep.22218. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847183/ (visited on 05/09/2022) (see page 92).

[204] F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. **PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells**. *Genome Biology* 20:1 (Mar. 2019), 59. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1663-x. URL: https://doi.org/10.1186/s13059-019-1663-x (visited on 02/09/2022) (see pages 93, 96).

[205] Quentin M. Anstee, Helen L. Reeves, Elena Kotsiliti, Olivier Govaere, and Mathias Heikenwalder. **From NASH to HCC: current concepts and future challenges**. en. *Nature Reviews Gastroenterology & Hepatology* 16:7 (July 2019). Number: 7 Publisher: Nature Publishing Group, 411–428. ISSN: 1759-5053. DOI: 10.1038/s41575-019-0145-7. URL: https://www.nature.com/articles/s41575-019-0145-7 (visited on 02/11/2022) (see page 97).

[206] Michele Alves-Bezerra and David E. Cohen. **Triglyceride metabolism in the liver**. *Comprehensive Physiology* 8:1 (Dec. 2017), 1–8. ISSN: 2040-4603. DOI: 10.1002/cphy.c170012. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6376873/ (visited on 10/12/2022) (see pages 97, 100).

[207] Karima Begriche, Julie Massart, Marie-Anne Robin, Fabrice Bonnet, and Bernard Fromenty. **Mitochondrial adaptations and dysfunctions in nonalcoholic fatty liver disease**. en. *Hepatology* 58:4 (2013), 1497–1507. ISSN: 1527-3350. DOI: 10.1002/hep.26226. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hep.26226 (visited on 10/12/2022) (see pages 97, 101, 102).

[208] James R Vinnai. **The Association Between Oxidative Stress, Cellular Differentiation And Galectins In Human Promyelocytic Leukemia Cells (HL-60)**. en (), 135 (see page 98).

[209] Masayuki Mogi, Akiko Toda, Kazuhide Iwasaki, Shogo Kusumoto, Hiromi Takehara, Makiko Shimizu, Norie Murayama, Hiroyuki Izumi, Masahiro Utoh, and Hiroshi Yamazaki. **Simultaneous pharmacokinetics assessment of caffeine, warfarin, omeprazole, metoprolol, and midazolam intravenously or orally administered to Microminipigs**. *The Journal of Toxicological Sciences* 37:6 (2012), 1157–1164. DOI: 10.2131/jts.37.1157 (see page 99).

[210] **The Genetics of Alcohol Metabolism - ProQuest**. en. In: URL: https://www.proquest.com/docview/222391732?fromopenview=true&pq-origsite=gscholar (visited on 05/19/2023) (see page 99).

[211] Ron Weathermon and David W. Crabb. **Alcohol and Medication Interactions**. *Alcohol Research & Health* 23:1 (1999), 40–54. ISSN: 1535-7414. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6761694/ (visited on 05/19/2023) (see page 99).

[212] Amy L Johnson and Kelly L Hayward. **Managing medicines in alcohol-associated liver disease: a practical review**. *Australian Prescriber* 44:3 (June 2021), 96–106. ISSN: 0312-8008. DOI: 10.18773/austprescr.2021.015. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8236868/ (visited on 05/19/2023) (see page 99).

[213] Mohamed A. Abdelmegeed, Youngshim Choi, Seung-Kwon Ha, and Byoung-Joon Song. **Cytochrome P450-2E1 promotes aging-related hepatic steatosis, apoptosis and fibrosis through increased nitroxidative stress**. *Free radical biology & medicine* 91 (Feb. 2016), 188–202. ISSN: 0891-5849. DOI: 10.1016/j.freeradbiomed.2015.12.016. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4761508/ (visited on 07/07/2023) (see page 100).

[214] Gina Song, Xueying Sun, Ronald N. Hines, D. Gail McCarver, Brian G. Lake, Thomas G. Osimitz, Moire R. Creek, Harvey J. Clewell, and Miyoung Yoon. **Determination of Human Hepatic CYP2C8 and CYP1A2 Age-Dependent Expression to Support Human Health Risk Assessment for Early Ages**. en. *Drug Metabolism and Disposition* 45:5 (May 2017). Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Article, 468–475. ISSN: 0090-9556, 1521-009X. DOI: 10.1124/dmd.116.074583. URL: https://dmd.aspetjournals.org/content/45/5/468 (visited on 07/07/2023) (see page 100).

[215] Jie Liu, Yuan-Fu Lu, J. Christopher Corton, and Curtis D. Klaassen. **Expression of cytochrome P450 isozyme transcripts and activities in human livers**. *Xenobiotica; the fate of foreign compounds in biological systems* 51:3 (Mar. 2021), 279–286. ISSN: 0049-8254. DOI: 10.1080/00498254.2020.1867929. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8215780/ (visited on 07/07/2023) (see page 100).

[216] Ulrich Klotz. **Influence of liver disease on the elimination of drugs**. en. *European Journal of Drug Metabolism and Pharmacokinetics* 1:3 (July 1976), 129–140. ISSN: 2107-0180. DOI: 10.1007/BF03189267. URL: https://doi.org/10.1007/BF03189267 (visited on 10/14/2022) (see page 101).

[217] R. K. Verbeeck and Y. Horsmans. **Effect of hepatic insufficiency on pharmacokinetics and drug dosing**. eng. *Pharmacy world & science: PWS* 20:5 (Oct. 1998), 183–192. ISSN: 0928-1231. DOI: 10.1023/a:1008656930082 (see page 101).

[218] Maria Greabu, Silviu Constantin Badoiu, Iulia-Ioana Stanescu-Spinu, Daniela Miricescu, Alexandra Ripszky Totan, Silvia Elena Badoiu, Michel Costagliola, and Viorel Jinga. **Drugs Interfering with Insulin Resistance and Their Influence on the Associated Hypermetabolic State in Severe Burns: A Narrative Review**. *International Journal of Molecular Sciences* 22:18 (Sept. 2021), 9782. ISSN: 1422-0067. DOI: 10.3390/ijms22189782. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8466307/ (visited on 10/14/2022) (see page 101).

[219] Bruno Fève and André J. Scheen. **When therapeutic drugs lead to diabetes**. en. *Diabetologia* 65:5 (May 2022), 751–762. ISSN: 1432-0428. DOI: 10.1007/s00125-022-05666-w. URL: https://doi.org/10.1007/s00125-022-05666-w (visited on 10/14/2022) (see pages 101, 102).

[220] Leon A. Adams, Scott W. White, Julie A. Marsh, Stephen J. Lye, Kristin L. Connor, Richard Maganga, Oyekoya T. Ayonrinde, John K. Olynyk, Trevor A. Mori, Lawrence J. Beilin, Lyle J. Palmer, Jeffrey M. Hamdorf, and Craig E. Pennell. **Association between liver-specific gene polymorphisms and their expression levels with nonalcoholic fatty liver disease**. en. *Hepatology* 57:2 (2013). _eprint: https://aasldpubs.onlinelibrary.wiley.com/doi/pdf/10.1002/hep.26184, 590–600. ISSN: 1527-3350. DOI: 10.1002/hep.26184. URL: https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.26184 (visited on 08/10/2020) (see page 102).

[221] Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. **Single-cell chromatin accessibility reveals principles of regulatory variation**. en. *Nature* 523:7561 (July 2015), 486–490. ISSN: 1476-4687. DOI: 10.1038/nature14590. URL: https://www.nature.com/articles/nature14590 (visited on 01/05/2019) (see page 103).

[222] Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. **Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing**. en. *Science* 348:6237 (May 2015), 910–914. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aab1601. URL: http://science.sciencemag.org/content/348/6237/910 (visited on 01/07/2019) (see page 103).

[223] Chongyuan Luo, Christopher L. Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, Brian Bui, Terrence J. Sejnowski, Timothy T. Harkins, Eran A. Mukamel, M. Margarita Behrens, and Joseph R. Ecker. **Single Cell Methylomes Identify Neuronal Subtypes and Regulatory Elements in Mammalian Cortex**. *Science (New York, N.Y.)* 357:6351 (Aug. 2017), 600–604. ISSN: 0036-8075. DOI: 10.1126/science.aan3351. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570439/ (visited on 01/08/2019) (see page 103).

[224] Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. **Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin**. en. *Cell* 183:4 (Nov. 2020), 1103–1116.e20. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.09.056. URL: https://www.sciencedirect.com/science/article/pii/S0092867420312538 (visited on 10/19/2022) (see page 103).

[225] Stephen J. Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C. Marioni, Oliver Stegle, and Wolf Reik. **scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells**. En. *Nature Communications* 9:1 (Feb. 2018), 781. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03149-4. URL: https://www.nature.com/articles/s41467-018-03149-4 (visited on 02/08/2019) (see page 103).

[226] Michaela Asp, Joseph Bergenstråhle, and Joakim Lundeberg. **Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration**. en. *BioEssays* 42:10 (2020), 1900221. ISSN: 1521-1878. DOI: 10.1002/bies.201900221. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201900221 (visited on 10/19/2022) (see page 104).

[227] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. **Visualization and analysis of gene expression in tissue sections by spatial transcriptomics**. en. *Science* 353:6294 (July 2016), 78–82. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaf2403. URL: http://science.sciencemag.org/content/353/6294/78 (visited on 01/29/2019) (see page 104).

[228] Johannes Wirth, Nina Compera, Kelvin Yin, Sophie Brood, Simon Chang, Celia P. Martinez-Jimenez, and Matthias Meier. **Spatial Transcriptomics Using Multiplexed Deterministic Barcoding in Tissue**. en. In: Pages: 2022.08.30.505834

Section: New Results. bioRxiv, Sept. 2022. DOI: 10.1101/2022.08.30.505834. URL: https://www.biorxiv.org/content/10.1101/2022.08.30.505834v1 (visited on 10/19/2022) (see page 104).

[229] Ashwini Kumar, Matti Kankainen, Alun Parsons, Olli Kallioniemi, Pirkko Mattila, and Caroline A. Heckman. **The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia**. *BMC Genomics* 18:1 (Aug. 2017), 629. ISSN: 1471-2164. DOI: 10.1186/s12864-017-4039-1. URL: https://doi.org/10.1186/s12864-017-4039-1 (visited on 09/06/2022) (see page 107).

[230] Shanrong Zhao, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack. **Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion**. en. *Scientific Reports* 8:1 (Mar. 2018). Number: 1 Publisher: Nature Publishing Group, 4781. ISSN: 2045-2322. DOI: 10.1038/s41598-018-23226-4. URL: https://www.nature.com/articles/s41598-018-23226-4 (visited on 09/06/2022) (see page 107).

[231] Hazuki Takahashi, Timo Lassmann, Mitsuyoshi Murata, and Piero Carninci. **5' end-centered expression profiling using Cap-analysis gene expression (CAGE) and next-generation sequencing**. *Nature protocols* 7:3 (Feb. 2012), 542–561. ISSN: 1754-2189. DOI: 10.1038/nprot.2012.005. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4094379/ (visited on 09/06/2022) (see page 107).

[232] Francis S. Collins and Leslie Fink. **The Human Genome Project**. *Alcohol Health and Research World* 19:3 (1995), 190–195. ISSN: 0090-838X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875757/ (visited on 09/06/2022) (see page 107).

[233] Eric S. Lander et al. **Initial sequencing and analysis of the human genome**. en. *Nature* 409:6822 (Feb. 2001). Number: 6822 Publisher: Nature Publishing Group, 860–921. ISSN: 1476-4687. DOI: 10.1038/35057062. URL: https://www.nature.com/articles/35057062 (visited on 09/06/2022) (see page 107).

[234] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. **Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq**. en. *Genome Research* 21:7 (July 2011). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 1160–1167. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.110882.110. URL: https://genome.cshlp.org/content/21/7/1160 (visited on 09/06/2022) (see page 108).

[235] **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features | Bioinformatics | Oxford Academic**. In: URL: https://academic.oup.com/bioinformatics/article/30/7/923/232889 (visited on 09/06/2022) (see page 108).

[236] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. **Near-optimal probabilistic RNA-seq quantification**. en. *Nature Biotechnology* 34:5 (May 2016). Number: 5 Publisher: Nature Publishing Group, 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. URL: https://www.nature.com/articles/nbt.3519 (visited on 05/19/2023) (see page 109).

[237] R Core Team. **R: A Language and Environment for Statistical Computing** (). URL: https://www.R-project.org/ (see page 110).

[238] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. **Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data**. eng. *Cell Systems* 8:4 (Apr. 2019), 281–291.e9. ISSN: 2405-4720. DOI: 10.1016/j.cels.2018.11.005 (see page 111).

[239] Leland McInnes, John Healy, and James Melville. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. *arXiv:1802.03426 [cs, stat]* (Feb. 2018). arXiv: 1802.03426. URL: http://arxiv.org/abs/1802.03426 (visited on 01/10/2019) (see page 116).

[240] Alexandra B Keenan, Denis Torre, Alexander Lachmann, Ariel K Leong, Megan L Wojciechowicz, Vivian Utti, Kathleen M Jagodnik, Eryk Kropiwnicki, Zichen Wang, and Avi Ma'ayan. **ChEA3: transcription factor enrichment analysis by orthogonal omics integration**. *Nucleic Acids Research* 47:W1 (July 2019), W212–W224. ISSN: 0305-1048. DOI: 10.1093/nar/gkz446. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6602523/ (visited on 05/26/2023) (see page 120).

[241] Steven Xijin Ge, Dongmin Jung, and Runan Yao. **ShinyGO: a graphical gene-set enrichment tool for animals and plants**. *Bioinformatics* 36:8 (Apr. 2020), 2628–2629. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz931. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7178415/ (visited on 05/26/2023) (see page 121).

[242] Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J. Daly, Nick Patterson, Jill P. Mesirov, Todd R. Golub, Pablo Tamayo, Bruce Spiegelman, Eric S. Lander, Joel N. Hirschhorn, David Altshuler, and Leif C. Groop. **PGC-103b1-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. en. *Nature Genetics* 34:3 (July 2003). Number: 3 Publisher: Nature Publishing Group, 267–273. ISSN: 1546-1718. DOI: 10.1038/ng1180. URL: https://www.nature.com/articles/ng1180 (visited on 05/26/2023) (see page 121).

[243] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 102:43 (Oct. 2005). Publisher: Proceedings of the National Academy of Sciences, 15545–15550. DOI: 10.1073/pnas.0506580102. URL: https://www.pnas.org/doi/10.1073/pnas.0506580102 (visited on 05/26/2023) (see page 121).

[244] Xiaobo Wang, Ze Zheng, Jorge Matias Caviglia, Kathleen E. Corey, Tina M. Herfel, Bishuang Cai, Ricard Masia, Raymond T. Chung, Jay H. Lefkowitch, Robert F. Schwabe, and Ira Tabas. **Hepatocyte TAZ/WWTR1 Promotes Inflammation and Fibrosis in Nonalcoholic Steatohepatitis**. en. *Cell Metabolism* 24:6 (Dec. 2016), 848–862. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2016.09.016. URL: https://www.sciencedirect.com/science/article/pii/S1550413116305010 (visited on 02/07/2022) (see page 121).

List of Publications

Articles in Refereed Journals

[1] **Single-nucleus RNA-seq2 reveals functional crosstalk between liver zonation and ploidy**. en. *Nature Communications* 12:1 (July 2021). Number: 1 Publisher: Nature Publishing Group, 4264. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24543-5. URL: https://www.nature.com/articles/s41467-021-24543-5 (visited on 02/04/2022).M. L. Richter, I. K. Deligiannis, K. Yin, A. Danese, E. Lleshi, P. Coupland, C. A. Vallejos, K. P. Matchett, N. C. Henderson, M. Colome-Tatche, and C. P. Martinez-Jimenez.

[2] **Single-cell metabolic profiling reveals subgroups of primary human hepatocytes with heterogeneous responses to drug challenge**. *Genome Biology* 24:1 (Oct. 2023), 234. ISSN: 1474-760X. DOI: 10.1186/s13059-023-03075-9. URL: https://doi.org/10.1186/s13059-023-03075-9 (visited on 03/01/2024).Eva Sanchez-Quant, Maria Lucia Richter, Maria Colomé-Tatché, and Celia Pilar Martinez-Jimenez.