Single-cell response to perturbations across biological scales: single organ, organ system and phenotypic individuals

by

Kellie Elizabeth Kolb

B.S., University of Wisconsin-Madison (2014)

Submitted to the Department of Chemistry
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Signature redacted

Signature of Author .....
Department of Chemistry
May 10, 2019

Signature redacted

Certified by.....
Alex K. Shalek
Pfizer-Laubach Career Development Assistant Professor of Chemistry
Thesis Supervisor

Signature redacted

Accepted by.....
Robert W. Field
Haslam and Dewey Professor of Chemistry
Chair, Departmental Committee on Graduate Students

Single-cell response to perturbations across biological scales: single organ, organ system and phenotypic individuals

by

Kellie Elizabeth Kolb

Submitted to the Department of Chemistry
on May 10, 2019 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Chemistry

Signatures

Signature redacted

.....

Alex K. Shalek
Pfizer-Laubach Career Development Assistant Professor of Chemistry
Thesis Supervisor

Signature redacted

.....

Matthew D. Shoulders
Whitehead Career Development Associate Professor
Thesis Chair

Signature redacted

.....

Shiv Pillai
Professor of Medicine
Thesis Committee member

Single-cell response to perturbations across biological scales: single organ, organ system and phenotypic individuals

by

Kellie Elizabeth Kolb

Submitted to the Department of Chemistry
on May 10, 2019 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Chemistry

Abstract

The biological processes that sustain a complex organism require the orchestrated dynamics of complex cellular ensembles. Several vital systems – such as the immune system, the digestive system and more – must process internal and external signals to maintain functional homeostasis in response to perturbations at the systems-level. To further understand how groups of cells collectively respond to perturbations, we have applied single-cell RNA-sequencing and complementary techniques to explore cellular behaviors within complex systems at multiple relevant biological scales: from within a single organ, to an organ system, to across several human individuals with differing genetic backgrounds linked by a shared phenotype.

More specifically, at the level of the organ, we have explored acute injury responses in the liver. We have identified and described a new compensatory phase of the liver response to injury, in which surviving hepatocytes upregulate their expression of critical liver function genes to maintain overall organ function. Next, we extended our approach from a focus on an acute injury targeting a single organ to exploring chronic damage resulting from a long-term high fat diet across multiple gastrointestinal and immune compartments. Our analysis revealed molecular pathways and changes in stem gene expression which may contribute to obesity-related disease. Finally, we characterized shared features across multiple unique human donors with a common phenotype, elite control of HIV-1. We identified and validated a subset of highly functional dendritic cells, and developed broadly applicable computational approaches to identify reproducible responses across donors and to nominate candidate targets for rationally modulating the system. Overall, our work demonstrates the utility of single-cell RNA-sequencing for uncovering important cellular phenotypes that inform systems-level responses at any biological scale.

Acknowledgements

I would like to thank my thesis advisor, Alex Shalek, whose suggestions support and guidance have greatly improved this work and my training. I would also like to thank my committee members, Matt Shoulders and Shiv Pillai; members of the Shalek Lab, especially Sanjay Prakadan, Ben Meade, Jose Ordovas-Montanes, Sam Kazer, Brittany Goods and Sarah Nyquist. Finally, I am grateful to my external collaborators Chad Walesky, Carolyn Winston, Jake Henderson, Wolfram Goessling, Miyeko Mana, Omer Yilmaz, Enrique Martin-Gayo, Xu Yu, Michael Cole, and Nir Yosef. I have really enjoyed the opportunity to be involved in highly collaborative projects and work with many excellent scientists.

Table of Contents

Lay Summary

Complex organisms, such has humans and mice, are made up of trillions of cells. These cells can be further categorized into many types and subtypes (immune cells, intestine cells etc.) which work together to perform the critical functions which support life. Additionally, these cells must respond to challenges, such as infection or injury, and work together to maintain the function the organs and systems that they make up in spite of stressors. To better understand how individual cell responses contribute to larger system or organ response to changes, we use single-cell RNA-sequencing, which tells us which genes a particular cell is activating. We have profiled responses of single cells from a single organ, the liver; from multiple related digestive organs; and immune cells from people with highly effective immune responses to infection.

In the liver, we explore how liver cells respond to liver injury. The liver has extraordinary regenerative ability, unlike any other organ. The liver is able to fully regenerate itself to 100% its original size, even after surgical removal of 70% or more of the original liver mass. To learn more about liver regeneration, we sequenced liver cells after liver injury in a mouse model. We used two liver injury models: the partial hepatectomy, surgical removal of 70% of the liver mass; and acetaminophen (the active ingredient in Tylenol) overdose, in which toxicity from the excess medicine damages cells in specific areas of the liver. We learned that shortly following liver injury, the remaining liver cells increase their functional output to make up for the work that was previously done by the liver cells that were killed in the injury. In doing so, the overall function of the liver as a whole stays about the same as it was before the injury. About 30-36 hours after injury, the cells in the liver divide to return the liver to its original size. We found that the liver function output of the individual liver cells that divide is not as high as the liver cells that are not actively dividing. This suggests a division of labor between hepatocytes that increase their output to maintain liver organ function, and hepatocytes that direct their output toward cell division to restore liver size. Finally, we identify proteins secreted from macrophages, a type of immune cell in the liver, which serve to support the liver cell responses.

We then expand our scope from exploring responses within a single organ to profiling multiple digestive organs which are all affected by a particular challenge: long term high fat diet. Obesity is linked to increased risk for many types of disease, including fatty liver and liver cirrhosis, and inflammation and cancer in the liver and intestines. We maintained mice on high fat diet which induces obesity for six months, then sequenced cells from the liver, small intestine, large intestine and immune sites. The high fat diet livers and intestines contained more immune cells and inflammation, which is known contribute to development of disease in these organs. We identify molecules which activate pathways in the high fat diet mice that may trigger changes that lead to disease. In the liver, the population of cells which activate stem cell programs is much higher than in the control diet. These stem-like cells are primed to grow and may develop into cancer over time. Additional work may compare immune cells in the liver, intestines and blood and explore whether these cells may travel between these different areas and facilitate communication between organs.

Finally, we identify and characterize a subgroup of immune cells that contributes to a highly effective immune response. In a tiny minority (~0.5%) of people infected with HIV known as elite controllers, the immune system is able to control the virus, maintaining undetectable levels of virus in the blood and preventing progression to AIDS, even without antiviral drugs. By learning more about how the immune system is able to control virus in these rare cases, we may find avenues to develop new treatments or vaccines. We sequenced a particular type of immune cell,

known as dendritic cells, from elite controllers. We found that a subgroup of these cells activated many antiviral pathways. We then devised a way to collect this subgroup of highly functional dendritic cells to learn more about them. The subgroup of highly functional dendritic cells was better able to stimulate other immune cells to multiply and fight viral infection than other dendritic cells not from this subgroup. We found that the highly functional dendritic cells were more abundant in elite controllers than donors who were not elite controllers. We identified a way to stimulate dendritic cells in the lab to make more of the highly functional dendritic cells. Our methods for identifying ways to stimulate cells to be more functional may be therapeutically useful.

Overall, our work demonstrates the utility of single-cell RNA-sequencing for uncovering important cellular behaviors that contribute to organ- and systems-level responses.

Statement of Contribution

The work presented here was conducted as part of highly collaborative projects in which I worked closely with several other scientists. Here, I specify my specific contribution to each project. For the work in Chapter 3, I did all of the sequencing data analysis and most of the library preparation. I worked with others on experimental design, and preparation of figures and manuscript. I also helped with smFISH and PCNA staining. For the work in Chapter 4, I did all of the sequencing data analysis, most of the library preparation, and some of the experimental design. For the work in Chapter 5, I did all of the library preparation and TLR stimulations. I also helped with sorting, data interpretation and manuscript and figure preparation. For the appendix work, I did the library preparation and data analysis.

Chapter 1: Introduction

Proper function of the many biological processes which sustain a complex organism requires the orchestrated dynamics of complex cellular ensembles. The ability to maintain these functions is founded upon interactions between the many cell types that comprise vital organs and systems and their coordinated responses to perturbations. In some cases, these cellular responses to stimuli may be beneficial (e.g. productive immune responses to pathogens or cancer; recovery from injury), while in other cases, a response may be inappropriate or go unchecked, ultimately becoming detrimental to the health of the organism (e.g. development of cancer resistance to drugs; autoimmune responses). Rapid development of single-cell RNA sequencing technologies enabling higher throughputs for exploration of many thousands of cells has facilitated advancements in our understanding of cellular responses. While much wok has been done to map the cell populations in tissues, organs and even whole organisms in health and disease and to characterize cellular responses to wide variety of stimuli, much more work remains to be done. To further understand how groups of cells respond to perturbations, we profile cell populations at the single cell level to explore how behaviors of distinct subgroups of cells within a heterogeneous mixture functionally contribute to the system-level response. In the following chapters, we will cellular explore responses within systems of increasing biological scale: acute injury in a single organ, the liver; chronic metabolic stress across an organ system, multiple gastrointestinal organs and immune compartments; and across multiple human individuals with differing genetic backgrounds and a shared phenotype, elite control of HIV.

1.1 Opportunities in single-cell RNA sequencing

Over the past several decades, substantial work has been done to catalog the cell types, states, and interactions that inform systems-level response behaviors[1-6]. However, more recent studies have shown that even seemingly identical cell populations can exhibit significant and functionally important heterogeneities[7-11]. While this degree of diversity challenges our understanding of how systems-level responses are structured, it also presents new opportunities to deepen our understanding of cellular responses to stimuli with an eye toward realizing strategies to modulate cellular composition and cell interactions toward a more desirable overall response.

Rapid development of scRNA-Seq approaches over the last decade has positioned this technique to make major contributions in advancing our understanding of cell types and responses. Other single cell techniques, such as FACS and smFISH, are limited in the number of genes or proteins that can be profiled in a single experiment. Population methods, such as bulk RNA Seq, may mask signals from functionally distinct subpopulations even within a seemingly homogenous population. While many techniques can only find what the experiment was designed to specifically look for, unbiased scRNA-Seq can lend surprising new insights. scRNA-Seq approaches have made possible, for the first time, an unbiased view of expression of all mRNAs over large numbers of individual cells. This type of data enables identification of unique subpopulations of cells within the larger population and unexpected transcriptional responses. Further, by examining each single cell individually, we can uncover important, novel subtypes and response groups of cells. This opens up exciting new opportunities to explore the cellular make-up of organs and organisms, reviewed in Chapter 2, and the responses of different cell groups to a perturbation, which will be the main focus of this work. In deepening our understanding of the cellular behaviors

that constitute the overall response to a stimulation, such as infection or injury, and the molecular drivers of these responses, we will develop a more comprehensive picture of how organisms respond to stressors and rationally identify targets to modulate this response for therapeutic effect.

1.2 Responses across biological scale

Profiling the relevant cells is essential to understanding cellular responses to perturbations, but the relevant scale on which these responses occur varies for different systems and stimulations. In studies regarding highly localized responses the relevant biological scale may be on the order of a single tissue or organ, while other types of perturbations may impart a larger systemic effect spanning multiple organs, necessitating profiling cells over a greater biological scale, such as a system of related organs.

In work on clinical human samples, it is often necessary to consider an even larger biological scale: characterizing a phenotype which spans across many unique individuals who are also affected by many other unrelated underlying factors. Compared to model organisms, variation in human donors is highly uncontrolled – age, sex, race/ethnicity, socioeconomic staus, diet, lifestyle, BMI, other health conditions, (co)infections and genetic background can all significantly impact observations in human cohorts. While it is possible to control for some of this variation by carefully selecting whom to enroll in a study, limitations on what patients and samples become available for research mean that screening on more than a few variables will likely not leave enough participants eligible. Further, in humans, access to cells of interest presents a significant challenge. These studies must often make use of whatever samples are easily obtained (e.g. peripheral blood) or clinically indicated biopsies and resections, while work with animal models

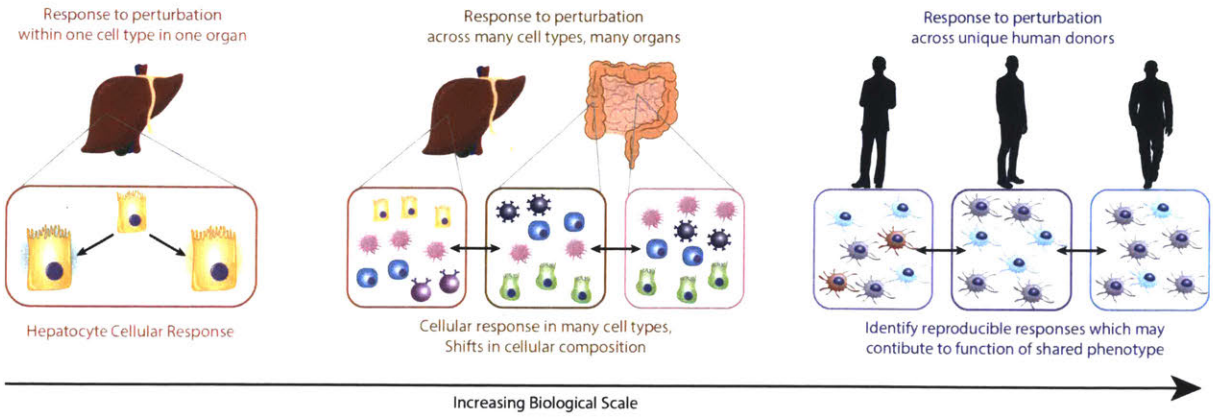

Figure 1-1 | Increasing biological scale: Single organ, organ system, multiple genetically unique individuals.
Single cells respond to perturbations, generating responses which must be assessed at the relevant biological scale. Hepatocytes respond to acute injury in the liver (left). Long-term consumption of a high fat diet can damage many gastrointestinal organs. We explore the responses of many cell types from many organs and compare high fat diet to control diet conditions (center). Dendritic cells from HIV-1 elite controllers (ECs) respond to viral stimulation *in* vitro. To more fully understand how DC responses may contribute to EC phenotype, we profile DCs from multiple ECs and identify reproducible responses, which are characteristic of the phenotype rather than unique to an individual donor(right).

can readily access cells from any compartment using whatever techniques are available. When using clonal model organisms under tightly controlled conditions, each organism under a given treatment condition may be considered a biological replicate; however, in human studies many uncontrolled factors can contribute to differences even between two relatively similar human donors. Despite these difficulties, the direct relevance of human samples sustains their continued appeal. To contend with the challenges associated with variation across multiple human donors with different genetic backgrounds and health histories, we can use scRNA-Seq data analysis to identify shared, reproducible responses that contribute to a group of donors' unifying phenotype.

1.3 Contributions of this work

As appropriate for the system and perturbation of study, we can apply scRNA-Seq methods to explore cellular responses at varied biological scales. In this work, we explore responses across different biological scales: from responses largely restricted to one relevant organ, to responses characterized across human individuals with different genetic backgrounds, histories, lifestyles, united by a shared phenotype of interest. With the application of scRNA-Seq methods and a suite of complementary validation techniques, we explore the responses of groups of single cells to acute, chronic perturbations and phenotypes, and how these cellular behaviors contribute to the overall systemic response.

In the next chapter, we provide background on the single cell sequencing field. We describe rapid development and scaling of scRNA-Seq over the last decade, and the benefits and drawbacks of various techniques. Development of higher-throughput methods have made evermore ambitious projects possible. We summarize the major efforts in the field to build "cell atlases" to catalog all cell types present at baseline and to characterize cellular responses to perturbations. We also explore the limitations and potential pitfalls of scRNA-Seq experiments and data analysis and approaches to mitigate these concerns.

In Chapters 3, 4 and 5, we apply scRNA-Seq to address biological questions, first directing our efforts to a localized, acute response – acute liver injury in Chapter 3. The liver possesses a fascinating, unparalleled ability to regenerate itself following injury. To better understand this ability, we apply scRNA-Seq methods to characterize responses to classic injury models within the liver organ at greater resolution and scope than previously achieved. We compare and contrast compensatory behaviors of hepatocytes in a toxin-induced and surgically-induced injury model, uncovering many shared features. In both toxin- and surgically-induced injury models, hepatocyte survivors upregulate liver function gene expression to compensate for functional

output from hepatocytes lost to injury. Additionally, we characterize the proliferating hepatocytes, which contribute to cellular and organ recovery. Our work contributes new insight into how a particular cell type, the hepatocyte, responds to and recovers from an injury perturbation.

In Chapter 6, we expand the scope of our work to more cell types and more organs to study the effect of a high fat diet on the gastrointestinal and immune system. In contrast to the acute liver injury study, in which cells respond to injury in a beneficial, restorative way, cellular responses to the chronic insult of a high fat diet are detrimental to overall system function and organism health and survival. Mice on a high fat diet can spontaneously develop liver problems – fatty liver, steatohepatitis, cirrhosis and hepatocellular carcinoma – and intestinal problems including inflammation and cancer[12]. To better understand these responses and to uncover potential molecular drivers, we apply scRNA-Seq to multiple compartments involved in responding to high fat diet: liver hepatocytes, liver non-parenchymal cells, proximal small intestine, distal small intestine, colon, spleen, bone marrow, and peripheral blood. In doing so, we identify cell circuits and pathways involved in high fat diet response, and identify candidate targets for therapeutic interventions.

Next, we extend our work from model organisms to humans in search of key cellular and molecular responses that contribute to a shared phenotype across different donors, thus expanding our biological scale to span genetically distinct individuals. We study the immune response to virus in dendritic cells from HIV elite controllers (ECs), a rare subset of the population which is able to control HIV infection without antiretroviral therapy. By increasing our understanding of how the EC immune system is able to control the virus, we may realize pathways to target for HIV therapies or vaccines. Because humans are necessarily less controlled than mice, with different genetic backgrounds and life histories, we must focus on shared immune responses of interest, and develop ways to computationally extract these shared features of

response which are characteristic of the phenotype rather than donor-to-donor variation. We develop a broadly applicable rational framework for identifying shared responses, isolating cell subgroups of interest, and nominating candidate targets for system modulation.

In the final chapter, we explore the possibilities for future work. Future studies will expand upon this work and take on new challenges. Our work described here has made contributions to the field of liver biology, deepened our understanding of the effects of a high fat diet, developed methods for identifying uniting characteristics within a phenotype, and applied scRNA-Seq methods to new problems. With modern high-throughput scRNA-Seq techniques, future studies will be well-equipped to expand the biological scale of their work to explore all relevant tissues, as we have done in Chapter 4, rather than focusing on a specific site out of method limitations. While all of the data presented in Chapters 2, 3 and 4 was collected from male mice or male human donors, females are known to have differing susceptibility to many types of disease, especially in the liver[13]. Future studies could build complementary female datasets to explore the molecular underpinning of these differences, and Appendix A summarizes some pilot work in this direction. Evolving computational approaches for data analysis, like the one put forth here in Chapter 5 will serve to identify meaningful, reproducible responses across donors and nominate targets for rational modulation of systems in future work. To address emerging challenges presented in datasets featuring more comprehensive sampling, new ways of integrating information over multiple tissues and identifying cross-talk signals must be realized. In another vein, future use of cross-species studies to combine the ease and control of mouse experiments with the relevance of clinical human samples may yield powerful new results. For example, as a follow-up to the high fat diet work presented in Chapter 4, a disease trajectory could be built across multiple time points, capturing all phases of chronic disease development in a mouse system while clinical human data could be mapped onto it to confirm the relevance of mouse data and better contextualize human data. More computational tools will be required to effectively

engage in making these cross-species comparisons, accounting for conserved and non-conserved elements of key genes, cell types, cell behaviors and pathways.

Single-cell RNA sequencing methods have made great strides in increasing throughput while analysis methods have developed to handle larger datasets and new problems over the last decade. The scRNA-Seq field has made and received significant contributions to and from many wide range of biological, technology development and computational fields as it has evolved. Numerous studies have shown the utility and power of including scRNA-Seq in their experimental design[14-16]. Now we can leverage these powerful techniques to learn more about many diverse systems than was ever possible before.

1.4 Introduction References

1 Allman, D. & Pillai, S. Peripheral B cell subsets. *Curr Opin Immunol* **20**, 149-157, doi:10.1016/j.coi.2008.03.014 (2008).

2 Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science* **332**, 677-678, doi:10.1126/science.1206351 (2011).

3 Kanno, Y., Vahedi, G., Hirahara, K., Singleton, K. & O'Shea, J. J. Transcriptional and epigenetic control of T helper cell specification: molecular mechanisms underlying commitment and plasticity. *Annu Rev Immunol* **30**, 707-731, doi:10.1146/annurev-immunol-020711-075058 (2012).

4 Merad, M., Sathe, P., Helft, J., Miller, J. & Mortha, A. The dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annu Rev Immunol* **31**, 563-604, doi:10.1146/annurev-immunol-020711-074950 (2013).

5 Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nat Immunol* **16**, 343-353, doi:10.1038/ni.3123 (2015).

6 Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663-1677, doi:10.1016/j.cell.2015.11.013 (2015).

7 Cohen, A. A. *et al.* Dynamic proteomics of individual cancer cells in response to a drug. *Science* **322**, 1511-1516, doi:10.1126/science.1160165 (2008).

8 Feinerman, O. *et al.* Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response. *Mol Syst Biol* **6**, 437, doi:10.1038/msb.2010.90 (2010).

9 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

10 Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369, doi:10.1038/nature13437 (2014).

11 Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496**, 461-468, doi:10.1038/nature11981 (2013).

12 Beyaz, S. *et al.* High-fat diet enhances stemness and tumorigenicity of intestinal progenitors. *Nature* **531**, 53-58, doi:10.1038/nature17173 (2016).

13 Biswas, S. & Ghose, S. Divergent impact of gender in advancement of liver injuries, diseases, and carcinogenesis. *Front Biosci (Schol Ed)* **10**, 65-100 (2018).

14 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

15 Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333-339, doi:10.1038/nature24489 (2017).

16 Mead, B. E. *et al.* Harnessing single-cell genomics to improve the physiological fidelity of organoid-derived cell types. *BMC Biol* **16**, 62, doi:10.1186/s12915-018-0527-2 (2018).

Chapter 2: Developments in single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-Seq) has enabled exciting new insights into cellular diversity and has demonstrated that functional diversity, even within seemingly homogenous cell populations, can drive responses to perturbations. Technology for scRNA-Seq has developed rapidly over the last decade, bringing about amazing new opportunities and challenges. In this chapter we consider the utility of single-cell RNA sequencing and the development and rapid scaling of scRNA-Seq methods, leading to the massively parallel, whole transcriptome methods popular today and detail the factors to weigh when selecting which of many current scRNA-Seq methods to use in a new experiment. We highlight some of the important work to build cell atlases as valuable maps of the cellular landscape. Additionally, we discuss how scRNA-Seq can be applied to understand cellular responses to perturbations, highlighting a few studies which take this approach. Finally, we discuss the limitations of scRNA-Seq methods and data analysis and some approaches to address these shortcomings. While far from exhaustive, this brief review provides an overview of some of the recent contributions and current work in the scRNA-Seq field and discusses some of the exciting questions and challenges that remain.

2.1 Utility of single-cell RNA sequencing in profiling systems-level response behaviors

Proper system function requires coordinated behaviors between many different cells. Over the past several decades, substantial strides have been made in cataloging the cell types and interactions that drive these behaviors. However, recent work has shown that even seemingly identical cells can exhibit significant heterogeneity with important functional consequences, challenging our current cell classification schemes as well as our understanding of how ensemble dynamics are truly structured[8]. Clearly, identifying the basic cell subsets and responses and their molecular drivers is essential for understanding and therapeutically manipulating how cellular systems responds to stimuli.

Fortuitously, the emergence of single-cell RNA-sequencing (scRNA-Seq) now enables profiling of the individual cells comprising a system of interest at a resolution and scope not previously possible, facilitating a more detailed picture of how individual cells and groups of cell contribute to function. scRNA-Seq techniques allows us to profile, genome-wide, the transcriptomes of individual cells and identify, from first principles, cell types and states that population methods may mask. In doing so, we may identify previously unappreciated subgroups of cells and uncover their functional contributions to the overall system. Many other techniques require the experimentalist to select genes or proteins to profile individually, thereby limiting the scope of what can be discovered to what the experiment is initially designed to profile. In contrast, scRNA-Seq provides an unbiased look at all the mRNAs expressed within a single cell, enabling us to discover surprising expression programs and transcriptional responses we would not have thought to explore based only on prior knowledge. Moreover, scRNA-Seq can be used to examine complex mixtures of cells and some techniques are compatible with low sample inputs, making scRNA-Seq ideal for studying isolates which contain complex mixtures of cell types or rare cell types (e.g. immune populations and rare proliferative/stem cells). As such, scRNA-Seq gives us

an exciting new lens through which to catalog the cellular composition of organs, organ systems and organisms, and to explore system function and dysfunction in health and disease.

We want not only to identify functionally distinct subpopulations of cells, but also to discover molecular pathways which drive the arising of cell groups and responses. In addition to identifying unique cell groups and the gene expression patterns that define them, further exploration of scRNA-Seq data through Gene Set (Enrichment) Analysis (GSA) can make use of existing biological knowledge in analysis of scRNA-Seq expression data[9]. The comparative analysis between a new scRNA-Seq data set and existing publicly available datasets can lead to deeper insights. For example, GSA can identify pathways and upstream drivers which contribute to the cellular circuits activated under a particular condition. Several programs, such as Qiagen's Ingenuity Pathway Analysis (IPA), and the Broad Institute's Gene Set Enrichment Analysis (GSEA) calculate overlaps between a set of differentially expressed genes that define a cell group of interest in a scRNA-Seq analysis and a collection of reference gene sets. A significant degree of overlap between the differentially expressed gene set and the reference gene set indicates that similar pathways, responses or upstream drivers may be active in both the cell group of interest in the sequencing data set and the reference experiment[9]. For example, a significant enrichment between a particular group of immune cells and a reference gene set of genes activated by interferon gamma would suggest that that particular group of immune cells may be responding to interferon gamma. A large collection of reference gene sets is publicly available in the Broad's Molecular Signatures Database (MSigDB)[10]. These and other reference gene sets may come from gene expression patterns observed in previous studies, curation from the literature, gene ontology, or user-generated sets to specifically interrogate responses to a collection of stimuli in a particular cell type[9]. Beyond using reference gene sets to identify pathways and responses at play in a group of cells, by focusing on upstream drivers of these pathways, we can nominate

potential routes to rational modulation of the system. By drugging to inhibit or stimulating to activate the upstream regulator, we may respectively up- or down-regulate a pathway that contributes to a given cellular response, thereby engineering the system to behave more in the direction of our choosing. In identifying and validating candidates for system modulation, we may identify targets for therapeutic intervention.

2.2 Increasing the throughput of scRNA-Seq techniques

Application of scRNA-Seq to many systems has led to many new exciting insights, and, in most cases, more cells means a more detailed, comprehensive view of the system and more statistical power to draw conclusions. In response to need for more cells and better data, scaling of scRNA-Seq has risen sharply over the last decade[7] (**Figure 2-1**). The throughput and scope of scRNA-Seq methods has increased exponentially from only a few genes in a single cell in the earliest experiments[11], to today's massively parallel methods capable of profiling the whole transcriptomes of thousands of cells in parallel[1,12], enabling ever more detailed surveys of the functional heterogeneity of complex populations of cells.

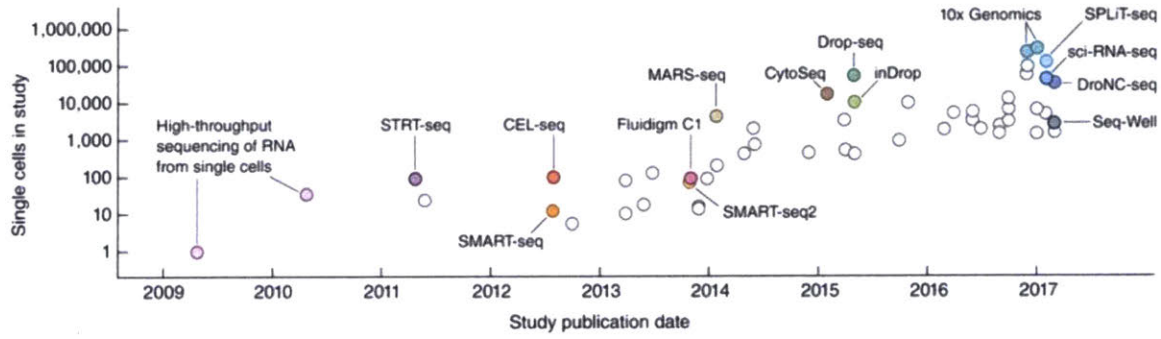

Figure 2-1 | Increasing throughput of scRNA-Seq techniques
Adapted from Svensson *et.al.* 2018[7]
Number of cells in selected scRNA-Seq publications versus time. As new higher-throughput protocols have been developed, studies sequencing increasing numbers of cells have become possible. Studies introducing key technologies are indicated.

To increase the gene coverage of single cell sequencing techniques beyond the handful of genes profiled in the earliest work[11], Tang *et. al.* applied next generation sequencing technology to profiling the transcriptomes of a single cell and obtained sequence information for 75% more genes than the microarray techniques popular at the time[13]. Building upon use of next generation sequencing to capture large numbers of genes, subsequent scRNA-Seq technology development sought improvements over the manual cell capture method used in earlier techniques, with the goal of improving capture efficiency and overall cell number. Plate-based methods were designed to generate sequencing libraries from single cells isolated by fluorescence-activated cell sorting (FACS) cells into multiwell plates. Flow panels for sorting cell types, especially immune subsets, had already been established for isolation of many targeted cell types, making plate-based methods an appealing scRNA-Seq technique for many biologists interested in further defining functional heterogeneity within the cell types already distinguished with FACS methods.

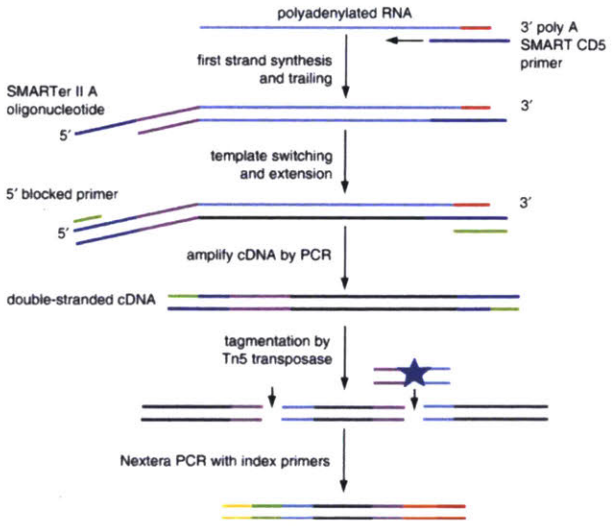

Figure 2-2 | SMART-Seq2 Schematic
Adapted from Trombetta *et.al.*[2-6]
SMART primer captures poly-A mRNAs and primes RT. RT enzyme adds tailing sequence by template switching and extension, 5' primer binds. Amplify cDNA by PCR, tagment by transposase (Nextera). Amplify with primers. Confirm quality and measure concentration for sequencing.

One widely adopted plate-based method, SMART-Seq2, initiates reverse transcription with a poly-T oligonucleotide designed to capture poly-A mRNAs[2,14](**Figure 2-2**). This oligonucleotide also contains a universal priming sequence, the SMART primer, which will later serve as a

handle for PCR. This method minimizes 3' bias by using a template-switching reaction at the 5' end to anneal the SMART primer at that end for PCR amplification such that only fully reverse transcribed mRNAs can be amplified. It also uses the same SMART priming sequence for both the forward and reverse reaction PCR directions, minimizing PCR artifacts from different primer affinities and hairpinning undesirable short templates. Other plate-based methods have also been developed, such as MARS-Seq which uses an *in vitro* transcription approach[15].

Commercial development of scRNA-Seq methods by Fluidigm produced the C1 system in 2012, which uses microfluidics to capture up to 96 single cells in tiny chambers. However, in many cases only a fraction of the chambers will successfully capture a cell[7]. Subsequent techniques further improved upon applications of microfluidics to capture cells.

In the In-Drop and Drop-Seq methods, individual cells are isolated in tiny droplets. A stream containing mRNA-capture beads and reagents for RT is merged with another liquid stream containing a single cell suspension[12,16]. Microfluidic control separates this stream into reverse emulsion droplets in oil, capturing one bead and one cell per droplet by poisson loading (most droplets will be empty and many others will contain only a bead or only a cell). The mRNA capture beads are functionalized with many oligos containing poly-T primers on their surface to bind the poly adenylated mRNAs. The bead oligos also contain a cell barcode, which is unique to each bead, for separating reads in the sequencing data output into the cells from which they came, and unique molecular identifier (UMI), which is unique to each individual oligo on the bead, to computationally collapse PCR amplified cDNAs back to the original one RNA molecule from which they amplified. These droplet-based techniques scaled scRNA-Seq to reach the ability to profile thousands of single cells in parallel. Due to the necessity of having at most one bead and one cell per droplet, both beads and cells must be Poisson loaded, resulting in large dead volumes and

many cells captured without beads lost, making this technique appropriate only for experiments with large numbers of cells available.

To iterate upon the bead-based, 3'-priming strategy of Drop-Seq, later techniques were developed to improve cell capture efficiency, making them suitable for lower input samples. Microwell-based methods, such as Microwell-Seq[17] and Seq-Well[1], pair individual cells and beads by settling them into tiny wells on a microarray device. By sizing the beads and wells such that only one bead will physically fit in each well, these techniques do not require Poisson loading of beads and can achieve bead loading in nearly every well (**Figure 2-3**). Cells still need to be Poisson loaded to minimize doublets, but with bead loading efficiency around 90%, few cells are lost to wells with no beads. In a typical Seq-Well experiment, around 15,000 cells are loaded, though the technique can also be run with less, and generally high-quality data will be obtained for over 1,000 cells, greatly improving cell capture efficiency and making this technique suitable for low-input samples[1]. Seq-Well further improves upon earlier microwell-based methods by applying a semi-permeable membrane which allows for buffer exchange, but prevents transfer of cells, beads and large molecules, such as mRNAs, between adjacent wells[1].

Commercialization of high-throughput techniques has already begun. 10X Genomics' Chromium platform, a commercialized version of In-Drop, has gained wide popularity. The ease of use of this system makes it accessible even to those without extensive experience in single-cell techniques and the ability to run up to eight samples in parallel increases experimental efficiency by multiplexing several experimental conditions into one run.

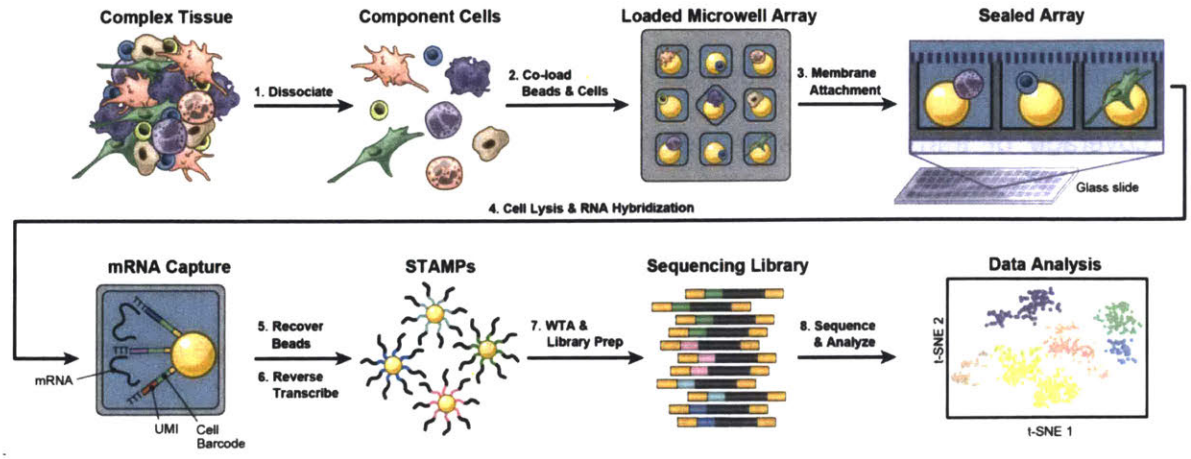

Figure 2-3 | Seq-Well Schematic
Adapted from Gierahn *et. al.* [1]
Complex tissue dissociated into single cell suspension. Beads and cells loaded onto microwell array and allowed to settle by gravity. Semipermeable membrane applied, cells lysed. Capture beads are functionalized with oligos containing a poly-T capture sequence, Unique Molecular Identifier (UMI) and cell barcode. mRNAs hybridize to bead oligos, beads recovered from microwell array device, reverse transcribed, and undergo whole transcriptome amplification (WTA). Sequencing libraries prepared from WTA by Nextera. Libraries sequenced and analyzed.

2.3 Selection of most appropriate scRNA-Seq method

With so many scRNA-Seq methods available it is important to consider which method is best suited for a particular experimental question. The microwell- and droplet-based methods both offer high-throughput processing and UMI barcoding for quantitative sequencing. Microwell-based methods are generally better suited than droplet-based methods for samples with limited cell numbers, as they do not require Poisson loading of cells. Cost is also an important factor, with high-throughput methods generally delivering a lower cost per cell. Even within high-throughput methods, cost can vary considerably. Seq-Well, a microwell-based method, processes a sample at much lower average cost per cell than the popular 10X Genomics Chromium platform; however, the Seq-Well processing is more labor-intensive.

While droplet- and microwell-based methods offer the obvious benefits of higher throughput and lower cost per cell, in some instances plate-based methods may be preferable. The high-throughput methods sacrifice coverage per cell for coverage of more cells and provide only 3'-end coverage of transcriptomes. In contrast, plate-based methods offer full transcriptome coverage and generally a higher sensitivity and overall coverage per cell. Full coverage may be desirable for some applications, such as calling splice variants. Some studies have made use of a combined approach, applying a high-throughput 3' method to sequence huge numbers of cells at relatively low coverage and cost per cell, and a FACS-enabled plate-based method to gain deeper information for select poorly-understood cell groups[18]. Additionally, in experiments requiring FACS to enrich for a targeted rare cell population, it may be preferable to sort these cells directly into plates rather than to further process them through the microfluidics or well loading steps of another method, particularly if the number of targeted cells available is very low (hundreds) or when the source material is very fragile.

The rapid development of scRNA-Seq tools has made many exciting studies possible. When designing a single cell study, it is imperative to carefully consider the merits of each technique and to select the one best suited the experimental questions and goals.

2.4 Mapping the cellular landscape: Cell Atlases

Thanks to extensive work in single cell genomics over the past several years to improve methods and survey cellular diversity, we now know more than ever about the cell types present within organs and some whole organisms at the single cell level. The creation of a comprehensive reference map – a "cell atlas" – of all the individual cells with a system has become a major pursuit of the single cell sequencing community. These cell atlases delineate the many types of cells

found in a particular tissue, organ, or organism, identify novel subgroups of cells with unique functions, and catalog marker genes for each group. This information deepens our knowledge of the system and serves as a valuable guide for other researchers needing to call cell types from their own sequencing data.

Many cell atlases, for mouse, human and other organisms, have been published in recent years, providing insight into the systems they explore and serving as valuable references for future work. Advent of high-throughput techniques has made feasible sequencing tens or even hundreds of thousands of single cells, rendering comprehensive profiling of entire systems of cells possible. In their 2015 Cell paper, Macoscko and colleagues introduce Drop-Seq and apply it to sequence and analyze the transcriptomes of over 44,000 mouse retinal cells, identifying 39 distinct populations and creating a cell atlas of known and candidate novel retinal cell types[12]. In the following years, many more cell atlases for other tissues and organs have been published. For example, Haber *et. al.* produced a cell atlas of the mouse small intestinal epithelium, featuring over 53,000 cells, characterizing previously unappreciated diversity, and uncovering new subsets in enteroendocrine and tuft cells [19].

Multiple studies have expanded the scope of the atlas to encompass whole organisms, including *C. elegans*[20] and mouse. Two whole mouse atlases were published in 2018: "Mapping the Mouse Cell Atlas by Microwell-Seq" covering 400,000 cells published in *Cell* by Han *et. al.* [21], and "Tabula Muris" published in *Nature* featuring over 100,000 cells and using a combination of high-throughput 3'-end droplet-based methods and full length transcript, flow-sorting enabled methods to obtain higher sensitivity data for selected cell types [18].

With the improved throughput of more recent scRNA-Seq techniques and high interest in atlas datasets, ambitious goals of identifying and describing increasingly extensive collections of cells at finer granularity have become attainable. The Human Cell Atlas, a coordinated global effort to create a collection of maps for every human cell, endeavors to "defin[e] the cellular basis for health and disease"[22]. Building the Human Cell Atlas parallels the Human Genome Project in many ways: its immense scope and collaborative approach and its potential to transform our understanding of biology and serve as a rich resource to facilitate future studies. The Human Cell Atlas will help to identify genes associated with disease, explore what cell types are present and where, uncover regulatory mechanisms which contribute to cell type development and function, and serve as a reference for future work[23,24].

2.4 Cellular responses to perturbations

To elucidate cellular contributions to systems level responses, we must evaluate how individual cells and groups of cells work together to respond to perturbations. Cell atlas work has made great strides in cataloguing the cells present in many systems, but typically focuses on surveying cell populations and diversity at baseline. Relevant atlases can serve as useful guides to interpreting data from experiments investigating cellular responses to stress or infection within a system. When responding to a stimulus, distinct cell groups and subgroups may respond in different ways, or not at all. In identifying cell types and characterizing the responses of functionally different groups of cells, we may begin to understand how the system-level response arises from the coordinated behaviors of all cell types involved.

scRNA-Seq methods enable us to investigate these cellular responses in greater detail and discover unexpected cell groups and responses. For example, early work in scRNA-Seq has shown that not all DCs respond identically to culture with LPS, despite originating from a seemingly homogenous population[5]. Avraham and colleagues used a combined FACS scRNA-Seq approach to investigate how variation in bacterial factors from invading *Salmonella* contribute to observed variation in individual macrophage responses to infection[25]. More recently, Kim *et. al.* used a combination of bulk and single cell sequencing approaches to investigate the evolution of chemoresistance in triple-negative breast cancer[26]. With scRNA-Seq, these studies and many others uncover contributions of previously unappreciated subgroups of responding cells.

2.5 Limitations of scRNA-Seq

While scRNA-Seq is an enormously useful tool, it also has limitations. Due to the very low input RNA material of a single cell (10-30 pg total RNA[27]), "drop-out" is a persistent challenge. Generally speaking, scRNA-Seq methods capture transcripts through a poly-T sequence on a priming oligonucleotide, which targets the poly-A tail of mRNA. This oligonucleotide then serves as a primer in the subsequent reverse transcription step. However, mRNA binding to poly-T oligonucleotides and reverse transcription reaction efficiencies are not 100%, leaving the possibility that some transcripts will not be captured or reverse transcribed, and, therefore, will ultimately be excluded from the resulting sequencing library. Since so little starting material exists in a single cell, missing even a few transcripts in library prep can impact the data, making the interpretation of a "zero" in the data ambiguous. It is difficult to know whether a zero in the data is biological (the gene was not expressed) or technical (the gene was expressed but not captured or not sequenced). Undersequencing a library can also contribute to dropout when sequencing

depth is insufficient to read all of the transcripts represented in the library. Sequencing to greater depth can recover more genes in undersequenced samples; however, increasing sequencing depth for thousands of single cells can be costly and is not always necessary to address the question at hand. Some computational data analysis techniques have attempted to address dropout and the sparsity of scRNA-Seq data matrices[28,29], though best-practices remain uncertain.

Dropout is especially problematic for lowly expressed genes, where a loss of one or two out of only two or three total transcripts greatly impacts the data (e.g. loss of one out of two transcripts results in a 50% change in expression quantification; or if all one or two transcripts are not captured, total loss of expression of the gene), while in more highly expressed genes, losing a few transcripts out of hundreds will have minimal impact on overall expression data. Additionally, transcripts which are physically short (e.g. interferon at 501nt[30], compared to 1-2 kb for typical transcripts) are often lost in the processing and purification steps of library preparation which are required to remove excess primer and generate a library of suitable quality for sequencing. As such, scRNA-Seq data is largely unreliable for lowly-expressed genes and genes with short transcripts.

To address challenges related to dropout, an analysis of scRNA-Seq data should characterize cell types, groups, subgroups and responses by use of gene sets, rather than individual genes, whenever possible. By focusing on sets of genes that are expressed as a group on a particular cell type or work together to coordinate a response or up-regulate a pathway, a scRNA-Seq analysis buffers itself from the effects of dropout. A gene set may include a collection of genes which are all markers for a particular cell type (e.g. intestinal stem cell markers: *Lgr5, Ascl2, Slc12a2, Axin2, Olfm4,* and *Gkn3*[19]), members of a transcriptional network of genes with a shared

upstream regulator which is activated in response to perturbation (e.g. interferon response genes) or any group of genes with expected co-expression in the targeted cell type or response. While the transcripts for a single gene may fall victim to dropout within a single cell, if a particular cell truly does belong to a given cell type or response group, it should express all or most of the genes included in the relevant gene set and at least some of these genes should be captured. When looking over the whole gene set, even cells with dropout data can be correctly identified by relying overall expression of coordinately expressed genes in the gene set.

scRNA-Seq captures an unbiased view of the transcriptional behavior of cells; however, usually it is the proteins which the mRNAs encode that actively drive the response behavior at the molecular level. While RNA and protein are usually roughly correlated, temporal differences in mRNA expression and translation, burst transcription kinetics, differences in RNA and protein half-lives, regulation of translation, and post-translational modifications can all contribute to drastically altering the functional output at the protein level from what was suggested in mRNA transcript space[31]. Since it is possible and relatively easy to sequence all mRNAs in a single cell by RNA sequencing, but not practical to sequence the full proteome across many single cells, scRNA-Seq techniques provide valuable insight, but the potential for signal regulation and modulation downstream of transcription must always be considered. Functional validations should be used as appropriate to confirm observations made from sequencing data.

While scRNA-Seq methods provide an exciting way to further understand cellular responses and dynamics in a system, it remains critical to validate discoveries with other techniques. Limitations of scRNA-Seq due to technical dropout, loss of shorter transcripts and potential differences in RNA and protein expression make further validation by other methods to more directly measure selected responses essential to draw strong conclusions. Despite limitations, scRNA-Seq can

provide great insights and unbiased, whole transcriptome data and can generate hypotheses in new and exciting directions. The broad scope of scRNA-Seq data enables generation and informs selection of hypotheses to further validate using more targeted approaches which are typically much more limited in scope. For example, GSA over differentially expressed genes from sequencing data may indicate that a particular pathway is upregulated during a cellular response and identify an upstream regulator. Targeted follow-up experiments may inhibit (by drugs or knockouts) or activate (by drug or ligand) the implicated upstream regulator, which should in turn modulate the downstream pathway and alter the response phenotype, validating the predicted contribution of the pathway from sequencing analysis. In another type of experiment, smFISH could be used to validate expression of a particular gene, expression patterns (e.g. co- or mutually exclusive-expression) of small groups of genes or physical spatial distribution of gene expression to corroborate gene expression patterns identified in scRNA-Seq[4,5].

A wide variety of validation techniques may be applied in cross-disciplinary studies to accompany and validate exciting scRNA-Seq data. This powerful combination of approaches yields new insights across many fields including infectious disease immunology, cancer, autoimmune conditions, acute injury and more. As scRNA-Seq becomes a more widely adopted technique, its contributions to our understanding the complex mixtures cell types and subtypes which contribute to organism function will continue to grow.

2.5 References

1 Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* **14**, 395-398, doi:10.1038/nmeth.4179 (2017).

2 Trombetta, J. J. *et al.* Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr Protoc Mol Biol* **107**, 4 22 21-17, doi:10.1002/0471142727.mb0422s107 (2014).

3 Feinerman, O. *et al.* Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response. *Mol Syst Biol* **6**, 437, doi:10.1038/msb.2010.90 (2010).

4 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

5 Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369, doi:10.1038/nature13437 (2014).

6 Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496**, 461-468, doi:10.1038/nature11981 (2013).

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

8 Cohen, A. A. *et al.* Dynamic proteomics of individual cancer cells in response to a drug. *Science* **322**, 1511-1516, doi:10.1126/science.1160165 (2008).

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

10 Institute, B. *GSEA MSigDB*, http://software.broadinstitute.org/gsea/msigdb (2019).

11 Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* **89**, 3010-3014 (1992).

12 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

13 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382, doi:10.1038/nmeth.1315 (2009).

14 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).

15 Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779, doi:10.1126/science.1247651 (2014).

16 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).

17 Yuan, J. & Sims, P. A. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Sci Rep* **6**, 33883, doi:10.1038/srep33883 (2016).

18 Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4 (2018).

19 Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333-339, doi:10.1038/nature24489 (2017).
20 Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667, doi:10.1126/science.aam8940 (2017).
21 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **173**, 1307, doi:10.1016/j.cell.2018.05.012 (2018).
22 Atlas, T. H. C. *The Human Cell Atlas*, 2019).
23 Consortium, T. H. C. A. The Human Cell Atlas White Paper. (2017).
24 Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451-453, doi:10.1038/550451a (2017).
25 Avraham, R. *et al.* Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. *Cell* **162**, 1309-1321, doi:10.1016/j.cell.2015.08.027 (2015).
26 Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893 e813, doi:10.1016/j.cell.2018.03.041 (2018).
27 Qiagen. *Resources*, https://www.qiagen.com/fr/resources/faq?id=06a192c2-e72d-42e8-9b40-3171e1eb4cb8&lang=en
28 van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e727, doi:10.1016/j.cell.2018.05.061 (2018).
29 Talwar, D., Mongia, A., Sengupta, D. & Majumdar, A. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci Rep* **8**, 16329, doi:10.1038/s41598-018-34688-x (2018).
30 Archive, E. N. *Coding: CAA26022.1 Homo sapiens (human) interferon precursor*, https://www.ebi.ac.uk/ena/data/view/CAA26022 (2019).
31 Genshaft, A. S. *et al.* Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol* **17**, 188, doi:10.1186/s13059-016-1045-6 (2016).

Chapter 3: Functional Compensation Precedes Recovery of Cell Mass Following Acute Liver Injury

This chapter is adapted from a manuscript in preparation:

Chad Walesky*, Kellie E. Kolb*, Carolyn Winston, Jake Henderson, Benjamin Kruft, Florian Mueller, Udayan Apte, Alex K. Shalek, Wolfram Goessling, "Functional Compensation Precedes Revovery of Cell Mass Folowing Acute Liver Injury," In preparation.

*Denotes equal authorship

The liver is responsible for many essential homeostatic functions, such as glucose homeostasis, protein and lipd metabolism, bile acid production, synthesis of serum proteins and toxin and xenobiotic metabolism, and is able to regenerate upon injury. While many factors have been identified that regulate the cellular proliferation the facilitates regeneration, how the liver maintains its vital functions preceding cellular recovery remains unknown. Here, we identify a new phase of functional compensation following acute liver injury prior to cell proliferation. Using single-cell RNA-sequencing and single-molecule fluorescent *in situ* hybridization techniques in two independent murine acute liver injury models, we discover up-regulated expression of injury response and core liver function genes, dependent on intact WNT/β-catenin signaling. We reveal that cell compensation and cell proliferation are inversely regulated, explaining the observed delay

in cell proliferation following injury. Our work describes a new mechanism by which the liver maintains essential physiological functions prior to the onset of cellular compensation and characterizes the hepatocytes that contribute to cellular recovery.

3.1 Background

The liver is a vital organ charged with a wide array of functions, including: homeostasis of glucose, protein, and lipid metabolism; production of bile; synthesis of critical serum proteins; and, metabolism of endogenous and xenobiotic toxins[1]. To accomplish these tasks, hepatocytes are organized into lobules across which blood flows from the portal vein and hepatic artery to the central vein establishing, in the process, oxygen and nutrient gradients. This, in turn, results in a corresponding zonal distribution and regulation of hepatocyte gene expression to support all of the liver's function; as a consequence, hepatocytes in the periportal zone have functions and gene expression profiles distinct from the midlobular and pericentral zones.

As a filter, the liver experiences frequent toxic insults resulting in cellular injury and death. Thus, the liver has evolved substantial regenerative capacity[2]. Extreme injury, however, can lead to decompensation, acute liver failure (ALF), and death[3]. Liver injury is frequently caused by selective toxins, such as acetaminophen (APAP), which results in zone-specific injury[4], and APAP overdose is the most common cause of ALF in the United States[5]. Another common cause of injury is surgical resection, which may be necessary to remove liver tumors[6]. Understanding the mechanisms governing liver regeneration in response to injury is crucial to facilitate novel treatment strategies for ALF.

To date, study of liver regeneration has singularly focused on the mechanisms underlying the re-establishment of lost cell mass through proliferation. In contrast to injury in other organs, such as

skin (~12 hrs), cell proliferation typically does not begin until approximately 24 hours and peaks around 30-48 hours post injury in mice[7,8]. It remains unclear how the organism survives the period immediately after losing massive amounts of functional liver tissue and why cellular recovery occurs relatively late.

Here, we investigated the functional compensatory responses of the liver during three key phases of liver regeneration (initiation phase, proliferation phase, and termination phase) in a resection (partial hepatectomy, PH) and a toxic (APAP) model of liver injury (**Figure 3-1A**). Utilizing a powerful combinatorial approach that couples Seq-Well, a platform for massively-parallel single-cell RNA-seq (scRNA-Seq) of clinical specimens ideally suited to fragile cells like hepatocytes, with single-molecule fluorescent in situ hybridization (smFISH), we define and validate transcriptional changes occurring during each of these three response phases[9]. We discover that, following injury, the remaining hepatocytes functionally compensate for lost functional liver mass preceding the peak of hepatic proliferation by increased transcriptional output of key hepatocyte genes. Importantly, hepatocytes demonstate an ability to alter their functional identity to maintain the expression of select genes despite injury-dependent loss of functional mass. By examining cell cycle signatures at different time points during the regenerative phase, we find that cycling cells do not participate in functional compensation to the same degree as non-cycling hepatocytes. We identify up-regulation of Wnt target genes in both cycling and non-cycling cells, and explore the contributions of Wnt/β-catenin to both proliferation and functional compensatory responses, demonstrating that hepatocyte functional compensation depends upon macrophage/Kupffer cell secreted Wnts. Overall, we identify and characterize a functional compensation phase in hepatocyte response to liver injury which preceeds cellular recovery, demonstrate that cycling hepatocytes do not functionally compensate to the same degree as noncycling hepatocytes and establish that macrophage Wnt secretion supports hepatocyte functional compensation.

Results

3.2 Transcriptional adaption after liver injury

To assess the global transcriptional responses of hepatocytes following acute liver injury, we used scRNA-seq (Seq-Well) to examine response dynamics in both the PH and APAP models, capturing the injury, regeneration, and termination phases of liver regeneration (**Figure 3-1B, C**). A total of 16,019 cells from 19 different experiments were profiled to an average sequencing depth of over 48,000 reads per cell (**Methods**). Subsequently, we filtered out immune and endothelial cell types and low quality cells from the dataset, retaining 10,762 high-quality hepatocytes on which we focus our subsequent analyses (**Methods**). Shared nearest neighbors clustering (SNN) visualized by t-Stochastic Neighbor Embedding (t-SNE) reveals hepatocyte populations that cluster by injury model and post-injury time point (**Figure 3-1D**). We observe baseline heterogeneity in untreated control samples, which form individual clusters by animal, differentiated largely by pheromone-related genes. The injury samples cluster together by time point, suggesting that individual livers become more similar to one another in their response to injury.

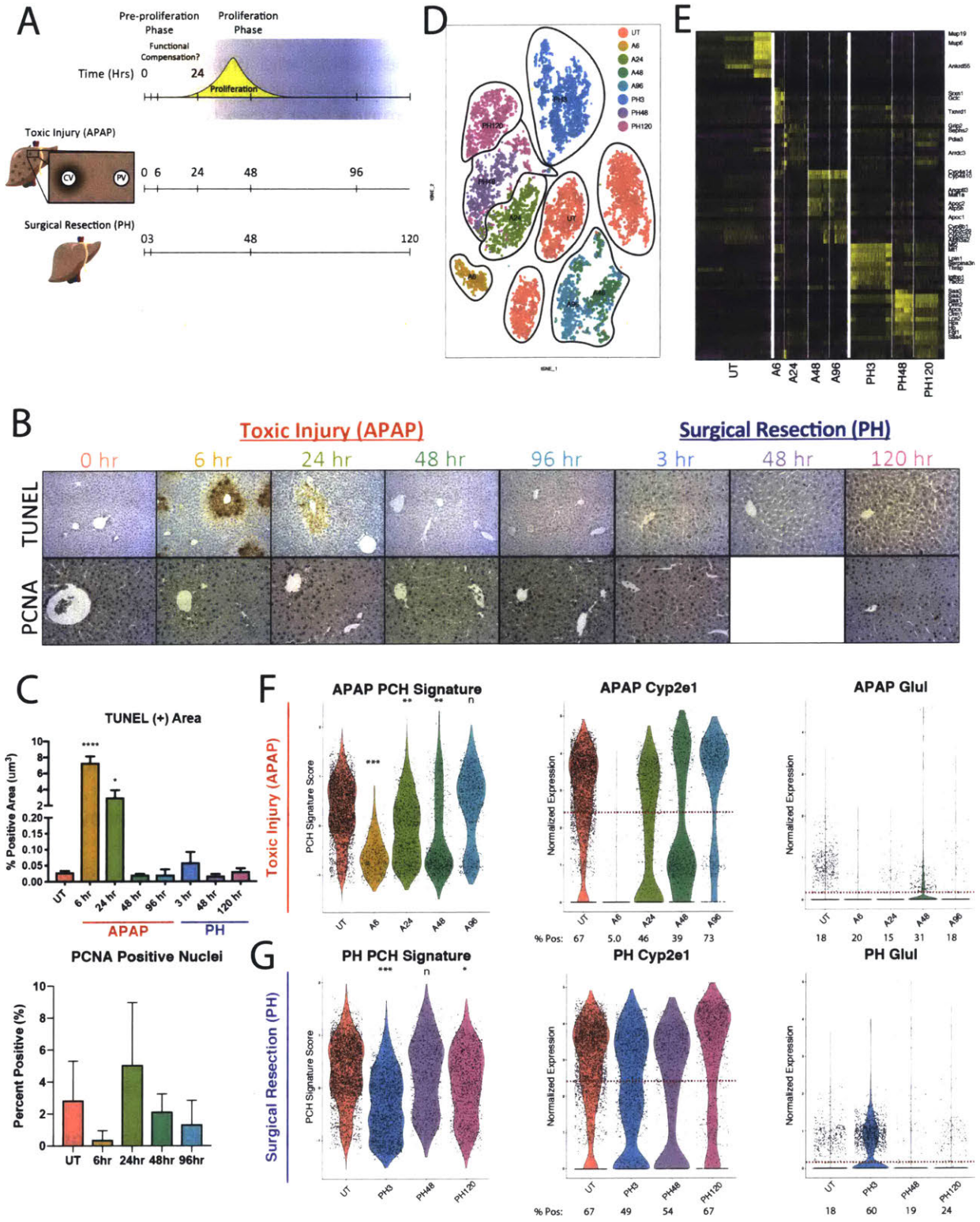

Figure 3-1 |Hepatocytes respond to toxic and surgical liver injuries.
Legend next page.

Figure 3-1 | Hepatocytes respond to toxic and surgical liver injuries.
Previous page.
(A) Time course depicting hypothesized functional compensation and proliferation phases of liver injury recovery (top). Samples collected from murine livers at multiple time points following APAP treatment (middle) or PH (bottom). **(B)** Staining of murine liver slices taken at vaired time points after liver injury showing apoptotic TUNEL positive (brown, top) and proliferative PCNA positive (brown, bottom) cells. **(C)** Bargraphs quanifying the total TUNEL and PCNA positive area. * $p < 0.1$; *** $p < 0.01$ **(D)** t-SNE of all high quality hepatocytes (**Methods**) in the scRNA-Seq data set. Cells are colored by treatment condition. SNN clusters outlined in black. **(E)** Heatmap of marker genes for all clusters loutlined in **D**. **(F,G)** Pericentral hepatocyte Signature Score (PCH Signature Score) (left). Violin plot of normalized expression of *Cyp2e1* (middle) and *Glul* (right); percent positive calculated as percentage of total cells in each condition above average normalized genes expression (dashed red line). Untreated (UT) and each post-treatment are plotted for APAP **(F)** and PH **(G)**. Effect size calculated by Cohen's d: $d < 0.2$ = n (negligible); $0.2 < d < 0.5$ = * (small); $0.5 < d < 0.8$ = ** (medium), $d > 0.8$ = *** (large).

To confirm clustering captures biological rather than technical variation, we performed differential expression to identify gene expression patterns unique to each cluster. We identify many genes related to liver function and response to injury and oxidative stress that define the clusters (**Figure 3-1D**), and technical gradients within rather than across clusters (nGene, nUMI, **Methods**). APAP injury results in typical pericentral necrosis at 6 hrs following APAP administration (hereafter A6; **Figure 3-1B,C**). Hepatocytes with pericentral hepatocyte signature (PCHSig) are absent at 6 hours post-treatment (**Figure 3-1F**, A6). At 24 hrs post APAP administration, however, the pericentral hepatocyte signature returns (**Figure 3-1F**, A24), while histology shows persistent necrosis in the pericentral area (**Figure 3-1B, C**, A24). This indicates that a previously existing cell population now expresses pericentral genes. Expression of pericentral genes *Cyp2e1,* responsible for metabolizing APAP, and *Glul*, which assimilates ammonia into glutamine, is highly restricted to the pericentral region. The frequency of *Cyp2e1*+ hepatocytes drops from 67% (Untreated, UT) to 5% (APAP 6 hour treatment, A6), returning back to 46% by 24 hrs (**Figure 3-1F**).

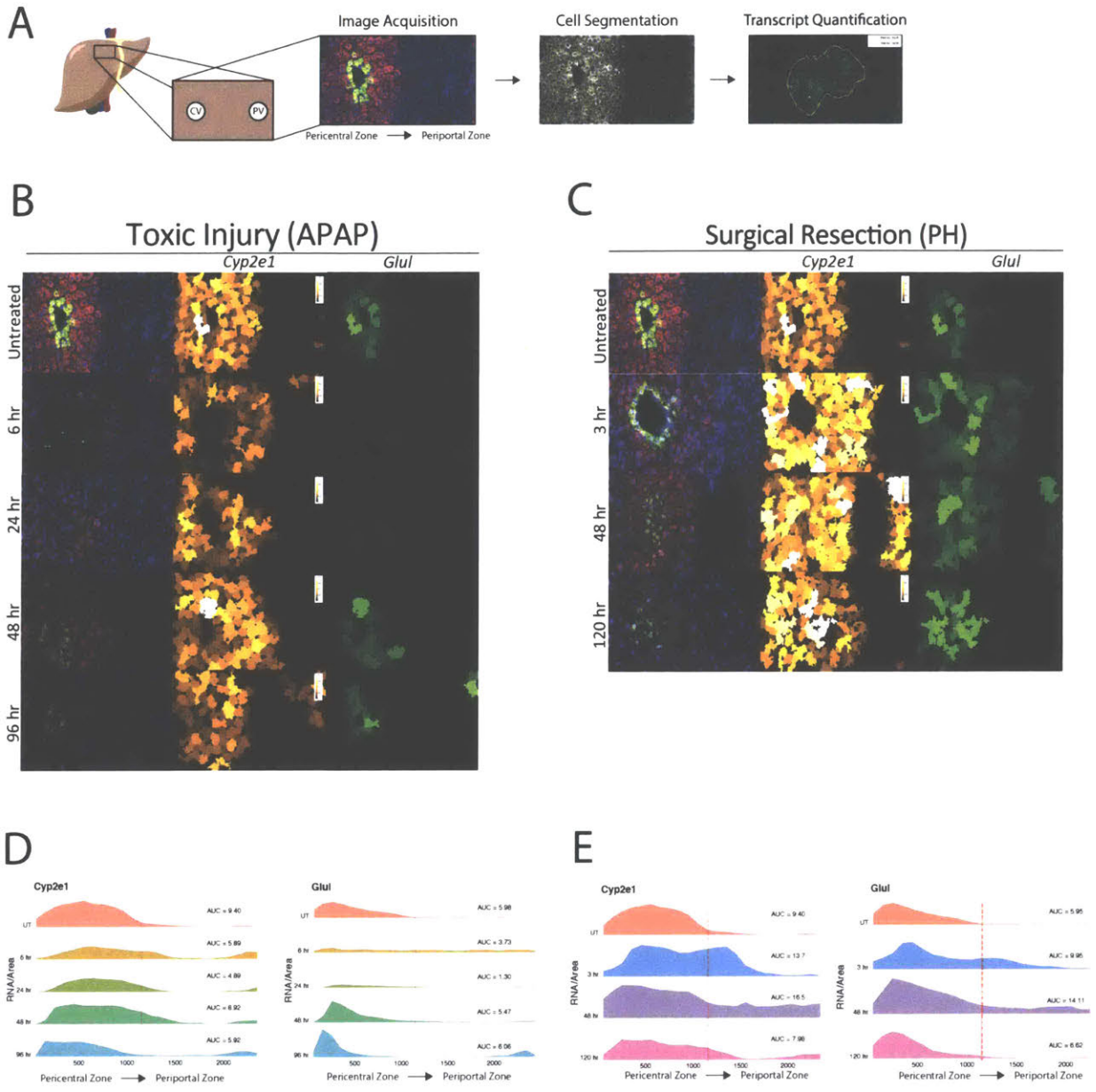

Figure 3-2 | smFISH for *Cyp2e1* and *Glul*.
(A) Schematic for staining and image quantification. **(B, C)** Imaging of liver slice showing *Cyp2e1* (red) *Glul* (green) and DAPI (blue) for untreated and each APAP-treated **(B)** or PH-treated **(C)** time point. (left column). Cell outlined and colored by number of *Cyp2e1* transcripts (brown, low; white, high) for each condition (middle column). Cell outlined and colored by number of *Glul* transcripts (black, low; bright green, high) for each condition (right column). **(D, E)** Quantification of gene expression intensity (y-axis) across the lobule (x-axis) for *Cyp2e1* and *Glul*. 90% of area under the curve (AUC) for UT is to the left dashed red line. Total AUC posted aboved each plot. APAP treated **(D)** and PH-treated **(E)**.

To validate these findings, we performed smFISH analysis to precisely quantify and spatially resolve gene expression (**Figure 3-2A**). We found that *Cyp2e1* reaches further into the midzone of the liver lobule following APAP treatment and necrosis of the pericentral region at 6 and 24 hrs; it recedes to the pre-injured area with complete resolution of the injury after 48 hrs (**Figure 3-2B,D**). *Glul* expression, meanwhile, is normally localized to a single layer of cells surrounding the central vein, which undergoes necrosis following APAP overdose. Surprisingly, we did not observe a difference in Glul+ cells following APAP exposure, demonstrating that Glul expression is maintained, through at a low level of expression across the entire liver lobule (**Figure 3-2B,D**, A6 and A24). *Glul* expression returns to normal by 48 hours after APAP dosage (**Figure 3-2B,D**, A48 and A96). These results demonstrate a compensatory expression of critical liver function genes immediately following zone-specific injury.

In contrast to APAP, PH does not result in zone-dependent injury. The dramatic tissue loss causes increased functional demand, as only ~30% of liver tissue remains compared to ~90% in APAP. Functional compensation is observed in this model, evident from a dramatic increase in *Glul*+ hepatoctyes (**Figure 3-1E**), from 18% (Control) to 60% (PH3). smFISH analysis confirms increased expression zones and total expression levels of both *Cyp2e1* and *Glul* (**Figure 3-2C,E**). Taken together, these observations suggest that hepatocytes can alter their transcriptional profile to maintain functions that are diminished due to injury. Further, our data suggest this can occur in a zone-dependent fashion, evident by midzonal hepatocytes up-regulating pericentral marker genes within the APAP model, or a zone-independent fashion relative to the extent of the injury (PH model).

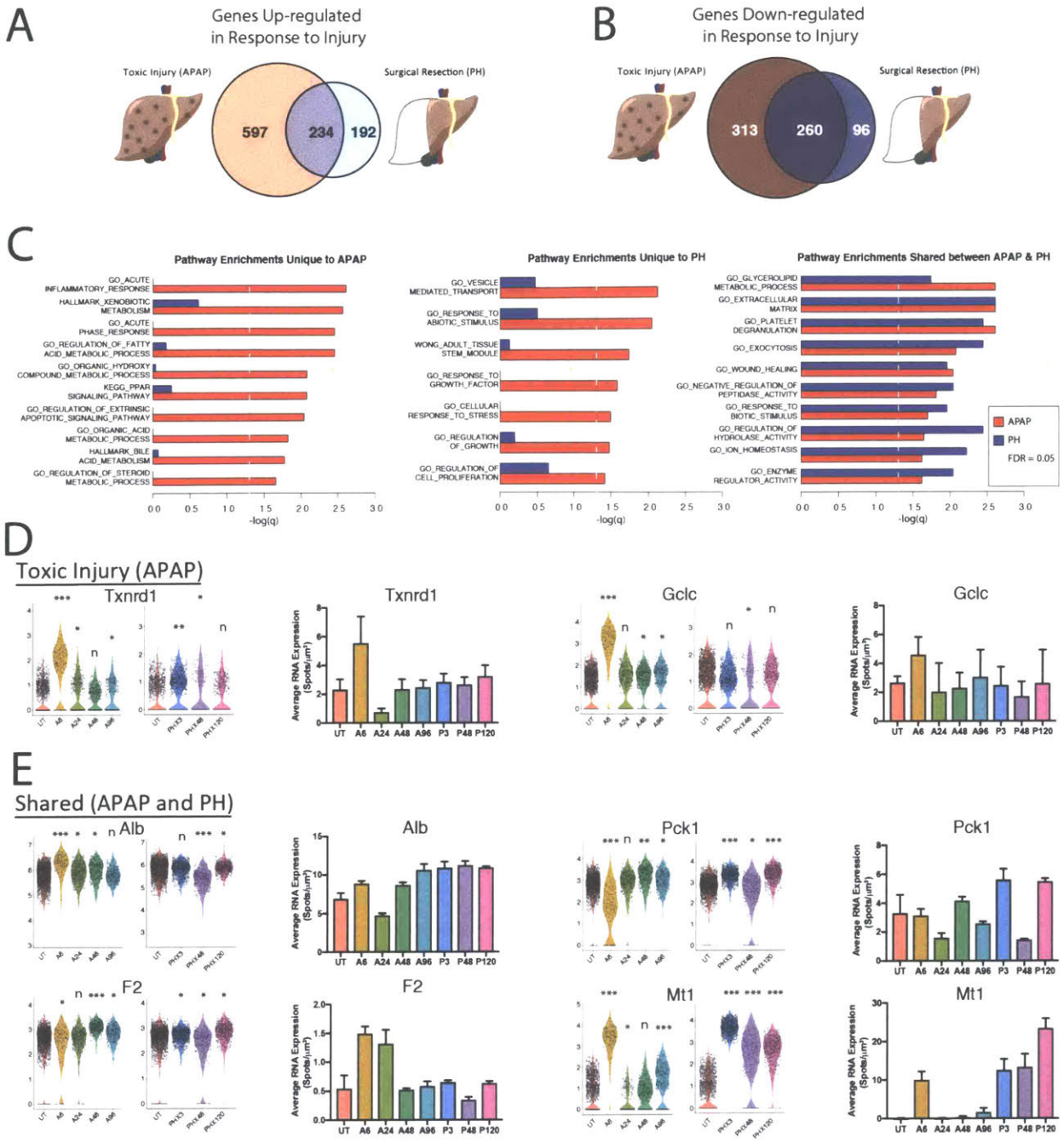

Figure 3-3 | Shared and unique gene expression responses in acute livery injury models. **(A)** Venn diagram showing genes significantly upregulated in response to APAP and/or PH treatment compared to untreated. **(B)** Venn diagram of genes downregulated. **(C)** Pathways with significant overlaps with differentially expressed genes. Significant pathways unique to APAP response (left), unique to PH response (middle) and significant in both responses (right). **(D)** Violin plot of individual genes (*Tsnrd1* and *Gclc*) significantly upregulated in APAP treatment response. Barplot of smFISH quantification of these genes. **(E)** Violin and barplots of individual genes (*Alb, Pck1, F2* and *Mt1*) with upregulation in both APAP and PH-response by smFISH. Effect size calculated by Cohen's d: d < 0.2 = n (negligible); 0.2 < d < 0.5 = * (small); 0.5 < d < 0.8 = ** (medium), d > 0.8 = *** (large).

3.3 Acute liver injury causes both injury-specific and non-specific responses

To further define shared and unique responses for each acute injury model, we calculated differentially expressed genes (DEG) between each treatment condition and untreated (UT), and then pooled results to reveal composite DEG results for APAP and PH (see **Methods**). A number of gene expression shifts can be attributed to injury-dependent affects; however, a large number of gene expression changes observed are shared between the two models sugguesting that many gene expression changes do not depend on the nature of the injury (**Figure 3-3A,B**). Among these DEG, gene set analysis (GSA) reveals an enrichment of pathways involved in toxic injury within the APAP model (**Figure 3-3C**)[10]. The PH model exhibits an enrichment of pathways involved in cell proliferation, which can most likely be attributed to differences in the extent of injury between the two models (**Figure 3-3C**). In the APAP model, there is a substantial anti-oxidant response, both by GSA and the expression of individual anti-oxidant response genes, such as *thioredoxin* (*Txnrd1*) and *glutamate-cysteine ligase subunit c* (*Gclc*) (**Figure 3-3D**), confirming the findings from our single-cell transcriptomic dataset. Importantly, smFISH data reveals an up-regulation of *Txnrd1* and *Gclc* across the entire liver lobule reaching into the periportal region. This suggests that some injury responses are not exclusive to the area of injury, but that all hepatocytes respond similarly to oxidative injury.

Albumin is the most abundant serum protein, and is produced by all hepatocytes across the liver lobule, with the highest expression in the periportal region. Acute injury in both models results in a dramatic up-regulation of albumin across the entire liver lobule (**Figure 3-3E**). This is consistent with the larger total loss of hepatocytes compared to APAP, resulting in a greater need for functional compensation. Similarly, other essential liver functions, examined here by expression of the gluconeogenesis gene *Pck1* and the coagulation factor *F2*, are compensated at a higher degree across the lobule after PH.

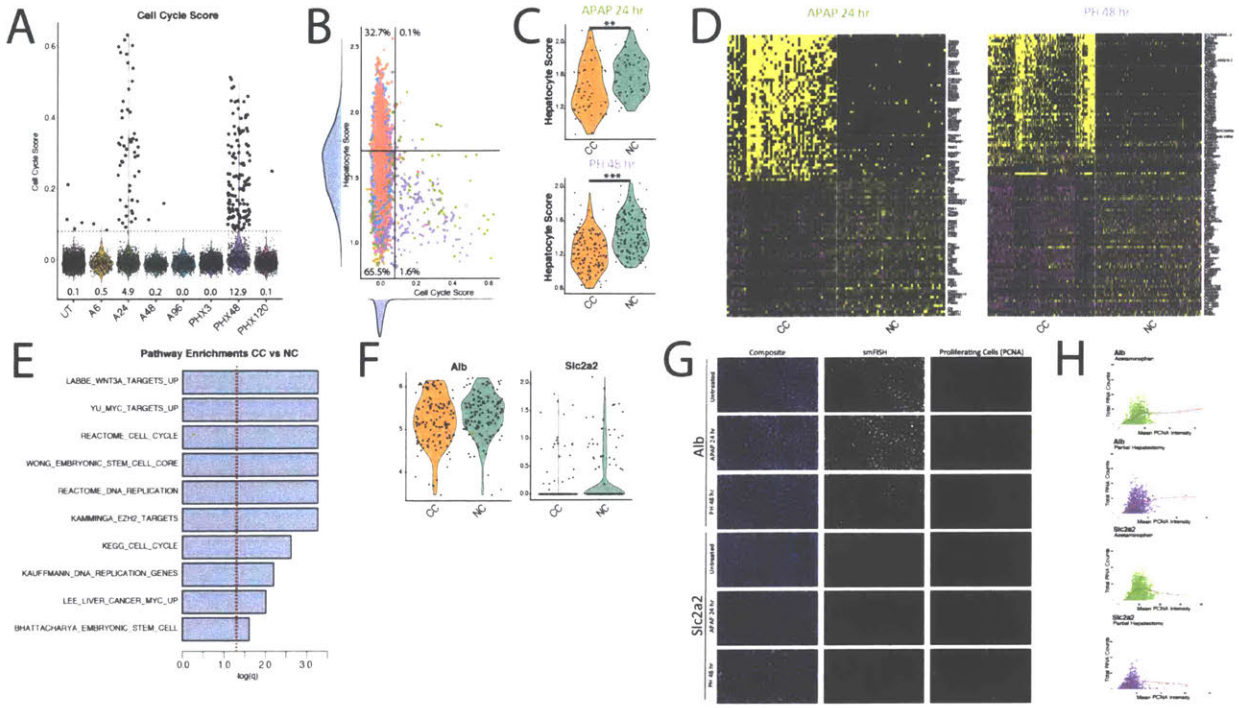

Figure 3-4 | Identification and characterization of of cycling cells.
(A) Violin plot of cell cycle score. Cycling cells (CC, larger black dots) are identified as having a cell cycle score two standard deviations above average (dashed red line). Percentage of cycling cells in each treatment listed below each violin. **(B)** Scatter plot of Hepatocyte Score versus Cell Cycle Score. Horizonal line represents average Hepatocyte Score calculated over all untreated cells. Vertical line represents two standard deviations above the average cell cycle score. **(C)** Violin plots on Hepatocyte Score for all APAP 24hr cycling cells (CC) and an equal number of non-cycling cells (NC) from APAP24 (top) and the same for PH48 CC and NC (bottom). Effect size calculated by Cohen's d: $d < 0.2$ = n (negligible); $0.2 < d < 0.5$ = * (small); $0.5 < d < 0.8$ = ** (medium), $d > 0.8$ = *** (large). **(D)** Heatmap of marker genes fo CC and NC in APAP 24hr (left) and PH 48hr (right). **(E)** Pathway analysis over genes differentially expressed between CC and NC in APAP 24hr and PH48 hr. **(F)** Violin plots of *Alb* and *Slc2a2* in CC and NC. **(G)** smFISH and PCNA costaining images. **(H)** Quantification of RNA expression and PCNA intensity. Functional hepatic markers are selectively maintained in proliferating hepatocytes. Alb shows a maintenance of expression (total RNA counts) in proliferating hepatocytes (mean PCNA intensity) while Slc2a2 reveals a negative correlation. Mean PCNA intensity (IF) and total RNA counts (smFISH) are plotted for individually segmented cells from three lobular areas/condition (UT, APAP 24 hr, PH 48 hr) with Loess regression (red line).

We observe a dramatic up-regulation in metallothionein (*Mt1*) in both injury models (**Figure 3-3E**). It has been suggested that *Mt1* may serve two purposes in tissue injury – protection against further oxidative damage and support for the proliferative response[8]. Further, *Mt1* has previously been shown to be up-regulated in the liver following PH[9,10]. We observe up-regulation of *Mt1* in all hepatocytes across the lobule, highlighting not only the rapidity of hepatic functional adaptation, but also, the plasticity of the majority of hepatocytes across the liver lobule. Expression of *Mt1* is upregulated to a greater degree in PH than APAP and remains elevated throughout the PH time course, where an increased proliferative demand is present due to the increased loss of cell mass in the PH model.

3.4 Cell proliferation impairs adaptation

Most work on liver regeneration has traditionally focused on the proliferative phase of repair. It is unknown, however, whether hepatocytes that are actively dividing can contribute to hepatocyte functional compensation. We identified cells that become transcriptionally active for the cell cycle in the scRNA-Seq dataset (**Figure 3-4A**), and analyzed hepatocyte-specific transcript output compared to those cells at all time points that are not cycling. Compared to non-cycling cells (NC), there is a significant down-regulation of the Hepatocyte Signature Score in cycling cells (CC) in both injury conditions (**Figure 3-4B; Methods**). DEGs reveal substantial differences between cycling and non cycling cells in both injury models. CCs express many classic cell proliferation markers, and exhibit down-regulation of classic hepatic function genes (A24 CC vs NC: $p < 1.4e-03$, Wilcoxon, Effect size d = 0.65, Cohen's d; PH48 CC vs NC: $p < 3.6$ e-11, Effect size d = 0.94, Cohen's d) (**Figure 3-4C,D**). Though proliferating hepatocytes score lower for hepatocyte markers, many of these genes are still expressed at an appreciable level with select markers not changing at all. For example, Albumin expression appears to be maintained in CCs (**Figure 3-4F**), which can be corroborated by smFISH analysis in combination with immunofluorescent co-

staining for the cell proliferation marker PCNA (**Figure 3-4G,H**). However, select genes, such as Slc2a2, appear to be dispensable during proliferation (**Figure 3-4F-H**). Taken together, these data suggest that proliferating hepatocytes have the ability to maintain expression of select hepatic functional markers while other hepatic genes appear to be expendable or fully compensated by the non-proliferating hepatocytes.

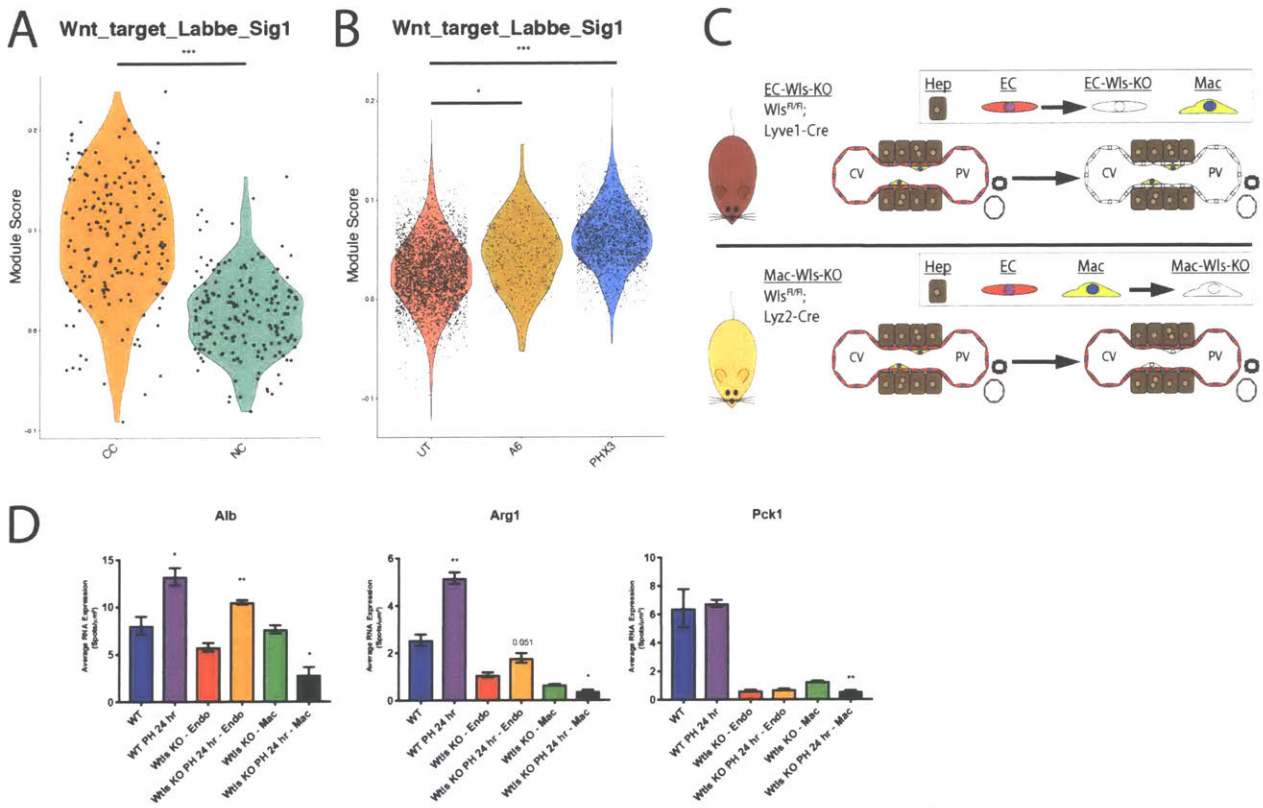

Figure 3-5 | Contribution of Wnt signaling to functional compensation.
Wnt target gene expression score over cycling cells (CC) and non-cycling cells (NC) from A24 and PH48 **(A)** and all hepatocytes grouped by treatment condition **(B)** **(A)(B)** Effect size calculated by Cohen's d: d < 0.2 = n (negligible); 0.2 < d < 0.5 = * (small); 0.5 < d < 0.8 = ** (medium), d > 0.8 = *** (large). **(C)** Wnt knockout mouse models. **(D)** *Alb* expression in untreated and PH 24hr for wild type (WT), β-catenin knock out (B-cat KO), and Wnt-less knock out (Wtls KO) by smFISH. **(E)** *Mt1* expression by smFISH. * p < 0.1, ** p < 0.05, *** p < 0.01.

To identify pathways and potential upstream regulators that may be involved in cell cycle activation, we performed GSA over DEGs calculated between CCs and NCs from A24 and PH48. The results reveal expected upregulation of many cell cycle-related pathways. We also observe enrichment for stem- and development- related pathways (e.g., EMBRYONIC_STEM_CELL, STEM_CELL_CORE, EZH2_TARGETS) and Wnt-related pathways (e.g., WNT3A_TARGETS_UP, MYC_TARGETS_UP, LIVER_CANCER_MYC), suggesting the proliferating hepatocytes may become more stem-like, and that Wnt signaling may be involved in these changes (**Figure 3-4E**). It has recently been shown that Wnt signaling is associated with normal hepatocyte turnover as well as liver regeneration[14]. These Wnts are thought to be derived from the endothelium and contribute to the activation of hepatic stem cell markers (*Axin2* and *Tbx3*). We have observed up-regulation of both *Axin2* and *Tbx3* in each acute injury model with positive cells reaching multiple cell layers into the midzone of the lobule.

3.5 Wnt Signaling mediates functional compensation

Given the demonstrated role of Wnt signaling in both establishing liver zonation and liver regeneration, we investigated whether Wnt signaling activates reprogramming of already present hepatocytes to maintain essential hepatic function[15–19]. Our Seq-Well data corroborates previous observation for increased Wnt activity in proliferating hepatocytes (**Figure 3-5A**)[19–21]. Further, our single-cell transcriptional dataset reveals activation of Wnt target gene expression in the majority of hepatocytes for both the APAP and PH models, which precedes the onset of cell proliferative activity (**Figure 3-5B**; A6, A24, and PH3).

To identify the dependence of the compensatory response on the Wnt/β-catenin pathway, we examined the contribution of both endothelial- (EC-Wls-KO) and macrophage- (Mac-Wls-KO) derived Wnts in the activation of functional compensation using previously described KO mouse

models, which both have intact β-catenin but lack Wnt secretion from respective cell populations (**Figure 3-5C**)[17,22]. Our data suggest that endothelial-derived Wnts are dispensable for functional compensation whereas macrophage-derived Wnts are required (**Figure 3-5D**).

3.6 Discussion

Overall, our study defines a novel functional compensatory mechanism after liver injury where essential liver functions are maintained prior to the onset of cellular restoration through proliferation. We find that hepatocytes upregulate transcription of important liver genes, often by adapting expression patterns traversing zonal boundaries. Select hepatic function genes appear to be dispensable in proliferating hepatocytes. Further, we define a novel dual role for the Wnt/β-catenin pathway in liver regeneration where the pathway not only promotes cell proliferation and a return to the pre-injured cell mass but also promotes functional compensation in order to maintain essential liver function prior to the proliferative response.

The liver is unique in that it maintains complex metabolic function throughout injury and subsequent regeneration[23,24]. This is due to the fact that most injury mechanisms induce a regenerative response where functionally active hepatocytes are thought to be the major contributor to cellular regeneration instead of an already present stem cell population[25–27]. By comparison, many complex mammalian tissues, like cardiac muscle and central nervous tissue, have little to no regenerative capability; others, like intestine, have a high rate of turnover/regeneration that is primarily achieved through a dedicated stem cell population with highly-differentiated/functionally-active cells contributing very little[28].

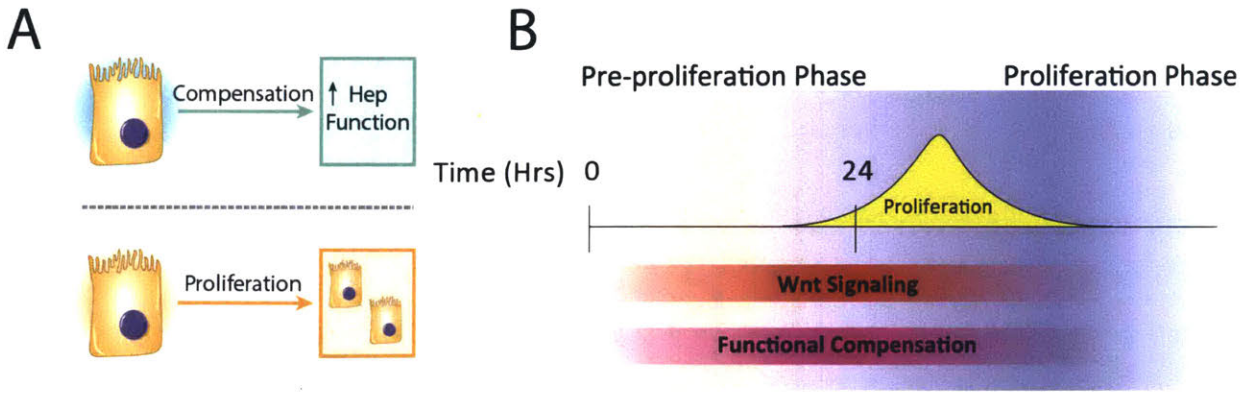

Figure 3-6 | Model of hepatocyte response to acute liver injury.
(A) Top compensating hepatocytes and proliferating hepatocytes are separate groups of cells.
(B) Wnt Signaling and contributes to Functional Compensation. Both precede proliferation.

Liver regeneration within the mouse model shows a peak of hepatocyte proliferation between 30-36 hrs for both PH and APAP-induced injury[29,30]. Cell cycle genes are activated well before hepatocyte proliferation begins (priming phase), as early as 10 min following injury[23,24,31]. However, cell cycle inhibitors, such as p21 and p27, are concurrently up-regulated early in liver regeneration and block progression of hepatocytes into the cell cycle[32,33]. It has been speculated that this co-expression of both stimulators and repressors of the cell cycle is what aids in the control of liver regeneration to a precise end point[23]. It is also plausible that this 'delay' in cell mass recovery helps to prevent genetically-damaged hepatocytes from re-populating the liver, allowing damaged cells time to succumb to injury or complete genomic repair. This could lend to a protective mechanism to reduce the liver's susceptibility to other pathologies, such as cancer. Therefore, it would make sense for the liver to evolve a functional compensatory mechanism in order to maintain essential liver function during this period of delayed proliferative response. The data we present throughout this manuscript supports this notion and further characterizes a novel phase of liver regeneration: the functional compensation phase.

The Wnt/β-catenin signaling cascade has been well-established to play a crucial role in regard to the proliferative response during liver regeneration[16,21,34,35]. Wnts appear to be derived from the endothelium, Kupffer cells/macrophages, and hepatocytes. Stimulation of the pathway can be observed as early as 5 min following PH where it is thought to primarily promote cell proliferation through an up-regulation of cell-cycle genes, such as Cyclin D1, via a β-catenin-dependent mechanism[34]. Hepatocyte-derived Wnts appear to be dispensable for proper hepatic zonation and liver regeneration; however, recent studies have highlighted the importance of endothelial- and macrophage-derived Wnts in the regenerative response. Endothelial-derived Wnts have been shown to be important for maintaining proper pericentral hepatic zonation and show a delayed regenerative response after PH while Macrophage-derived Wnts have no effect on normal hepatic zonation but are equally important for liver regeneration[17,36]. Here, we describe a novel role for the Wnt/β-catenin pathway in the functional compensatory response, which highlights a dual role for it in liver regeneration: maintenance of critical liver function and promotion of the cell proliferation response (**Figure 3-6**). Further, we describe that Kupffer cell/macrophage-derived Wnts appear to be essential for the functional compensatory response while Wnts from endothelial cells appear to be largely dispensable for this role in the midzone and periportal areas, where a large amount of compensation is occurring.

In conclusion, we describe a novel functional compensatory phase of liver regeneration that precedes cell mass recovery in the murine liver following either toxic-induced liver injury (APAP) or surgical resection (PH). We further describe a novel role for the Wnt/β-catenin pathway in promotion of functional compensation via the liver macrophage population. This work further highlights the potential for the Wnt/β-catenin pathway as a target for therapeutic intervention in acute liver failure and other liver pathologies where maintenance of liver function is paramount.

3.7 Methods

Animals

Three-month-old male and female C57BL/6J mice, purchased from Jackson Laboratories (Bar Harbor, ME, USA), were used in these studies. All animals were housed in Association for Assessment and Accreditation of Laboratory Animal Care – accredited facilities at Brigham and Women's Hospital (Boston, MA) under a standard 12-hour light/dark cycle with access to chow and water *ad libitum*. The Institutional Animal Care and Use Committee at Brigham and Women's Hospital approved all studies.

Acetaminophen (APAP) Exposure

Mice were fasted 12 hours before administration of APAP. APAP was dissolved in warm 0.9% saline, and mice were injected with 300 mg/kg APAP, i.p. Food was returned to the mice after APAP treatment. Mice were then used for isolation of primary hepatic cells for single cell RNA-sequencing or tissue harvest for further downstream analysis.

Partial Hepatectomy

Partial hepatectomy surgeries were performed as previously described. Mice were euthanized at 3 hrs, 48 hrs, and 120 hrs post-partial hepatectomy by cervical dislocation under isoflurane anesthesia and livers were harvested for downstream analysis. Further, mice were used for isolation of primary hepatic cells at 3 hrs, 48 hrs, and 120 hrs post-partial hepatectomy.

Isolation of Primary Hepatocytes and Non-parenchymal Cells

Mouse hepatic cells were isolated by a modification of the two-step collagenase perfusion method (1). Cells were isolated from untreated (n=3), APAP-treated mice (n=2 at 6, 24, 48, and 96 hours following APAP exposure), and mice subjected to partial hepatectomy (n=3; 3 hrs, 48 hrs, and 5

days). The digestion step was performed using Liver Digest Medium (Cat. # 17703034; ThermoFisher Scientific; Pittsburgh, PA, USA). Cells were immediately loaded for Seq-Well.

Tissue Harvest

Untreated (n=3 for each sex) and APAP-treated mice (n=2 at 6, 24, 48, and 96 hours following APAP exposure) were euthanized by cervical dislocation following carbon dioxide exposure. Blood was then collected via cardiac puncture and livers were excised for further downstream processing. Serum samples were obtained from the blood and used for analysis of alanine aminotransferase (ALT) activity using commercially available kits (ThermoFisher Scientific, Pittsburgh, PA, USA). Part of the liver tissue was fixed in 10% neutral buffered formalin for 48 hrs and further processed to obtain paraffin blocks and 5 μm thick sections. A piece of liver was frozen in OCT and used to obtain 10 μm fresh frozen sections. The remainder of liver tissue was snap frozen in liquid N_2 and stored at -80°C until used for further processing.

Library Preparation and Sequencing

Sequencing libraries were prepared from the single cell suspension using the Seq-well method as described in Gieran *et. al.* 2017. Briefly, a microwell array was loaded with polyT capture beads (Chemgenes). Then 200ul of media containing 15,000 single cells from the suspension prepared were loaded onto the array and allowed to settle into the wells by gravity. Membrane sealing, lysis, hybridization, reverse transcription, exonuclease digestion, second strand synthesis, PCR, and library construction by Nextera were all performed as previously described. Resulting libraries were quantified by Qubit and tape station (Agilent), and sequenced on an Illumina NextSeq 500 (UT and APAP samples, 2 arrays per run) or a NovaSeq (PH samples, 10 arrays per run).

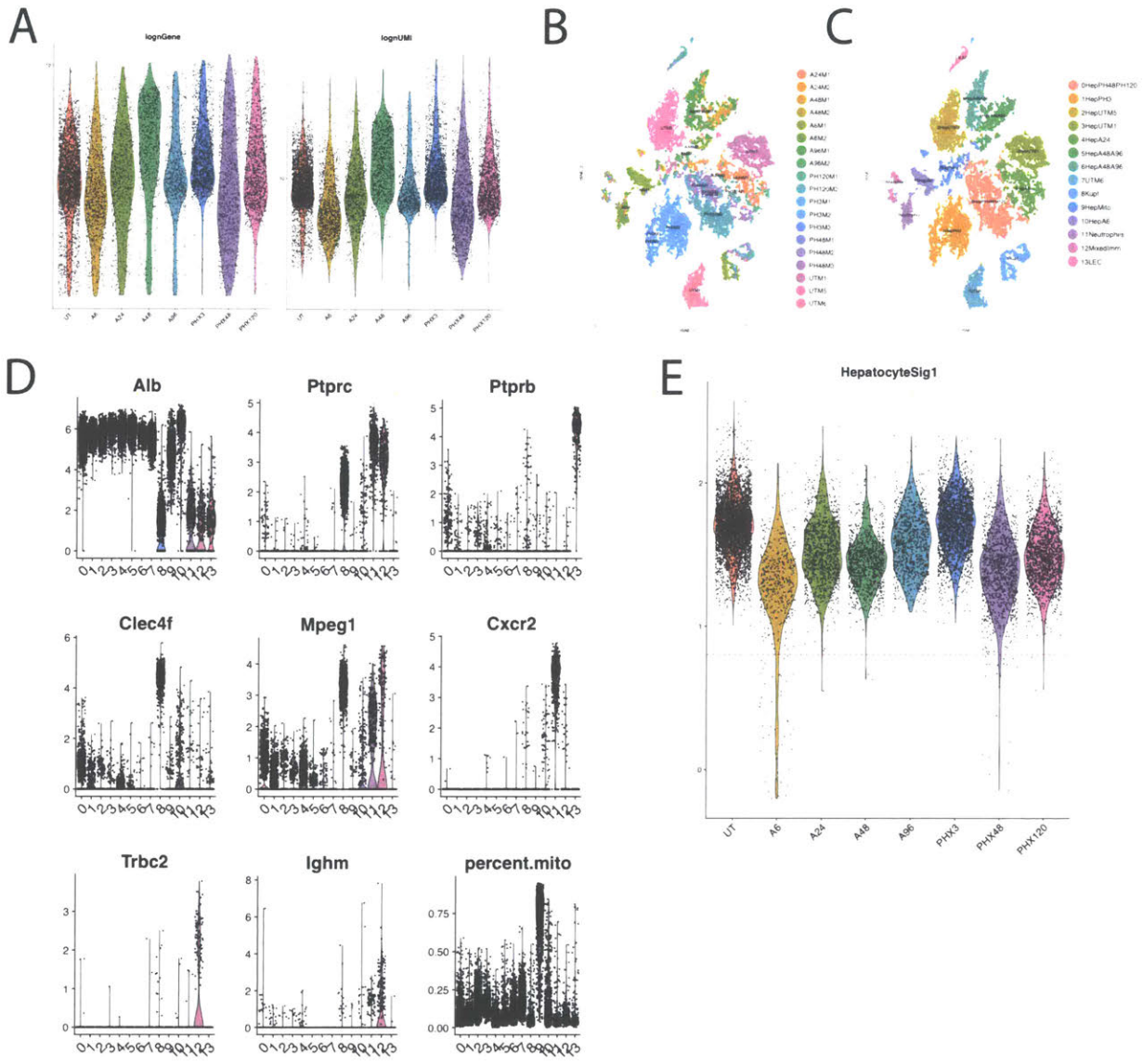

Figure 3-7 | scRNA-Seq Data Processing.
(A) log(nGene) and log(nUMI) for each treatment condition. **(B)** t-SNE colored by mouse of origin. **(C)** t-SNE colored by cluster. Clusters are numbered from most to fewest member cells and annotated by cell type. **(D)** Violin plots for marker gene expressionand percent mitochondrial content (percent.mito) in each cluster. **(E)** Hepatocyte Signature Scores for cells in good quality hepatocyte clusters, grouped by treatment condition. Cells scoring less than 3 standard deviations below the mean (dashed red line) were filtered out as non-hepatocytes. Remaining cells were included in the high quality hepatocyte dataset for further analysis.

Single-cell Sequencing Data Processing

Sequencing data was demultiplexed and aligned to mm10 with STAR aligner. Libraries were sequenced to an average depth of 48,000 reads per cell.

Events with fewer than 400 genes were discarded from the genes x cells data matrix as non-cells, with 16,019 cells remaining. Data was log normalized and TPM-like (base 10,000) normalized and analyzed using the Seurat version 2.3.2 package in R. Distributions for number of genes (nGene) and number of unique molecular identifiers (nUMI) were fairly even across treatment conditions (**Figure 3-**7A). We performed principal components analysis and selected the top 13 PCs for tSNE dimensional reduction. We then performed shared nearest neighbors (SNN) clustering, and identified 14 distinct clusters in the data (**Figure 3-7B,C**). We then performed differential expression across the clusters and plotted expression of marker genes for known liver cell populations (**Figure 3-7D**). We identify nine high-quality hepatocyte clusters, separated by treatment condition; one low quality hepatocyte cluster with high percent mitochondrial content and low nGene and nUMI; a kupffer cell cluster; a liver endothelial cell (LEC) cluster; a neutrophil cluster; and a mixed immune cluster, which appears to contain T cells, B cells and monocytes. We calculated a hepatocyte signature score using AddModuleScore in Seurat over multiple highly expressed hepatocyte genes which span the lobule: *Apoa1, Glul, Acly, Asl, Cyp2e1, Cyp2f2, Ass1, Alb, Mup3, Pck1, G6pc, Fabp1.*

In order to focus on hepatocyte responses, we subsetted our data to include only the nine high-quality hepatocyte clusters. Following subsetting, we observed a remaining few cells scoring low on the hepatocyte signature. We filtered out any cells with a Hepatocyte Signature score less than 3 standard deviations below the average as non-hepatocytes (**Figure 3-7E**). These non-hepatocytes originate primarily from the A6 sample, which has the largest immune infiltration in response to injury and the highest fraction on non-parenchymal cells in the total sample. The

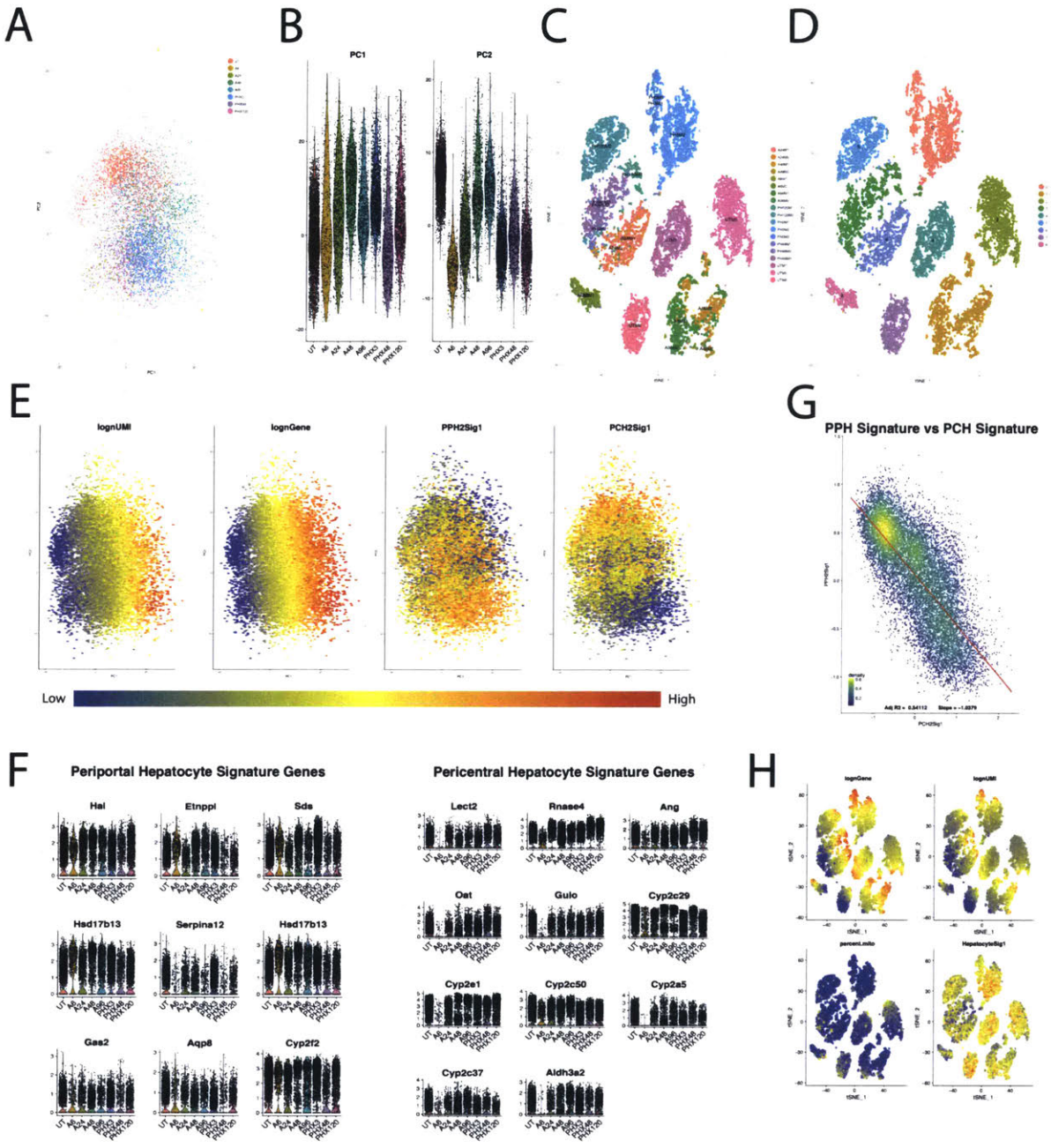

Figure 3-8 | Hepatocyte dataset analysis.
(A) Principle Components Analysis (PCA) of hepatocyte dataset, PC1, PC2. Cells (dots) colored by treatment condition. **(B)** Violin plot of PC1 and PC2 scores for each cell, grouped by treatment condition. **(C)** t-sne, colored by mouse of origin. **(D)** t-SNE colored by SNN clustering assignment. **(E)** PCA (PC1, PC2), colored by lognUMI, lognGene, Periportal Hepatocyte (PPH) Signature, and Pericentral Hepatocyte (PCH) Signature. Blue, low; yellow, medium; red, high. **(F)** Violin plots of genes used to calculate PPH Sig and PCH Sig, grouped by treatment condition. **(H)** t-sne colored by lognGene, lognUMI, percent mitochondrial content (percent.mito) and Hepatocyte Signature Score. Blue, low; yellow, medium; red, high.

filtered non-hepatocytes are likely non-parenchymal cells incorrectly assigned to a hepatocyte cluster by SNN. Following these filtering steps, we retained 10,833 high-quality hepatocytes for analysis.

Single-cell Sequencing Data Analysis (Hepatocyte Data)

We performed dimensional reduction and clustering again on our filtered hepatocyte only dataset. Principal component 1 (PC1) captures technical variation (nGene, nUMI) in the data (**Figure 3-8A, B, E**). This is not surprising for a dataset comprised of a single cell type. Each of our treatment conditions scores similarly on PC1 (**Figure 3-8B**). PC2 captures pericentral-periportal variation (**Figure 3-8B,Eb)**. We identify pericentral and periportal genes in PC2 loading. We also note periportal-pericentral variation captured in PC4. To more clearly visualize pericentral-periportal variation we scored cells on this metric. To generate a list of pericentral genes, we calculated gene by gene correlations and selected moderately expressed genes with large variability in expression across the dataset which correlated positively with *Cyp2e1*, a canonical pericentral gene. To generate a periportal gene list we selected genes negatively correlated with *Cyp2e1* (**Figure 3-8F**). We then calculated the pericental hepatocyte (PCH) score and periportal hepatocyte (PPH) score using AddModuleScore for these genes. We then confirmed that PCH Score and PPH Score are inversely correlated as expected. We observe a pericental-periportal gradient across PC2 using these scores (**Figure 3-8E**). To generate a single score that captures pericentral-periportal character, we subtracted the PCH Score from the PPH Score to create the PPH-PCH Score, in which pericentral hepatocytes will score negatively and periportal hepatocytes will score positively. We confirm that PCH Score and PPH score are inversely related (**Figure 3-8G**).

To better visualize the data, we performed t-SNE dimensional reduction (**Figure 3-8C**). We performed shared nearest neighbors clustering (SNN) to identify groups of similar cells (**Figure**

3-8D). Hepatocytes from all samples look rather similar in lower PCs which describe shared variation, such as technical differences or cross-lobule variation, while the higher PCs capture inter-sample variation. We calculated percent variation captured per PC and generated an elbow plot to determine the correct number of PCs to use in further analysis. We selected the top 13 PC to include in our analysis, which well separated samples by treatment condition and does not appear to be driven by technical effects. We observe a technical gradient across each cluster (which is orthogonal to the pericentral-periportal gradient across each cluster), but the clusters themselves do not appear technically driven (**Figure 3-8H**).

Heatmap genes were found using FindAllMarkers in Seurat, Wilcox test, min.percent = 0.10, thresh.use = 0.25. Mitochondrial (mt-) and hemoglobin (Hbb-, Hba-) genes were removed from the list prior to heatmap plotting.

Shared and unique by injury model gene lists were assembled by combining DE results across all time points for each injury. We ran differential expression using a Wilcoxon test between each treatment condition (A6, A24, A48, A96, PH3, PH48, PH120) individually and untreated (UT). We then combined results across all time points within each injury model. For genes that appeared in the DE results in multiple time points to be combined, we retained the DE result with the largest magnitude average log fold-change for that gene to generate composite DE results for each injury model.

We ran pathway analysis on the composite DE results using the piano R package. Reference gene sets were downloaded from MSigDB (Broad Institute). We used geneSetStat = "fisher", adjMethod = "fdr", and signifMethod = "geneSampling". We then parsed the results to identify shared and unique reference gene sets for each injury. Any reference gene set with a q-value greater than 0.05 was discarded as insignificant. We then identified reference gene sets with

significant overlaps with only APAP and with only PH composite DE results. To focus on truly unique responses, we filtered out any reference gene set from the unique tables which had a q-value < 0.2 for the other injury model. We then identified shared responses by compiling all reference gene sets with a q-value of <0.05 in both APAP and PH. Selected reference gene set -log(q) are plotted in **Figure 3-3**.

To identify cycling cells in the data, we calculated Cell Cycle Score using AddModuleScore in Seurat over the cell cycle markers found in Tirosh *et. al.* 2015. We classified cells with a Cell Cycle Score 2 standard deviations above the average as cycling cells (**Figure 3-4**). To better compare cycling and non-cycling cells (CC and NC, respectively) we subsetted the data to create a dataset containing all 51 CCs from the A24 condition and an equal number of NCs also from A24; similarly, we created a dataset containing all 123 CC from PH48 and an equal number of NCs also from PH48. Pathway analysis was done on a DE result obtained from comparing 174 CC from A24 and PH48 against an equal number of NCs from these time points. Piano was run as described above. We plot -log(q) values for selected reference gene sets with a q value < 0.05. Wnt Target Labbe Sig was calculated using AddModuleScore and the reference gene set LABBE_WNT3A_TARGETS_UP which was identified as significant in Piano gene set enrichment analysis.

Immunohistochemical Analysis

Histology was performed by the histology core at Beth Israel Deaconess Medical Center using standard procedures and automated workflow. Samples were processed and embedded following fixation in 10% neutral buffered formalin for 48 hrs. Samples were embedded in paraffin and sectioned at 5 μm thick. Immunohistochemistry was performed on a Leica autostainer (Leica Biosystems) with enzyme treatment (1:1000) using standard protocols. The antibody used for assessment of cell proliferation was Mki-67, and cell death was ApopTag Peroxidase *In Situ*

Apoptosis Detection Kit (Millipore, Cat. # S7100). Sections were then counterstained with hematoxylin, dehydrated, and film cover slipped. Four representative images were captured per slide. TUNEL-positive area and Mki-67-positive cells were measured and averaged across the four images for each sample using Fiji.

Single Molecule Fluorescent *in Situ* Hybridization (smFISH)

smFISH was conducted using RNAscope technology (RNAscope Fluorescent Multiplex Kit; Cat. # 320850; Advanced Cell Diagnostics; Neward, CA, USA). Fresh frozen sections (10 μm thick) were used following the manufacturer's guidelines. A 4x4 40x maximum intensity projection was created following capture of a 10 uM z-stack (0.5 μM per slice). This resulted in multiple liver lobules available for analysis within a single section. Images were cropped to the size of a single liver lobule and cellular outlines were defined using CellProfiler. smFISH signal was then quantified using FISH-quant.

Statistical Analysis

We calculated p-values for shifts in gene expression or module scores using the Wilcox test, Bonferroni corrected for multiple testing. Gene set enrichment results in piano were calculated using Fisher's test and the gene sampling method and corrected by FDR.

3.8 References

1. Kietzmann, T. Metabolic zonation of the liver: The oxygen gradient revisited. *Redox Biol.* **11**, 622–630 (2017).
2. Michalopoulos, G. K. Hepatostat: Liver regeneration and normal liver tissue maintenance. *Hepatology* **65**, 1384–1392 (2017).
3. Montrief, T., Koyfman, A. & Long, B. Acute liver failure: A review for emergency physicians. *Am. J. Emerg. Med.* (2018). doi:10.1016/j.ajem.2018.10.032
4. Mossanen, J. C. & Tacke, F. Acetaminophen-induced acute liver injury in mice. *Lab. Anim.* **49**, 30–6 (2015).
5. Budnitz, D. S., Lovegrove, M. C. & Crosby, A. E. Emergency Department Visits for Overdoses of Acetaminophen-Containing Products. *AMEPRE* **40**, 585–592 (2011).
6. Couri, T. & Pillai, A. Goals and targets for personalized therapy for HCC. *Hepatol. Int.* **13**, 125–137 (2019).
7. Michalopoulos, G. K. Principles of liver regeneration and growth homeostasis. *Compr. Physiol.* **3**, 485–513 (2013).
8. Thirumoorthy, N., Manisenthil Kumar, K.-T., Shyam Sundar, A., Panayappan, L. & Chatterjee, M. Metallothionein: an overview. *World J. Gastroenterol.* **13**, 993–6 (2007).
9. Oliver, J. R., Mara, T. W. & Cherian, M. G. Impaired hepatic regeneration in metallothionein-I/II knockout mice after partial hepatectomy. *Exp. Biol. Med. (Maywood).* **230**, 61–7 (2005).
10. Jakovac, H. *et al.* Metallothionein expression and tissue metal kinetics after partial hepatectomy in mice. *Biol. Trace Elem. Res.* **114**, 249–68 (2006).
11. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2(+) cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–5 (2015).
12. Zhao, L. *et al.* Tissue repair in the mouse liver following acute carbon tetrachloride

depends on injury-induced Wnt/β-catenin signaling. *Hepatology* (2019). doi:10.1002/hep.30563

13. Yang, J. *et al.* β-catenin signaling in murine liver zonation and regeneration: a Wnt-Wnt situation! *Hepatology* **60**, 964–76 (2014).
14. Preziosi, M., Okabe, H., Poddar, M., Singh, S. & Monga, S. P. Endothelial Wnts regulate β-catenin signaling in murine liver zonation and regeneration: A sequel to the Wnt-Wnt situation. *Hepatol. Commun.* **2**, 845–860 (2018).
15. Planas-Paz, L. *et al.* The RSPO-LGR4/5-ZNRF3/RNF43 module controls liver zonation and size. *Nat. Cell Biol.* **18**, 467–79 (2016).
16. Leibing, T. *et al.* Angiocrine Wnt signaling controls liver growth and metabolic maturation in mice. *Hepatology* **68**, 707–722 (2018).
17. Nejak-Bowen, K. N. & Monga, S. P. S. Beta-catenin signaling, liver regeneration and hepatocellular cancer: sorting the good from the bad. *Semin. Cancer Biol.* **21**, 44–58 (2011).
18. Monga, S. P., Pediaditakis, P., Mule, K., Stolz, D. B. & Michalopoulos, G. K. Changes in WNT/beta-catenin pathway during regulated growth in rat liver regeneration. *Hepatology* **33**, 1098–109 (2001).
19. Carpenter, A. C., Rao, S., Wells, J. M., Campbell, K. & Lang, R. A. Generation of mice with a conditional null allele for Wntless. *Genesis* **48**, 554–8 (2010).
20. Fausto, N. Liver regeneration. *J. Hepatol.* **32**, 19–31 (2000).
21. Su, A. I., Guidotti, L. G., Pezacki, J. P., Chisari, F. V & Schultz, P. G. Gene expression during the priming phase of liver regeneration after partial hepatectomy in mice. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11181–6 (2002).
22. Schaub, J. R., Malato, Y., Gormond, C. & Willenbring, H. Evidence against a stem cell origin of new hepatocytes in a common mouse model of chronic liver injury. *Cell Rep.* **8**, 933–939 (2014).

23. Yanger, K. *et al.* Adult hepatocytes are generated by self-duplication rather than stem cell differentiation. *Cell Stem Cell* **15**, 340–349 (2014).

24. Tarlow, B. D., Finegold, M. J. & Grompe, M. Clonal tracing of Sox9+ liver progenitors in mouse oval cell injury. *Hepatology* **60**, 278–89 (2014).

25. Iismaa, S. E. *et al.* Comparative regenerative mechanisms across different mammalian tissues. *npj Regen. Med.* **3**, 1–20 (2018).

26. Nevzorova, Y. A., Tolba, R., Trautwein, C. & Liedtke, C. Partial hepatectomy in mice. *Lab. Anim.* **49**, 81–8 (2015).

27. Bhushan, B. *et al.* Pro-Regenerative Signaling after Acetaminophen-Induced Acute Liver Injury in Mice Identified Using a Novel Incremental Dose Model. *Am. J. Pathol.* **184**, 3013–3025 (2014).

28. Albrecht, J. H. & Hansen, L. K. Cyclin D1 promotes mitogen-independent cell cycle progression in hepatocytes. *Cell Growth Differ.* **10**, 397–404 (1999).

29. Albrecht, J. H. *et al.* Involvement of p21 and p27 in the regulation of CDK activity and cell cycle progression in the regenerating liver. *Oncogene* **16**, 2141–50 (1998).

30. Bhushan, B. & Apte, U. Liver Regeneration after Acetaminophen Hepatotoxicity Mechanisms and Therapeutic Opportunities. *Am. J. Pathol.* **189**, 719–729 (2019).

31. Ding, B.-S. *et al.* Inductive angiocrine signals from sinusoidal endothelium are required for liver regeneration. *Nature* **468**, 310–5 (2010).

32. Tan, X., Behari, J., Cieply, B., Michalopoulos, G. K. & Monga, S. P. S. Conditional deletion of beta-catenin reveals its role in liver growth and regeneration. *Gastroenterology* **131**, 1561–72 (2006).

33. Yang, J. *et al.* Beta-catenin signaling in murine liver zonation and regeneration: A Wnt-Wnt situation! *Hepatology* **60**, 964–976 (2014).

Chapter 4: Identifying Cellular Changes to the Gastrointestinal System Induced by High Fat Diet

In this chapter, we expand our scope from a targeted acute injury to one organ, to chronic damage affecting many organs. Unlike the acute liver injurie models in the previous chapter, where the cellular response contributes to maintenance of normal function and organ recovery, here the cells respond to a perturbation – six months on a high fat diet – in a way that is detrimental rather than restorative to organismal health. A high fat diet and obesity are known to increase risk for inflammation and cancer in the liver and gut, and indeed we observe these outcomes in our high fat diet mouse model system. To better understand the cellular changes driving these outcomes, we apply scRNA-Seq to samples across multiple gastrointestinal and reference immune compartments (the liver hepatocyte-enriched; the liver non-parenchymal cell-enriched; proximal small intestine; distal small intestine; colon; bone marrow; spleen; and peripheral blood) to profile shifts in cell type composition and cellular behavior in response to six months on a high fat diet compared to control diet. We encounter changes in the liver and gut concordant with known biology and discover pathways whose activation or deactivation may help drive these changes.

4.1 Background

Obesity is linked to increased risk for many health problems, including multiple types of cancers, heart disease and osteoporosis[2]. Rising obesity rates around the world[3,4,5] create a pressing need for a greater understanding of how obesity and diet contribute to poor health at the cellular and molecular level. The influence of diet over organismal health has long been appreciated, and many epidemiological and model organism studies have linked obesity with increased risk for inflammation, fibrosis and cancer in many organs, including the colon[6], liver[7], and pancreas[8]. Cancer, the second leading cause of death in the United States, has a strong nutritional component, with an estimated one-third of cancers due to dietary factors, such as obesity[2]. Studies have suggested that changes in the adult stem cells which maintain homeostasis in mnay organs may lead to the development of tumorigenic stem cells, and that these changes are linked to obesity[6]. However, our understanding of the molecular pathways driving these alterations and actionable targets to therapeutically modulate these tumorigenic shifts in stem cells have only recently begun to take shape.

In 2016, Bayaz and colleagues published work uncovering a propensity for pro-obesity high fat diets (HFDs) to increase stemness in intestinal progenitors in mice and identified activation of pathways that may lead to development of cancer[1]. These researchers maintained mice on a pro-obesity high fat diet for 9-14 months. They found that HFD induces a peroxisome proliferator-activated receptor delta (PPAR-δ) signature in stem and progenitor cell populations in the intestine. In HFD, crypts contained higher numbers of intestinal stem cells (ISC) at the expense of more differentiated cells (lower numbers of Paneth cells) (**Figure 4-1A**). In an organoid culture system, they found HFD and PPAR-δ agonist-treated samples possessed enhanced organoid forming capacity, suggestive of enhanced tumorigenic potential of these cells (**Figure 4-1B**). Additionally, they were able to recapitulate the effects of HFD by treating CD intestinal-

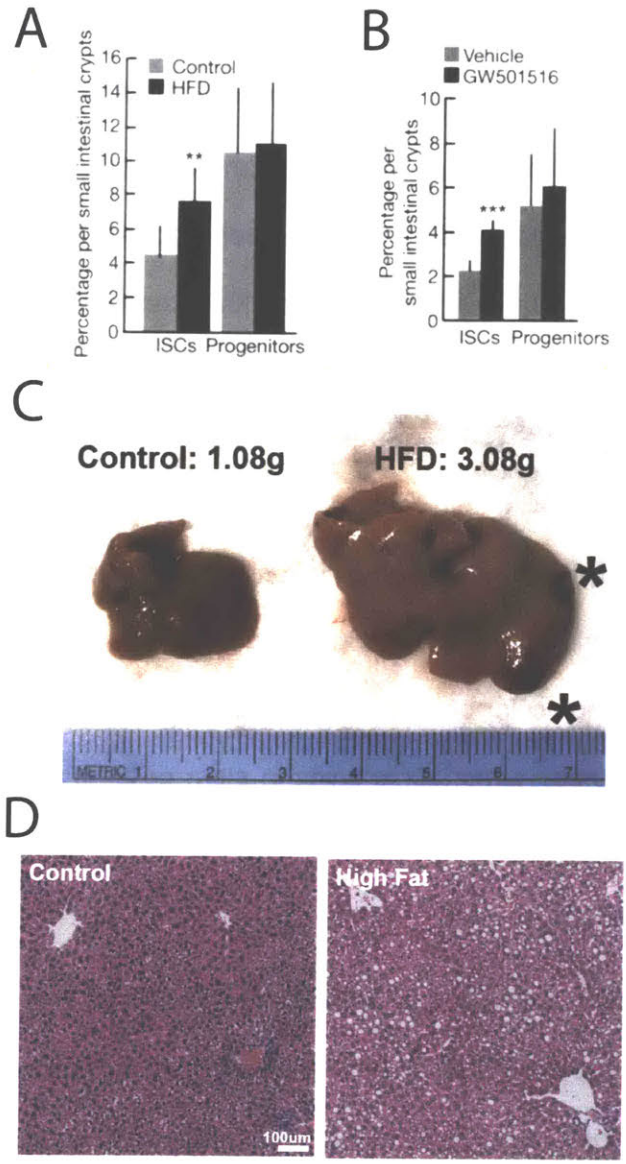

Figure 4-1 | Previously reported effects of HFD on gut and liver.

A and B adapted from Beyaz et. al. [1]

(A) Frequencies of ISCs (Lgr5-GFP^{hi}) and progenitors (Lgr5-GFP^{low}) in the entire small intestine ($n = 10$) as measured by flow cytometry. **(B)** Frequencies of flow-sorted ISCs (Lgr5-GFP^{hi}) and progenitors (Lgr5-GFP^{low}) ($n = 5$) from the entire small intestine of vehicle and GW501516-treated (PPAR agonist) mice. **(C)** Gross anatomy of HFD mouse liver, asterisks denote spontaneous tumors. **(D)** H&E histology of CD and HFD livers. Fat accumulation evident in the HFD.

derived organoids with lipids found in the high fat diet, supporting the notion that it is the interaction with these lipids that initiates the observed changes, and demonstrating the ability to model this phenomenon *in vitro.*

In addition to the described HFD-induced changes in the gut, the Yilmaz Lab also observed liver disease – including fatty livers, steatohepatitis and spontaneous hepatocellular carcinomas (HCCs) – in their HFD mice (**Figure4-1C,D**)[9]. Indeed, non-alcoholic steatohepatitis (NASH) due to obesity and metabolic syndrome has emerged as a major risk factor for cirrhosis and HCC, and is expected to surpass hepatitis C viral infection as the major cause of liver disease by 2022[7,9,10]. Alarmingly, HCC is the fastest-growing cancer in terms of number of diagnoses in the U.S., and few treatment options currently exist[11].

Unlike many cancers in which the cell-of-origin and tumor progression are well defined, the cellular and molecular origins of liver cancer are largely unknown. Even the identity of hepatocyte subsets with progenitor activity or tumorigenic potential have not been definitively described[12]. While there may not be a defined stem cell in the liver, possibly all hepatocytes have the ability to become stem cells under proper conditions, though this remains uncertain. Diverse liver cell types possess regenerative capacity in injury models[13,14]; for example, after limited acute injury, populations of mature hepatocytes can self-duplicate or proliferate to replenish the damaged liver and, after severe injury, bipotent biliary cells can give rise to both mature hepatocytes and biliary cells[15]. These findings indicate that distinct subsets of liver cells harbor regenerative potential which may be induced in a context dependent manner. The identity of these progenitors and their molecular adaptations to HFD that render the organ vulnerable to oncogenic transformation require comprehensive characterization. Possibly, HFDs contribute to tumorigenesis in the liver by altering the regenerative capacity of progenitors in the liver, as was discovered in previous

work in the intestine[1]. To directly explore these points, we seek to define progenitor cell subsets induced or modulated by HFD that may serve as the cell-of-origin for liver tumorigenesis, and to determine mechanisms through which lipid-enriched diets may influence tumorigenic potential in the liver. Identifying and targeting the pathways that contribute to changes in stem-like or progenitor cells and enhance tumorigenesis – whether PPAR, like in the intestine[1] or some other pathway unique to the liver – in obesity has the potential to dampen obesity-linked tumor incidence and disease progression in the gastrointestinal system.

Here, we profile eight samples originating from multiple organs from three HFD and two control diet mice using the scRNA-Seq method, Seq-Well[16]: liver hepatocyte-enriched, liver non-parenchymal cell-enriched, proximal small intestine, distal small intestine, colon, bone marrow, spleen, and peripheral blood to generate our pilot dataset. We observe diet-induced changes within multiple cell types. In our sequencing data analysis, we identify pathways and regulators associated with HFD-induced changes in the profiled compartments. These include upregulation of PPAR pathways in HFD intestinal samples (in line with previous reports[1]), inconsistent PPAR upregulation in HFD liver, transitions toward steatosis and lipid accumulation in HF hepatocytes, and shifts in immune populations in the gut and liver. Additionally, we pilot a protocol for growing organoids from our hepatocyte samples and identify a possible enhanced growth phenotype in HFD hepatocyte-derived organoids.

RESULTS

4.2 scRNA-Seq captures many cell types across tissues

To profile cellular responses to pro-obesity HFD, we performed Seq-Well on mice maintained on a HFD (60% of calories from fat) as described in Beyaz *et. al.*[1] for six months. Diet-induced cellular changes are likely in progress by six months, with mice progressing to more severe

manifestations of obesity-associated metabolic changes and gastrointestinal disease by around nine to 14 months[1,17]. Obesity is linked to cancer and inflammation in both the gut and liver; therefore, we profile samples from multiple gastrointestinal and complementary immune sites to gain a fuller picture of the effects of HFD spanning multiple organs.

Samples from peripheral blood (PB), bone marrow (BM), spleen (Sp), liver hepatocyte-enriched (Hep), liver non-parenchymal-enriched (NPC), proximal small intestine (Prox), distal small intestine (Dis) and Colon (Col) were processed to single cell suspension and loaded onto a Seq-well array (**Table 4-1, Methods**). Prior to loading, crypts from proximal small intestine, distal small intestine and colon were isolated, dissociated into a single cell suspension and sorted into CD45+ and EPCAM+ populations to enrich for immune cells in the sample. The sorted populations (20,000 EPCAM+, 5,000 CD45+) were mixed together and loaded onto an array. Libraries were then prepared and sequenced on a Nova-Seq.

Following data processing and filtering, we obtained a total of 42,684 cells. To visualize our data, we performed dimensional reduction by Principal Components Analysis (PCA) and t-Stochastic Neighbor Embedding (t-SNE). We identified groups of similar cells using Shared Nearest Neighbor (SNN) clustering, and generated module scores from marker genes highly expressed in various cell types to identify the cell type present in each cluster (**Figure 4-2A-D**, **Methods**). We identify several clusters and multiple types of intestinal cells: stem/transamplifying (STA), Enterocyte, Enteroendocrine (EEC), Goblet, Paneth and Tuft. STA and Enterocyte clusters separate mainly by point of origin: proximal, distal, or colon (**Figure 4-2**). We observe immune cell (B cell, T cell) clusters populated by cells from many different compartments. Liver-resident Kupffer and hepatocyte clusters emerge, with clear separation by diet condition in hepatocytes. Finally, we identify bone marrow- and spleen-specific clusters of immature immune cells.

	nGene	nUMI	nCell filter
CD2Col	2212	4889	1737
CD2Dis	2260	6129	1332
CD2Hep	2531	15127	1328
CD2NPC	670	2639	1326
CD2PB	453	1019	449
CD2Prox	4090	13068	1756
CD4BM	3406	10330	1868
CD4Col	2997	9707	1241
CD4Dis	3740	11319	1607
CD4Hep	239	1786	335
CD4NPC	464	1878	742
CD4PB	1755	5002	1075
CD4Prox	3693	10280	1965
CD4Sp	2870	7767	1543
HF2Col	2435	8610	935
HF2Dis	3309	10264	1292
HF2Hep	681	4713	964
HF2NPC	786	2763	1371
HF2PB	1409	5685	878
HF2Prox	2471	8137	1170
HF3BM	1757	4275	1804
HF3Col	646	1917	608
HF3Dis	1251	2748	1279
HF3Hep	216	837	231
HF3NPC	189	899	342
HF3PB	675	1393	745
HF3Prox	2749	6316	1813
HF4Col	3455	9932	1296
HF4Dis	3551	9690	1961
HF4Hep	540	1700	1347
HF4NPC	458	1696	960
HF4PB	1973	4497	1622
HF4Prox	3416	8702	1683
HF4Sp	2899	7201	2079

Table 4-1 | Sample metrics
Samples processed from two control diet (CD2, CD4) and three high fat diet (HF2, HF3, HF4) mice. Samples were prepared from bone marrow (BM), colon (Col), distal small intestine (Dis), liver hepatocyte-enriched (Hep), liver NPC-enriched (NPC), peripheral blood (PB), proximal small intestine (Prox) and spleen (Sp). Due to technical challenges not all samples were obtained from all mice. Number of genes (nGene) and number of unique molecular identifiers (nUMIs) were calculated for each sample over all events called in alignment. Number of cells remaining after filtering for >500 transcripts and >200 genes (nCell filter) reported for each sample.

We apply quality metrics: number of genes (nGene), number of unique molecular identifiers (nUMI, number of RNA molecules captured) and percent mitochondrial content (percent.mito; NB high mitochondrial content can indicate cell membrane disruption from excessively harsh processing and diminished data quality[18]); and we identify two low quality clusters mainly originating from colon and from liver which we omit from further analysis (**Figure4-2C,D)**. We also note lower quality in the HFD hepatocyte clusters relative to other cell types. Cells isolated from the livers of HFD animals are incredibly delicate, likely due to increased volume of fats, and strongly encapsulated within the more fibrotic tissue found in HFD. This increased tissue fibrosis and larger gross liver size necessitated harsher profusion and digestion conditions, as well as longer treatment time to liberate single cells for analysis. It has also been noted in the literature that hepatocyte mitochondrial content can be very high and that hepatocytes appear highly susceptible to damage from processing[18,19]. It has been postulated that these large fragile cells' membranes are more easily disrupted which may further inflate mitochondrial content due to loss of cytosolic mRNAs[18]. Here, metabolic changes induced by HFD may also contribute to shifts in mitochondrial gene expression.

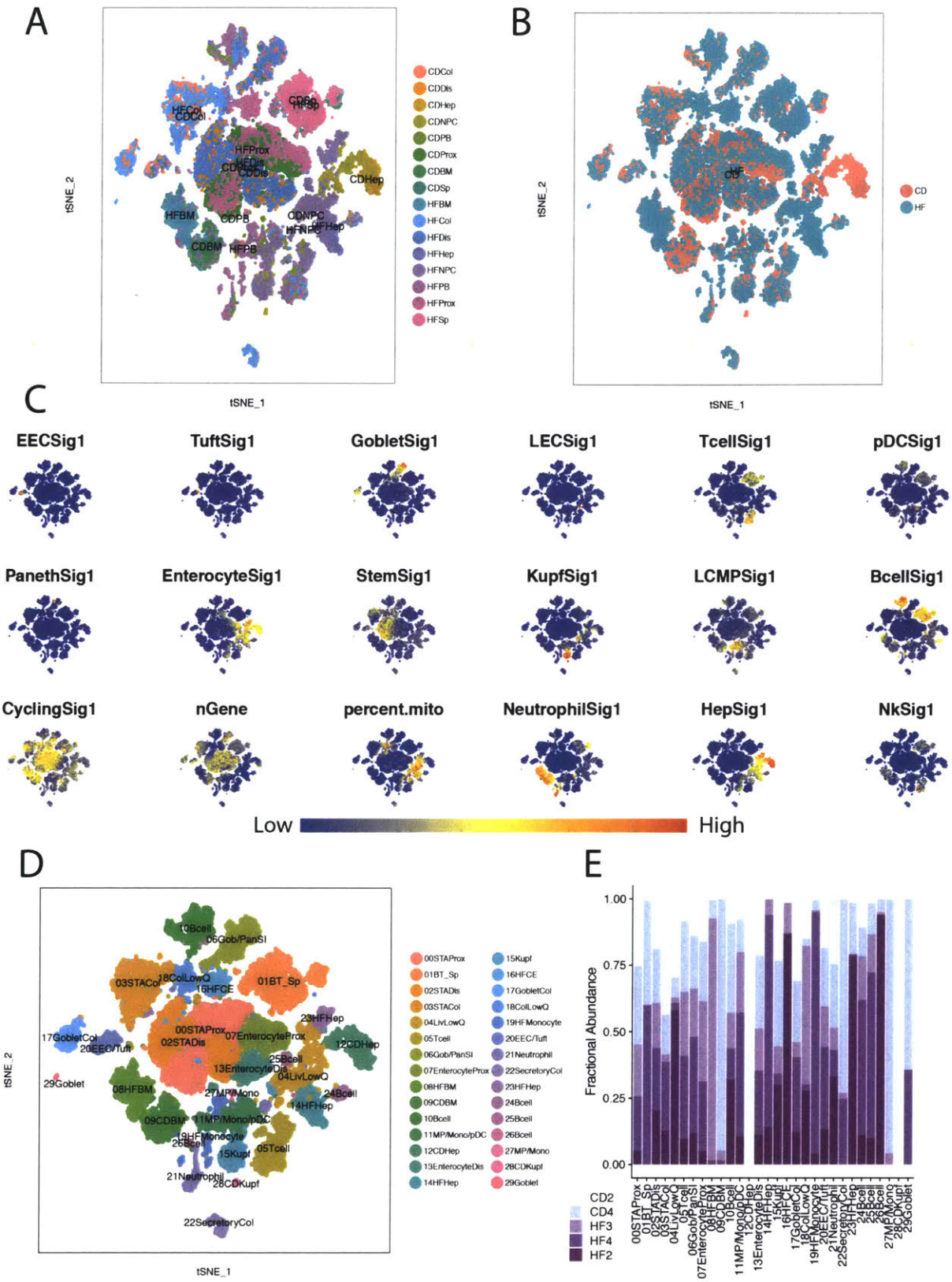

Figure 4-2 | Identification of cell types in full dataset.

Legend next page.

Figure 4-2 | Identification of cell types in full dataset.
Previous page.
(A) t-SNE of all sequenced cells passing initial filter, colored by compartment of origin. **(B)** t-SNE of full dataset colored by diet condition, CD (red) or HFD (blue). **(C)** t-SNEs colored by module score calculated over marker genes for expected cell types, and number of genes captured (nGene), and percent mitochondrial content (percent.mito). Blue, low; yellow, intermediate; red, high. **(D)** t-SNE showing SNN clustering (numbered with 0 being the cluster with the most cells, to 29, the cluster with the fewest). Clusters are annotated with cell type and, for samples primarily from a particular sample, major sample type of origin. **(E)** Stacked barplot showing fractional abundance of cells from each mouse in each cluster. HF mice are shown in purple, CD mice in blue.

4.3 HFD-induced changes in the gut

To more clearly assess diet-induced shifts in the gut, we subsetted the dataset to include only samples that originated in the proximal, distal or colonic regions, and filtered out the low-quality colon cluster and the irreproducible HF2 Proximal cluster (**Methods**). We performed dimensional reduction and SNN clustering again on this subsetted data, and assigned cell type identities to each cluster as we did for the full dataset. Each cluster is populated with cells from HFD and CD mice, yet we noticed some diet-based variation within clusters, especially in the enterocyte clusters (**Figure 4-3AB**).

Previous work has reported that PPAR signaling drives differences between HFD and CD intestinal cells at 9-14 months on the diet[1]. We calculated a PPAR signaling score over our gut cells to determine whether this pathway is already activated in the intestines at 6 months on HFD in each of the cell type clusters captured (**Methods**). Indeed, in some of our cell type clusters we observe upregulation of PPAR target genes (KEGG_PPAR_SIGNALING_PATHWAY, Broad

MSigDB [20]) in the HFD compared to the CD condition, as previously reported[1] (**Figure 4-3C**). Interestingly, we find the strongest upregulation of the PPAR program in the proximal enterocyte cells (d = 2.13), upregulation in the proximal stem (d = 1.94) and transamplifying cells (d = 1.94), and little to no upregulation in the distal and colon samples as well as for cell types other than enterocyte/transamplifying/stem (d = 0.629 to 0.03) (**Figure 4-3C**). These patterns are represented in each of the multiple mice in this dataset (data not shown). This supports the report by Beyaz and colleagues of an HFD-induced increase in PPAR target gene expression in the small intestine and colon at around one year on HFD[1]. Our data show significant upregulation of PPAR targets has begun by six months on HFD in the proximal region, but suggests that changes in the distal or colonic regions may involve lower levels of PPAR activation, occur more slowly, or involve pathways other than PPAR at this time point.

To explore other pathways which may be involved in the distal small intestine and colon we performed pathway analysis with Ingenuity Pathway Analysis (IPA) from Qiagen. We found decreased activity of RB1, a tumor suppressor[21], in HFD enterocytes from the proximal region (z-score -2.945, p-value 5.51e-03) and stem/transamplifying cells from the colon (z-score -3.537, p-value 1.41e-11). We also found upregulation of RELA, involved in NF-κB signaling and inflammation[22] in colon stem/transamplifying HFD (z-score 2.779, p-value 1.85e-05).

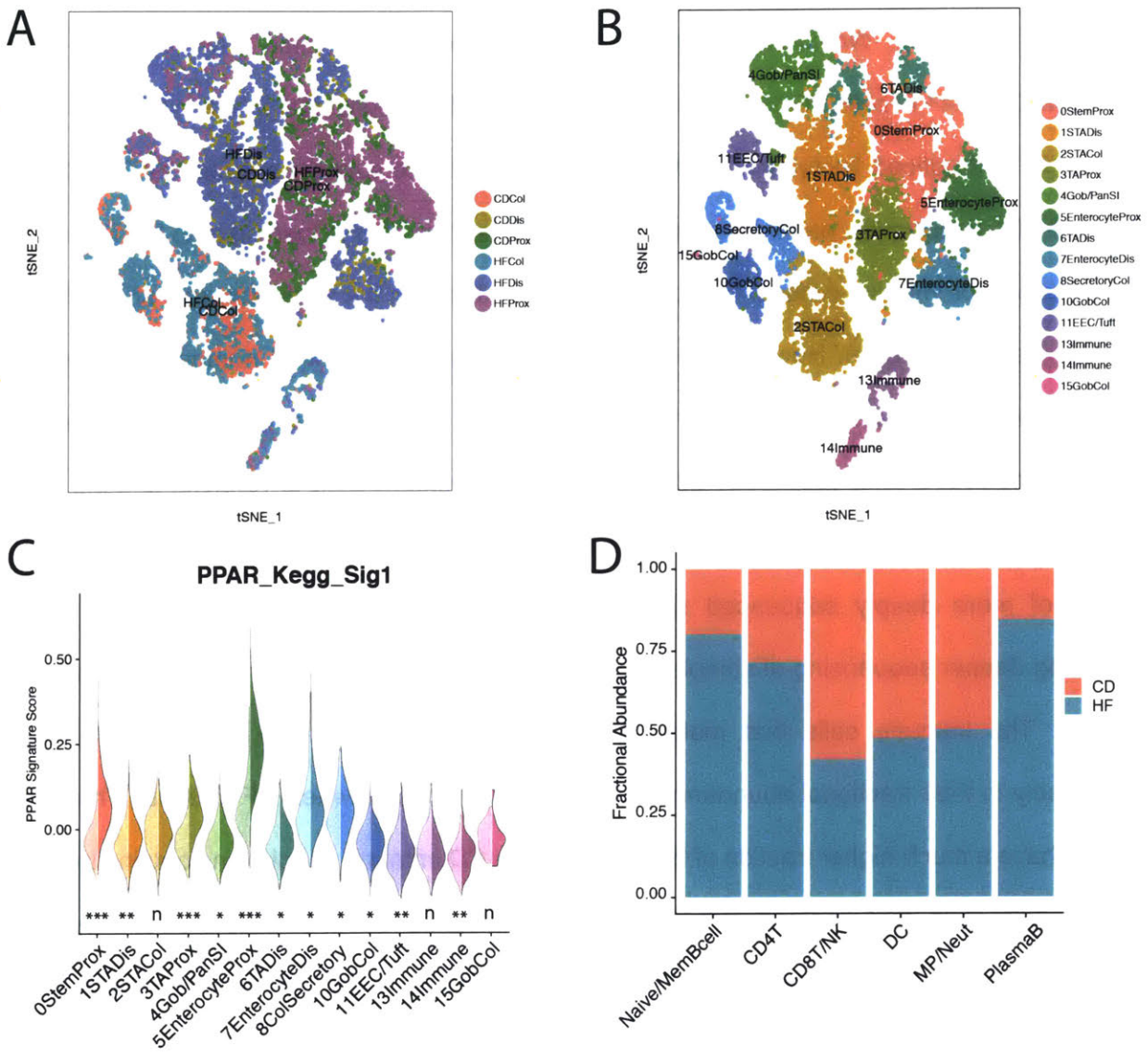

Figure 4-3 | Analysis of gut-originating populations.
(A) tsne over gut-originating samples only, colored by gut location (colon, distal small intestine, proximal small intestinal) and diet (CD, HF). **(B)** tsne with SNN clustering, clusters numbered from most to fewest cells. Clusters are annotated with cell type and sample of origin. **(C)** PPAR signature score calculated for CD (pale colored, left) and HF (bright colored, right) cells in each cluster. Effect size calculated by Cohen's d: d < 0.2 = n (negligible); 0.2 < d < 0.5 = * (small); 0.5 < d < 0.8 = ** (medium), d > 0.8 = *** (large). **(D)** Fractional abundance of HF and CD cells for each type of immune cell in gut dataset.

Next, we subsetted our gut immune cell cluster to further refine our cell type cluster assignments by iterative clustering. We identify several cell types, including B cells (naïve/memory and plasmablast) CD4+T cells, CD8+T cells/NK cells, dendritic cells (DC), macrophages and neutrophils, and noticed fluctuations in the frequencies of these subsets between HFD and CD (**Figure 4-3D**). Importantly, each of the gut samples was sorted prior to loading onto the Seq-Well array to enrich for immune cells, and each array was loaded with the same 1:4 ratio of CD45+ to EPCAM+ cells (**Methods**). However, we ultimately obtained inconsistent numbers of immune cells in each of our samples. This is likely due to a combination of variable relative viability among cell types and sequencing depth. The absolute number of immune cells is variable, and the ratio of immune to non-immune ranges from 14% immune in HF2 to 2% in HF4. There appears to be a trend of more deeply sequenced samples (CD2, HF2) containing more immune cells, suggesting deeper sequencing of samples from experiments 3 and 4 may increase immune cell numbers. The immune cells that make up the immune component in each sample vary considerably in their fractional abundance of immune cell types between HF and CD. The HFD samples have a much higher fraction of B cells while the CD immune population contains more T cells, dendritic cells, and macrophages. This variability may represent an infiltration of B cells or efflux of T cells and macrophages in HFD, or the reverse in CD. Since our protocol accepts a set number of cells an input, an increased infiltration of one cell type will result in a decrease in the fractional abundance of others in the data, making absolute abundance is difficult to determine. Our data suggests some shift in immune composition, but additional experiments, such as flow analysis, are needed to quantitatively ascertain the abundance of various immune subsets in HFD and CD guts.

4.4 HFD-induced changes in the liver

Obesity is known to increase risk for both intestinal and liver disease. The HFD mice in this study do begin to develop liver problems by 6 months on the diet and, in some cases, progress to spontaneous HCC at later time points (**Figure 4-1C**). To dissect and study HFD-induced transformations in the liver at single-cell resolution, we applied Seq-well to liver samples from HFD and CD. Biological changes in the HFD liver make hepatocytes more sensitive to processing due to fat accumulation while, at the same time, making the liver larger, more fibrotic and difficult to dissociate, presenting challenges in processing. For these reasons, our HFD liver data is of lower quality (lower nGene, lower cell number, higher percent mitochondrial content) than CD liver, but still interpretable (**Table 4-1, Figure 4-4A**). We have already made several adjustments to the protocol to improve data to this point (**Methods**), but future iterations may make additional adjustments to improve HFD liver data quality.

We subsetted our dataset to include only samples originating in the liver, performed dimensional reduction and reclustering (**Methods**). We identify expected liver cell types: Hepatocytes, liver endothelial cells (LECs), Kupffer cells, macrophage/monocytes, pDCs, T cells, B cells and neutrophils (**Figure 4-4A,B**). Intriguingly, the hepatocytes form distinct clusters separating by diet, with some diet-based shifts evident in other cell types as well (**Figure 4-4C**).

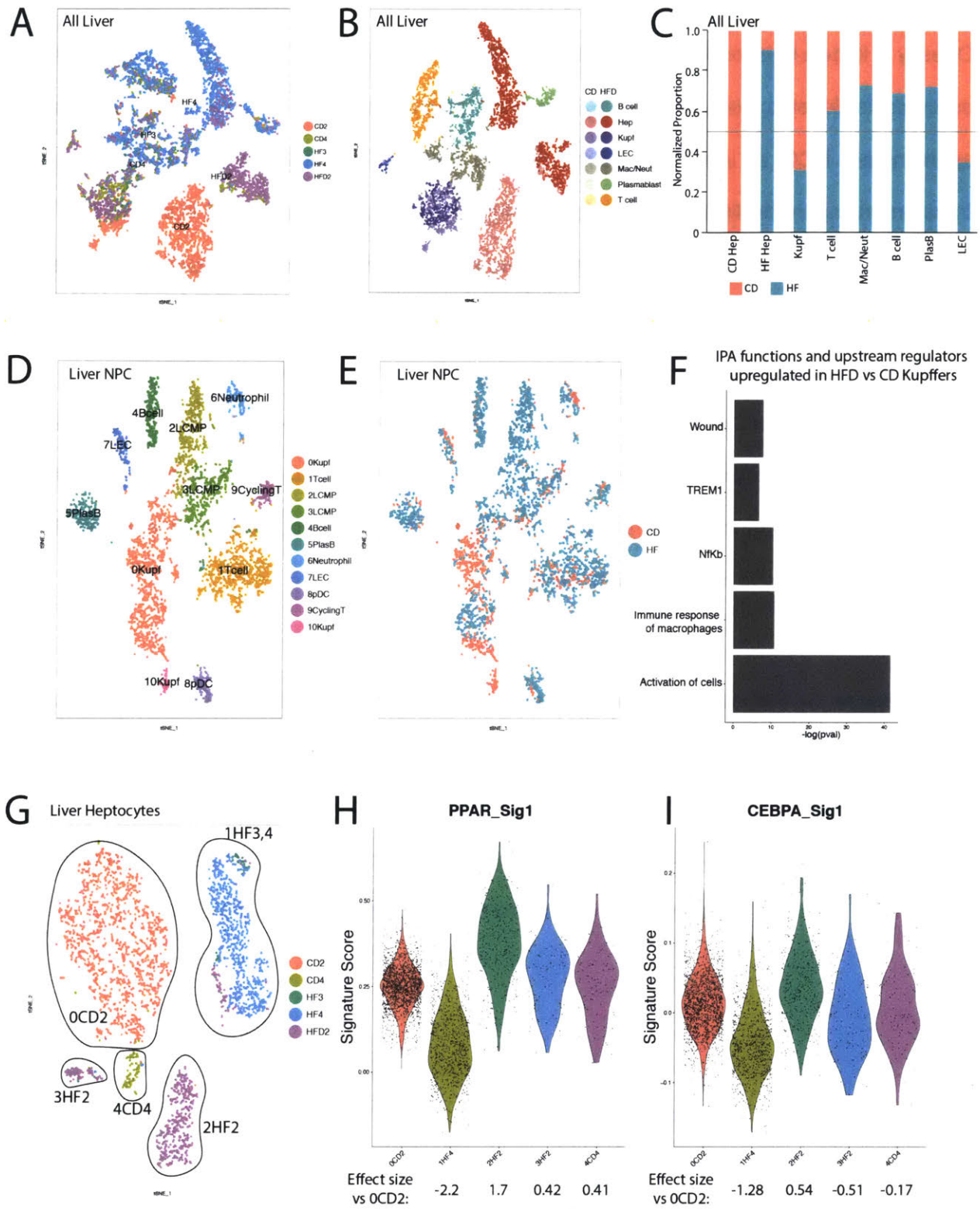

Figure 4-4 | Analysis of liver-originating populations.

Figure 4-4 | Analysis of liver-originating populations.
Previous page.
(A) t-SNE of liver-originating samples, colored by sample type and diet condition. **(B)** SNN clustering. Clusters annotated with cell type. Cells originating from control lighter colored; cells originating from HFD vibrant colored. **(C)** Stacked barplot of fractional abundance of cells from each mouse in each identified liver sample cluster. HFD purple; CD blue. **(D)** Iterative clustering over non-parenchymal liver cells (NPCs). SNN clustering and cell type annotation. **(E)** t-SNE of NPC liver cells colored by diet condition. **(G)** Iterative clustering over hepatocytes. Colored by mouse of origin. SNN clusters outlined in black. **(H)** PPAR activation signature score over hepatocyte clusters. **(I)** CEBPA activation signature score over hepatocyte clusters. Effect size calculated by Cohen's d.

We performed iterative clustering over the non-parenchymal cells of the liver to gain greater resolution in calling the cell types represented. We identify Kupffer cells, liver capsule macrophages (LCMP), pDCs, Neutrophils, liver endothelial cells (LEC), B cells, and T cells (**Figure 4-4D**). Kupffer cells were the most plentiful cell type in the NPC dataset and appear to separate slightly by diet condition (**Figure 4-4E**). We ran IPA over genes differentially expressed between HFD and CD kupffer cells. Within "Diseases & Functions", we found upregulation in HFD of "Immune response of macrophages" (z-score 1.778, p-value 1.65e-11), "Activation of cells" (z-score 2.294, p-value 2.42e-42) and "Wound" (z-score 2.219, p-value 1.97e-08). HFD also showed upregulation of the activity of several upstream regulators such as pro-inflammatory NF-κB (z-score 2.179, p-value 3.98e-11) and TREM1 (z-score 2.938, p-value 2.31e-07). Taken together, these results present HFD kupffer cells as more activated and more inflammatory than in CD (**Figure 4-4F**).

To specifically analyze cellular responses in the hepatocyte data, we selected hepatocyte clusters, filtered on a mitochondrial content cutoff of 50%, as has been reported previously[18], and performed iterative clustering over the remaining cells (**Methods**). We identify a large cluster of hepatocytes originating mainly from CD2, a large cluster from HF4 and HF3, a smaller cluster

from HF2, a small cluster from CD4 and another small cluster from HF2 (**Figure 4-4G**). We performed differential expression between the HFD and CD hepatocytes and ran pathway analysis on the resulting differentially expressed genes through IPA. Results from IPA "Diseases & Functions" identifies upregulation of "Liver steatosis" (z-score 3.522, p-value 2.93e-21), "Hepatic steatosis" (z-score 3.522, p-value 2.93e-21), "Inflammation of liver" (z-score 1.857, p-value 1.50e-09), Oxidative stress (z-score 3.657, p-value 1.10e-11) and "Accumulation of cholesterol" (z-score 2.320, p-value 3.95e-07), as well as a decrease in "Synthesis of lipid" (z-score -3.501, p-value 6.89e-36) in the HFD compared to CD hepatocytes. IPA upstream regulators show a downregulation in HFD of activity of SREPF2 (also known as Srebp2, z-score -4.883, p-value 1.67e-24), a transcription factor responsible for activating synthesis and uptake of cholesterol and fatty acids[23]. This aligns well with the expected biology of the HFD liver, confirming that we have captured interpretable data.

We next asked whether activation of the PPAR pathway in the HFD condition occurs in the liver as it does in the gut. We find PPAR target genes upregulated in HF2 compared to CD2, but down regulation of PPAR in HF3 and HF4 compared to CD2 **(Figure 4-4H)**. Samples from experiment 2 were resequenced to achieve greater sequencing depth, while some samples from experiments 3 and 4, especially HF4Hep, are undersequenced, possibly affecting PPAR target gene expression sensitivity. We will address this possibility by resequencing these samples to improve depth. Alternatively, PPAR activation may occur in some HFD mice, but not others at the six month time point. In search of other regulators which may contribute to HFD-induced changes in the liver, we combed our Upstream Regulator IPA results for potential drivers of HFD-induced changes in the liver. IPA identifies significant downregulation in HFD of activity of RB1 (z-score -5.82, p-value 8.43e-14), a transcription factor with tumor suppressive function[21], and down regulation, particularly in HF4, of activity of CEBPA (z-score -4.749, p-value 4.93e-13), a

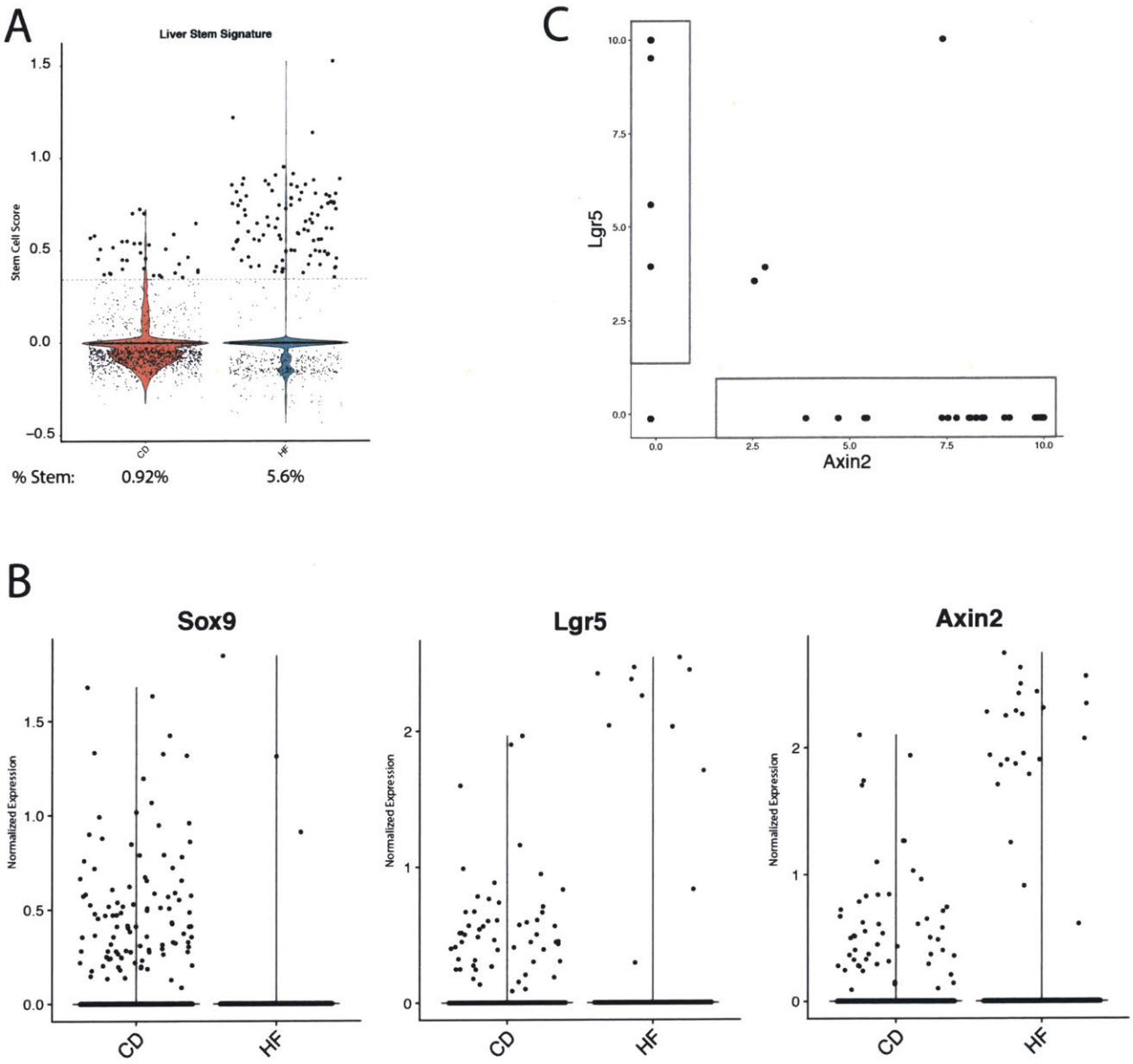

Figure 4-5 | Stem cell gene expression in hepatocytes.
(A) Expression of liver stem cell genes module score in CD and HFD hepatocytes. Liver stem cells called as scoring two standard deviations above the average (red line). Percentage of stem cells in each sample listed below. **(B)** Violin plots of expression of selected genes from the stem cell module. **(C)** Biaxial plot of Axin2 vs Lgr5 expression in identified stem cells.

transcription factor involved in cell cycle regulation, lipid and glucose metabolism in the liver, and leptin expression and body weight homeostasis, whose function is known to be suppressed in HCC and other types of liver disease[24] **(Figure 4-4I)**. IPA also identified upstream regulators whose function increased under HFD conditions, including NCOR1 (z-score 2.6, p-value 2.85e-10), which can contribute to thyroid hormone resistance, and hormonal and metabolic changes[25,26]. Additional work is needed to further explore and validate the potential contributions of these pathways to HFD-induced changes in the liver.

Changes in proliferative potential and "stemness" can prime cells to grow in dysregulated ways, possibly leading to cancer. Although a dedicated liver stem cell population has never been definitively identified[12], we postulate that a subset of hepatocytes may activate stem function in response to HFD and progress toward the development of HCC. Thus, we scored hepatocytes on a stem cell signature to identify changes is stem-like expression and searched for changes in these stem-like cells which may lead to HCC. More specifically, we scored hepatocytes on expression of stem marker genes *Lgr5, Axin2, Sox9, Ascl2, Tbx3* and *Gkn3* (**Methods**). We identified hepatocytes which have activated a stem cell program as cells scoring at least two standard deviations above average. **(Figure 4-5 A)**. A much higher percentage of HFD hepatocytes score as stem cells than CD hepatocytes (5.6% vs 0.92%), supporting the notion that HFD may increase stemness in the liver, similarly to what has been reported in the gut[1]. High expression of the stem signature in hepatocytes was driven mainly by expression of Sox9, *Lgr5* and/or *Axin2.* HFD appears to dysregulate expression of these genes, with suppressed expression of the stem gene *Sox9* and increased expression of *Lgr5* and *Axin2* in HFD compared to CD (**Figure 4-5B**). Interestingly, *Lrg5* and *Axin2*, the stem genes most highly expressed in HFD hepatocyte stem cells, are expressed largely mutually exclusively, in contrast to the gut where they are coexpressed (**Figure 4-5C**). Many of the genes correlated with *Lgr5* expression and

Axin2 expression in the hepatocyte dataset are involved in cytokinesis and cell cycle pathways[20], supporting the notion that cells expressing these genes may possess increased proliferative potential. Further identification and characterization of changes in stemness within hepatocytes will serve to pinpoint the cellular origins of HCC, which remain poorly defined.

4.5 Liver Organoids

Organoids can serve as a useful model system for evaluating perturbations *in vitro* and assessing the stemness of input samples. In the intestine, HFD samples possessed greater capacity to form and grow organoids, a characteristic of their enhanced stemness[1]. This same characteristic may enhance their ability to progress to tumors. Here, we pilot a recently published hepatocyte organoid protocol[27] with our HFD and CD hepatocyte-enriched samples to assess their relative abilities to form and grow organoids.

We seeded organoids in matrigel and noted clear morphological differences between the HFD and CD hepatocytes, suggestive of their biology. Despite efforts to seed equal numbers of cells for both conditions, in this pilot experiment, the HFD was seeded much less efficiently than the CD (more cells matrigel at day 0 for CD than HFD) (**Figure 4-6A**). As expected, only a few organoids formed and grew very slowly[27]. We continued to grow the organoids under the prescribed conditions and noted heterogeneous morphology. Some organoids appeared solid and bumpy or branched (the hepatocyte organoid morphology) while others appeared cystic and spherical (the cholangiocyte or biliary morphology). Over time, the cultures shifted to contain all cholangiocyte morphology organoids in all samples (**Figure 4-6B**). We observed some individual

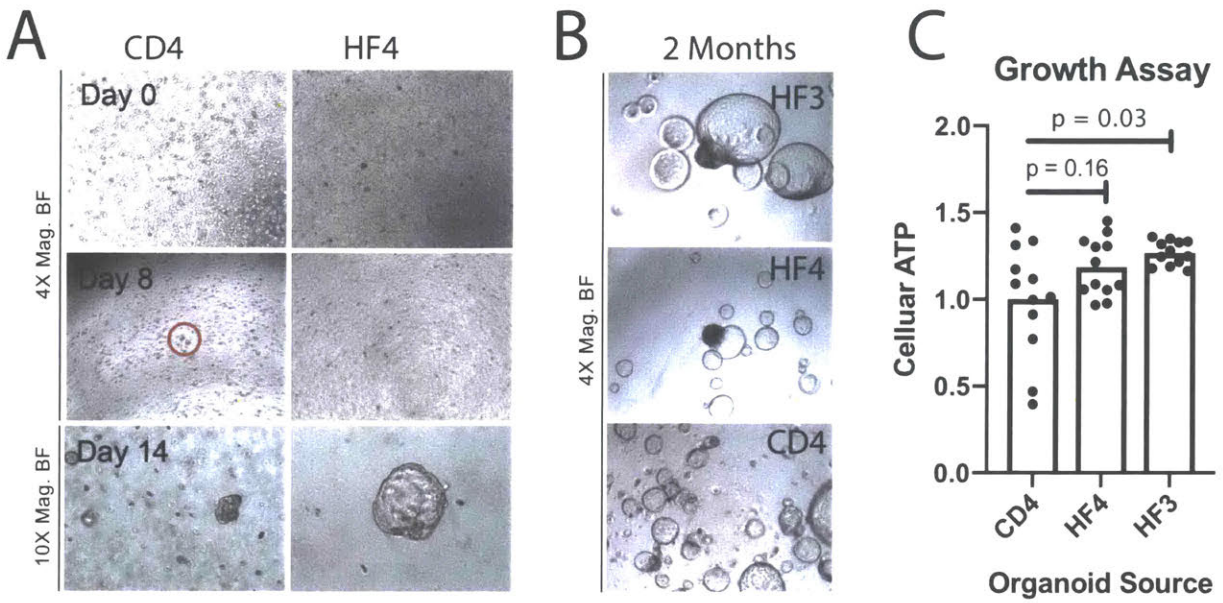

Figure 4-6 | Hepatocyte-derived organoid growth.
(A) Heptocytes seeded in matrigel at 0 days. Growth at 8 and 14 days, small, growing organoids circled in red in CD Day 8. **(B)** Organoids after 2 months in culture. **(C)** ATPase growth assay on organoids after 2 months in culture. ANOVA with corrections for multiple comparisons.

hepatocyte morphology organoids shifting to cholangiocyte morphology in culture; additionally, the culturing conditions apply a selective pressure in favor of the faster growing cholangiocyte-like organoids. The large, spherical cholangiocyte organoids may break apart during passaging and seed more of these organoids in the new matrigel. The published protocol did not report this cholangiocyte shift phenomenon, which is likely a product of the much older mice used in our study (a few weeks old in the published protocol versus seven to eight months in the experiments here). Bidirectional ability for biliary and hepatic cells to regenerate one another has been reported *in vivo*[15], suggesting the possibility of interconverting between these cell types under proper organoid culture conditions (which have yet to be determined). By around two months in culture,

the liver organoids began growing much more rapidly, likely due to adaptation to culture conditions and selective pressure for cells able to grow rapidly *in vitro*. We performed an ATPase growth assay and detected a subtle enhanced growth phenotype in the HFD hepatocyte-derived organoids (CD4 vs HF4 p-value = 0.16, CD4 vs HF3 p-value = 0.03, ANOVA with multiple testing correction) even after months in culture (**Figure 4-6C**). Additionally, we performed Seq-Well on the liver organoids to determine how faithfully they recapitulate the transcriptional profiles of the hepatocytes from the same animals which were immediately profiled (HF3, HF4, CD4 hepatocyte-enriched samples). Data processing and analysis for this experiment are ongoing.

4.6 Follow-up and ongoing experiments

Our current pilot dataset analysis has characterized many cell types from multiple compartments, identified biologically meaningful shifts in transcript expression, and nominated pathways which may participate in driving these changes; yet, more work remains to be done to further explore and validate these observations.

In agreement with earlier work from the Yilmaz Lab[1], we observe upregulation of PPAR in HFD intestinal samples. In our data, PPAR upregulation occurred mainly in the proximal region. We also note significant upregulation of PPAR activity in enterocyte cells from the proximal region, not mentioned in previous work which was focused on stem and progenitor cells. Importantly, our experiments were performed at an earlier time point than in the published work (6 months vs 9-14 months on diet), thus representing earlier initiation of PPAR activation at least in some regions. Additional experiments, such as the organoid, and imaging experiments described in the earlier publication[1], could be performed at the earlier time point used here to validate our findings.

Beyond the gut, we also observe diet-induced changes in the liver. While upregulation of PPAR activity was inconsistent in the liver, it remains a possible driver of liver changes in need of further investigation. In the gut, treatment with a PPAR agonist recapitulated the effects of HFD, and a similar experiment could be performed in the liver or liver organoids to determine whether a PPAR agonist can recapitulate HFD effects in the liver as well. In addition to PPAR, we identify other pathways which may be activated or deactivated in the liver. Modulation of these upstream regulators through agonists or inhibitors will provide insight into the role of these pathways in driving changes in the HFD liver.

In both liver and gut, we note possible shifts in immune composition, but these shifts are difficult to interpret in existing data due to confounding factors (CD45 sorting in gut samples, technical effects, undersequencing of some samples). Flow analysis to quantify the abundance of immune subsets in the liver and gut will determine whether immune populations, such as B cells in the gut or macrophages in the liver, infiltrate into these organs in HFD. Additionally, we have sequencing data from reference immune sites in the peripheral blood, bone marrow and spleen. Comparing between immune cells within the liver or gut to immune cells outside these organs will reveal how these immune cells respond to the inflammatory or oncogenic environment in the HFD gastrointestinal system. "Spill over" genes from the free RNAs in the media in cell loading complicate comparisons of a given cell type across sample types. Application of a computational tool such as SoupX[28] for background correction to remove this contamination is needed to properly compare immune cells across organs.

Infiltrating immune cells may travel between the liver and gut and support cross-talk between gastrointestinal and immune sites. Further analysis of bone marrow, peripheral blood and spleen samples may identify immune responses to HFD outside the GI system if such responses exist.

In our dataset HFD and CD bone marrow samples do cluster separately, but technical differences in sequencing depth dominate the differences between the HFD and CD data in this compartment. After deeper sequencing, we will be equipped to better compare these samples.

Analysis of sequenced liver organoids is ongoing and when complete will determine how closely the organoids recapitulated the biology of the hepatocytes from the HFD and CD livers. Results may guide further optimization of the organoid culturing protocol. Future organoid experiments will work toward better normalizing seeding efficiency to facilitate comparison of organoid forming efficiency between diets. Ideally, we will be able to grow organoids with a consistent hepatocyte phenotype and controlled seeding from HFD and CD samples. Withdrawal of some growth factors, such as WNTs, from organoid culture during the first few days could select for cells already primed for growth or proliferation in the *in vivo* environment and may confer a stronger growth advantage to the HFD-derived organoids. If so, this would demonstrate the enhanced ability of HFD hepatocytes to survive and grow *in vitro* and possibly form tumors *in vivo*. In the published work with gut organoids from HFD and CD, culturing CD-derived organoids with lipids recapitulated the effects of HFD. Similarly, culturing CD hepatic organoids with these lipids will determine whether these fats also affect these liver organoids in analogous ways.

Extensions of this work to future projects may include building a dataset over a full-time course of 3, 6, 9 and 12 months, repeating experiments with female mice to explore sex-differences in HFD responses, and extending our work to human samples are discussed in detail in Chapter 6. The work described here and these extensions will deepen our understanding of the effects of obesity and diet on the gastrointestinal system and development of diet-induced cancer, and point toward potential therapeutic targets. Further validation and development of these candidate targets may one day lead to improved treatment options for NASH, HCC, and intestinal cancers.

4.7 Methods

Mice

Mice were maintained on a high fat diet (HFD) or control diet (CD) for 6 months, as described previously[1]. Liver samples (hepatocyte-enriched and NPC-enriched) were obtained as described in Chapter 3. Intestinal samples (proximal small intestine, distal small intestine and colon) were processed to enrich for crypts, then dissociated to single cell suspensions. Single-cell suspensions were sorted on a Sony SH800 flow sorter into CD45+ (immune) and Epcam+ to increase input of immune cells. One array was loaded for each intestinal sample with a sorted population of 5,000 immune cells and 20,000 Epcam+ cells. Counting of sorted populations showed that only about half as many cell as expected are in the sorted populations, so the arrays were loaded with close to the target of 15,000 cells.

Library preparation and sequencing

Samples were run according to the Seq-Well[16] version 2 protocol with second strand synthesis with the following adjustments: increased loading from 10,000 to 15,000 cells, media for loading and sealing was changed from RPMI to Hepatocyte media for liver samples and crypt media for gut, cell loading time was increased from 5 minutes to 15-20 minutes for liver samples and 10 minutes for all other samples. We note that prompt processing of all samples, especially liver samples, is essential to obtain quality so each sample was processed as soon as it was ready, rather than waiting for several samples to run in parallel.

Libraries were sequenced on a Nova-Seq (Illumina) at 12 libraries per run. Undersequenced libraries from experiment 2 were sequenced again to improve sequencing depth. Resequencing of some samples in Experiments 3 and 4 to increase sequencing depth is still needed. Average nGene <1000 may indicate undersequncing. Sequencing output was aligned to mm10 by STAR aligner. Events with fewer than 500 transcripts captured were discarded as non-cells. Remaining

cells were filtered on >200 genes expressed. More stringent filtering results in excessive loss of HFD hepatocytes. This filtering does allow some low-quality cells/events into the dataset, but these cells readily cluster together and are filtered out from subsequent analysis.

Sequencing data analysis

Filtered data was analyzed using primarily the R package Seurat version 2 from the Satija Lab. We performed dimensional reduction by Principal Components Analysis (PCA) over the dataset. We selected significant PCs from the Elbow Plot and performed t-Stochastic Neighbor Embedding (tsne) over the selected PCs and Shared Nearest Neighbor (SNN) clustering over those same PCs. Differential expression was performed using the "FindMarkers" function and the Wilcoxon statistical test.

We created cell type signatures using the AddModuleScore function in Seurat and a list of marker genes for each expected cell type. These module scores were used to assign cell types to SNN clusters. Marker genes for cell types were obtained from Haber *et. al.* for intestinal cell types[29] and Halpern *et.al.* for liver cell types[30]. We also created module scores for selected pathway gene lists, such as KEGG_PPAR in the same way.

To gain further resolution in our analysis, we performed iterative clustering. In very large datasets cell types or subtypes which are small in number compared to the total often do not drive enough of the total variation to clearly cluster out by SNN. By subsetting the data to include only a smaller selection of cells, we can increase resolution to call more subtle differences or identify rarer cell types within this subset group as variation driven by the small group of cells is now enough of the total to separate clearly by SNN. We perform iterative clustering over groups selected by sample of origin, and cell type and cluster(s).

We perform pathway analysis using Ingenuity Pathway Analysis (IPA) from Qiagen on selected differentially expressed gene lists to identify biological processes which may vary between the compared groups. DEGs were identified as described above, and filtered to include only genes with a p-adjusted value of <0.1 and an average log-fold change of >0.25 for input to IPA. Core analysis was run using default settings. Interesting IPA results were curated manually.

Organoid Culture

Hepatocyte organoid culturing was performed as described previously[27]. For organoid Seq-Well experiment, organoids were dissociated to single cell suspension and loaded 15,000 cell per array. On array each was run for organoids from HF3, HF4 and CD4. ATPase growth assay was performed after approximately 2 months in culture. Aspirate media from well, add 65 uL CTG3D (Promega) to each well, seal plate and shake at room temperature 30 minutes. Transfer 15ul to white 384 wp (in triplicate), read at 1sec lum interval time on lumenesence plate reader.

Statistics

Effect size for expression of module scores was calculated Cohen's d. Significance in the organoid growth assay was calculated using ANOVA with correction for multiple testing.

4.8 References

1 Beyaz, S. *et al.* High-fat diet enhances stemness and tumorigenicity of intestinal progenitors. *Nature* **531**, 53-58, doi:10.1038/nature17173 (2016).

2 Cordain, L. *et al.* Origins and evolution of the Western diet: health implications for the 21st century. *Am J Clin Nutr* **81**, 341-354, doi:10.1093/ajcn.81.2.341 (2005).

3 Ramachandran, A. & Snehalatha, C. Rising burden of obesity in Asia. *J Obes* **2010**, doi:10.1155/2010/868573 (2010).

4 Kain, J., Uauy, R., Vio, F. & Albala, C. Trends in overweight and obesity prevalence in Chilean children: comparison of three definitions. *Eur J Clin Nutr* **56**, 200-204, doi:10.1038/sj.ejcn.1601301 (2002).

5 Vandegrift, D. & Yoked, T. Obesity rates, income, and suburban sprawl: an analysis of US states. *Health Place* **10**, 221-229, doi:10.1016/j.healthplace.2003.09.003 (2004).

6 Mihaylova, M. M., Sabatini, D. M. & Yilmaz, O. H. Dietary and metabolic control of stem cell function in physiology and cancer. *Cell Stem Cell* **14**, 292-305, doi:10.1016/j.stem.2014.02.008 (2014).

7 Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of NAFLD development and therapeutic strategies. *Nat Med* **24**, 908-922, doi:10.1038/s41591-018-0104-9 (2018).

8 Pinte, L., Balaban, D. V., Baicus, C. & Jinga, M. Non-alcoholic fatty pancreas disease - practices for clinicians. *Rom J Intern Med*, doi:10.2478/rjim-2019-0005 (2019).

9 Argo, C. K. & Caldwell, S. H. Epidemiology and natural history of non-alcoholic steatohepatitis. *Clin Liver Dis* **13**, 511-531, doi:10.1016/j.cld.2009.07.005 (2009).

10 El-Serag, H. B. Hepatocellular carcinoma: recent trends in the United States. *Gastroenterology* **127**, S27-34 (2004).

11 Waghray, A., Murali, A. R. & Menon, K. N. Hepatocellular carcinoma: From diagnosis to treatment. *World J Hepatol* **7**, 1020-1029, doi:10.4254/wjh.v7.i8.1020 (2015).

12 Grompe, M. Liver stem cells, where art thou? *Cell Stem Cell* **15**, 257-258, doi:10.1016/j.stem.2014.08.004 (2014).

13 Forbes, S. J. & Newsome, P. N. Liver regeneration - mechanisms and models to clinical application. *Nat Rev Gastroenterol Hepatol* **13**, 473-485, doi:10.1038/nrgastro.2016.97 (2016).

14 Gilgenkrantz, H. & Collin de l'Hortet, A. Understanding Liver Regeneration: From Mechanisms to Regenerative Medicine. *Am J Pathol* **188**, 1316-1327, doi:10.1016/j.ajpath.2018.03.008 (2018).

15 Raven, A. *et al.* Cholangiocytes act as facultative liver stem cells during impaired hepatocyte regeneration. *Nature* **547**, 350-354, doi:10.1038/nature23015 (2017).

16 Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* **14**, 395-398, doi:10.1038/nmeth.4179 (2017).

17 Winzell, M. S. & Ahren, B. The high-fat diet-fed mouse: a model for studying mechanisms and treatment of impaired glucose tolerance and type 2 diabetes. *Diabetes* **53 Suppl 3**, S215-219 (2004).

18 MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383, doi:10.1038/s41467-018-06318-7 (2018).
19 Weibel, E. R., Staubli, W., Gnagi, H. R. & Hess, F. A. Correlated morphometric and biochemical studies on the liver cell. I. Morphometric model, stereologic methods, and normal morphometric data for rat liver. *J Cell Biol* **42**, 68-91 (1969).
20 Institute, B. *GSEA MSigDB*, <http://software.broadinstitute.org/gsea/msigdb> (2019).
21 Dyson, N. J. RB1: a prototype tumor suppressor and an enigma. *Genes Dev* **30**, 1492-1502, doi:10.1101/gad.282145.116 (2016).
22 Ricca, A. *et al.* relA over-expression reduces tumorigenicity and activates apoptosis in human cancer cells. *Br J Cancer* **85**, 1914-1921, doi:10.1054/bjoc.2001.2174 (2001).
23 Madison, B. B. Srebp2: A master regulator of sterol and fatty acid synthesis. *J Lipid Res* **57**, 333-335, doi:10.1194/jlr.C066712 (2016).
24 Reebye, V. *et al.* Gene activation of CEBPA using saRNA: preclinical studies of the first in human saRNA drug candidate for liver cancer. *Oncogene* **37**, 3216-3228, doi:10.1038/s41388-018-0126-2 (2018).
25 Fozzatti, L. *et al.* Resistance to thyroid hormone is modulated in vivo by the nuclear receptor corepressor (NCOR1). *Proc Natl Acad Sci U S A* **108**, 17462-17467, doi:10.1073/pnas.1107474108 (2011).
26 Mottis, A., Mouchiroud, L. & Auwerx, J. Emerging roles of the corepressors NCoR1 and SMRT in homeostasis. *Genes Dev* **27**, 819-835, doi:10.1101/gad.214023.113 (2013).
27 Hu, H. *et al.* Long-Term Expansion of Functional Mouse and Human Hepatocytes as 3D Organoids. *Cell* **175**, 1591-1606 e1519, doi:10.1016/j.cell.2018.11.013 (2018).
28 Young, M. B., S. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv* (2018).
29 Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333-339, doi:10.1038/nature24489 (2017).
30 Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat Biotechnol* **36**, 962-970, doi:10.1038/nbt.4231 (2018).

Chapter 5: A Reproducibility-based computational framework identifies an inducible, enhanced antiviral dendritic cell state in HIV-1 Elite Controllers

This chapter is adapted in accordance with BMC's open access policy from the following article published in Genome Biology:

Martin-Gayo, E*., Cole, M.B*., Kolb, K.E*., Ouyang, Z., Cronin, J., Kazer, S.W., Ordovas-Montanes J., Lichterfeld, M., Walker, B.D., Yosef, N., Shalek, A.K., and Yu, X.G., "A Reproducibility-Based Computational Framework Identifies an Inducible, Enhanced Antiviral State in Dendritic Cells from HIV-1 Elite Controllers," *Genome Biology* 19, 10 (2018).

* Denotes equal authorship

Human immunity relies on the coordinated responses of many cellular subsets and functional states. Inter-individual variations in cellular composition and communication could thus potentially alter host immune function. Here, we explore this hypothesis by applying single-cell RNA-Seq to examine viral responses among the dendritic cells (DCs) of three elite controllers (ECs) of HIV-1 infection. To overcome the potentially confounding effects of donor-to-donor variability, we present a generally applicable computational framework for identifying reproducible patterns in gene expression across donors who share a unifying classification. Applying it, we discover a highly functional antiviral DC state in ECs whose fractional abundance after *in vitro* exposure to HIV-1 correlates with higher $CD4^+$ T cell counts and lower HIV-1 viral loads, and that effectively primes polyfunctional T cell responses *in vitro*. By integrating information from existing genomic

databases into our reproducibility-based analysis, we identify and validate select immunomodulators that increase the fractional abundance of this state in primary peripheral blood mononuclear cells (PBMCs) from healthy individuals *in vitro*. Overall, our results demonstrate how single-cell approaches can reveal previously unappreciated, yet important, immune behaviors and empower rational frameworks for modulating systems-level immune responses that may prove therapeutically and prophylactically useful.

5.1 Background

Effective immune responses are founded upon the orchestrated dynamics of complex cellular ensembles. While substantial work has been done to catalog the cell types, states, and interactions that inform these behaviors, [2-8] greater resolution is still needed to fully understand cell subsets and behaviors. Recent studies have uncovered significant and functionally relevant heterogeneities even within seemingly identical cell populations [2,12-16]. This unprecedented degree of cellular diversity points to new opportunities to redefine the structure behind systems-level immune responses and identify potential therapeutic or prophylactic strategies rooted in modulating immune composition and interactions.

One powerful approach for uncovering correlates of immune fitness is to study individuals that demonstrate exceptionally effective immune phenotypes [18], such as resistance to or immunological control of HIV-1 infection. Analysis of T cells from persons resistant to HIV-1 infection has linked genetic variation in the *CCR5* locus to reduced risk [19]. Similarly, studies of elite controllers (ECs) – a rare (~0.5%) subset of HIV-1 infected individuals who naturally suppress viral replication without combination antiretroviral therapy (cART) [20,21] – have highlighted the importance of specific *HLA-B* variants and enhanced cytotoxic $CD8^+$ T cell responses [22,23]. Although compelling, these findings have proven insufficient

to explain the frequency of viral control in the general population; additional cellular components or interactions could be implicated in coordinating effective host defense. Moreover, these studies have not suggested clinically actionable targets for eliciting an EC-like phenotype in other HIV-1-infected individuals. Other work has demonstrated enhanced crosstalk between the innate and adaptive immune systems of ECs [24-26]. For example, we recently reported that heightened cell-intrinsic responses to HIV-1 in primary myeloid dendritic cells (mDCs) from ECs lead to more effective priming of HIV-1-specific $CD8^+$ T cell responses *in vitro* [25]. Nevertheless, the master regulators driving this mDC functional state, the fraction of EC mDCs that assume it, its biomarkers, and how to potentially enrich for it are unknown.

The recent emergence of single-cell RNA-Seq (scRNA-Seq) affords a direct means of identifying and comprehensively characterizing functionally important subsets of cells and their complex underlying biology. As scRNA-Seq has matured into a mainstream technology, new questions about how to model single-cell variation continue to arise. To date, computational modeling approaches have typically described single-cell heterogeneity as a combination of gene-intrinsic effects (i.e., fundamental molecular noise), and gene-extrinsic ones, capturing both cell-intrinsic features (e.g., differences in intracellular protein levels, epigenetic state, mutation status, extracellular environment) and library-intrinsic technical artifacts (e.g. drop-out effects). Yet, in single-cell studies that utilize samples from across multiple donors (e.g., multiple ECs), these gene-extrinsic sources can be further subdivided into those that are unique to specific donors and those that are shared. The category of donor-dependent variation ranges from donor-specific cell subsets or large differences in cell-type composition to more subtle expression differences in constituent cell-types. If the goal of a study is to generate hypotheses

relating to a common phenotype, such as EC, strategies for prioritizing shared features can benefit from quantitative characterizations of reproducibility across multiple donors. In developing methods to focus on shared characteristics, we address questions of how to identify characteristics that contribute to a particular phenotype by operating at broad biological scales across multiple human individuals with differing genetic backgrounds health histories but a shared phenotype, in this case elite control of HIV.

Here, we apply single-cell RNA-Seq (scRNA-Seq) to evaluate heterogeneity of transcriptional responses of mDCs ($CD14^{-}$, $CD11c^{Hi}$, $HLA\text{-}DR^{+}$) from three EC individuals after *in vitro* exposure to a VSV-G pseudotyped HIV-1 virus or media control. To overcome the potentially confounding effects of donor-dependent biological and technical variation, we propose a broadly applicable strategy that combines reproducibility-based computational analyses with targeted experimentation to resolve, characterize, and modulate common response states across multiple donors (**Figure 5-1**). More specifically, we utilize existing tools for single-cell data analysis, including SCONE [27] and FastProject [28], and implement an IDR (Irreproducible Discovery Rate)-based framework [29] in scRAD (Single-Cell Reproducibility Across Donors; https://github.com/YosefLab/scRAD) to identify reproducible response states, pathways, and biomarkers that are consistently detected after viral exposure across multiple donors who share a unifying classification such as EC. Our analysis reveals remarkable functional heterogeneity among mDCs, described by several discrete transcriptional response states. We discover one reproducible state that displays gene expression features consistent with profound functional activation and heightened antiviral activity. This subset of mDCs, enriched among cells expressing the surface molecules PD-L1 and CD64, is: i) is induced more efficiently in ECs than in HIV-1 chronic progressors (CPs) or healthy donors (HDs) after *in*

vitro viral exposure; ii) associated with both higher $CD4^+$ T cell counts and lower HIV-1 viral loads; iii) more effective at stimulating T cell proliferation *in vitro*; and, iv) more efficient in inducing HIV-1-specific polyfunctional cytotoxic $CD8^+$ T cells – all canonical correlates of antiviral immunity in EC [30]. By leveraging *scRAD* to re-examine publicly-available transcriptomic datasets, we further identify and experimentally investigate key regulatory molecules and adjuvants for modulating the acquisition of this functional mDC response state in the general population, with potential therapeutic and prophylactic implications. Together, our results highlight how single-cell analytic approaches can identify shared drivers of enhanced immunity across a phenotype and empower rational strategies for altering ensemble cellular responses.

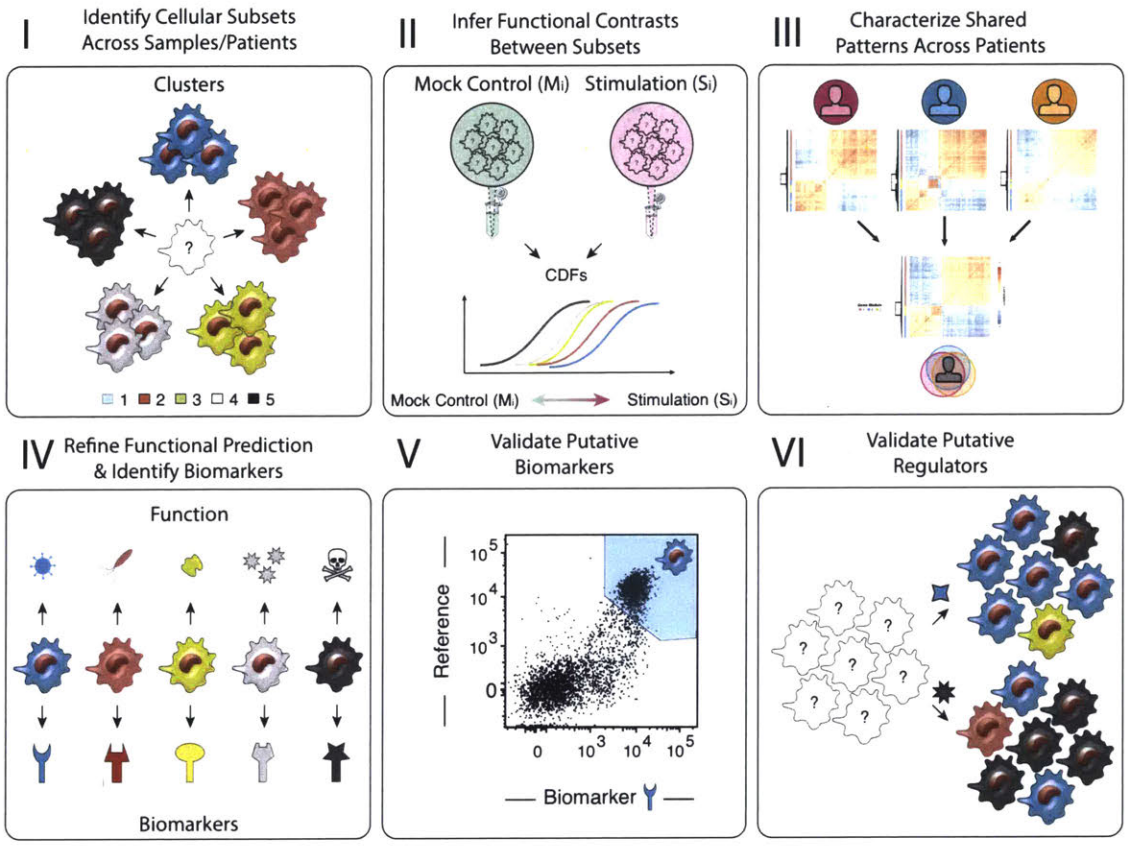

Figure 5-1 | A rational framework used to resolve, characterize and then modulate response states across multiple patient sources. Our generally applicable framework is founded upon the following workflow: (**I**) Resolve the individual mDC subtypes and states that comprise the system under study (**Methods**); (**II**) Define putative functions for each and identify biologically meaningful contrasts using existing databases; (**III**) Characterize patterns of differential expression that are common across patients; (**IV**) nominate potential biomarkers and relevant cellular circuitry based on accumulated knowledge; (**V**) Isolate and characterize interesting subsets; (**VI**) validate inferred regulators.

RESULTS

5.2 Shared EC mDC Subsets Revealed by Single-Cell RNA-Seq

In order to identify shared features of mDC ($CD14^-$, $CD11c^{Hi}$, $HLA\text{-}DR^+$) innate immune responses to HIV-1 across ECs, we performed scRNA-Seq [3,13,14,31,32] on peripheral blood mononuclear cells (PBMCs) from three ECs (p1, p2, p3) exposed *in vitro* to either a VSV-G pseudotyped HIV-1 virus or a media control for 48 hours (**Figure 5-2A**, **Methods**) [33]. Stimulating PBMCs, rather than isolated mDCs, mimics some of the critical physiological interactions that occur between mDCs and other immune cell types *in vivo*, while the use of a VSV-G pseudotyped HIV-1 particles enhances mDC infection efficiency [10]. Given the potential bias of viability sorting, which may discard dying dendritic cells (DCs) undergoing viral stress responses, we opted for *in silico* viability gating rather than sorting on live/dead markers or stains. Following incubation, we sorted single mDCs ($CD14^-$, $CD11c^{Hi}$, $HLA\text{-}DR^+$) into 96-well plates and performed SMART-Seq2-based scRNA-Seq [34]. After estimating gene expression levels, we applied elements of the *SCONE* [27] normalization pipeline to filter out single-cell samples with poor alignment characteristics and normalize the remaining data to minimize the impacts of these characteristics on expression quantification (**Figure 5-3**, **Methods**). Subsequently collected viability-sorted mDC data exhibited only a 2-3 fold gain in the fraction of high-quality cells, suggesting that incubated primary cells from HIV-1 infected patients represent a fragile source material (**Figure 5-3**). In total, we obtained high-quality expression data in 188 virus- and 130 media-exposed cells by sequencing to an average depth of 700,000 reads (**Methods**).

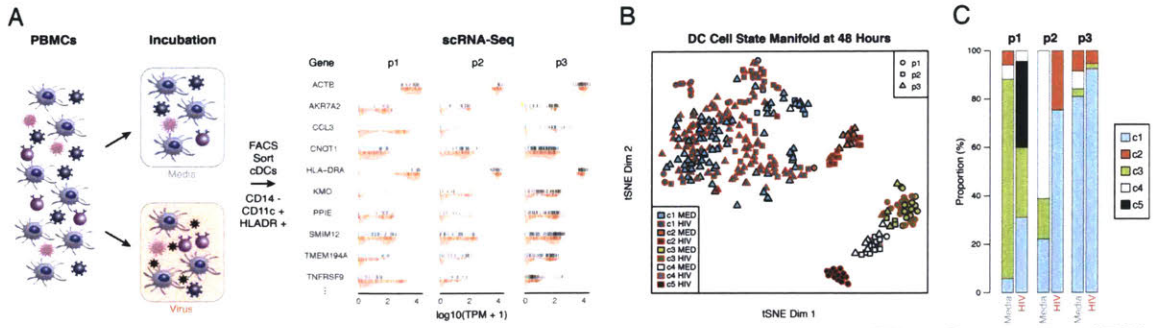

Figure 5-2| scRNA-Seq identifies 5 response clusters among Elite Controller (EC) Myeloid Dendritic Cells (mDCs). (A) Left: Schematic representation of experimental system. After incubation with virus or a media control for 48 hours, mDCs were isolated from PBMCs by FACS and profiled by scRNA-Seq. Right: Violin plots of single-cell expression levels for ten select genes for each EC donor (p1, p2, p3). Vertical lines represent individual cellular values; the upper (grey) half of the violin shows the distribution of values for the media control and the bottom (red) shows the same for virus-exposed cells. **(B)** t-Distributed Stochastic Neighbor Embedding (t-SNE) of all FACS sorted mDCs across three EC subjects passing quality filters (**Methods**; p1: circles, p2: triangles, p3: squares). Virus exposed cells are outlined in red; media exposed cells have no outline. Cells separate into five distinct clusters (c1-5; **Methods**). **(C)** Stacked bar plot depicting the percentage of total mDCs in each cluster for each patient under media and viral exposure conditions.

Dimensional reduction of normalized expression estimates (**Methods)** with tSNE representation illustrates how cells from each of the 3 EC donors span a common expression state-space: cells from different donors often share similar expression profiles, forming mixed clusters. Unsupervised k-medoids clustering revealed five distinct transcriptional response states (clusters 1-5, c1-5; **Figure 5-2B, Methods**), with all but one state (c5) observed in all 3 donors. Linear regression analysis identified a small number of genes exhibiting significant cluster-independent associations with patient and exposure (131 and 14 genes, respectively). On the other hand, the fractional abundance of c1-c4 varied significantly across the three donors and two exposure conditions (**Methods**). Among these, the c1 response state was consistently enriched among virally exposed mDCs (p-value = 8.5×10^{-6}, logistic regression) while c3 and c4 were more

common among media-exposed cells (p-value = 1.3×10^{-4} and 1.1×10^{-5} respectively, logistic regression) (**Figure 5-2C**).

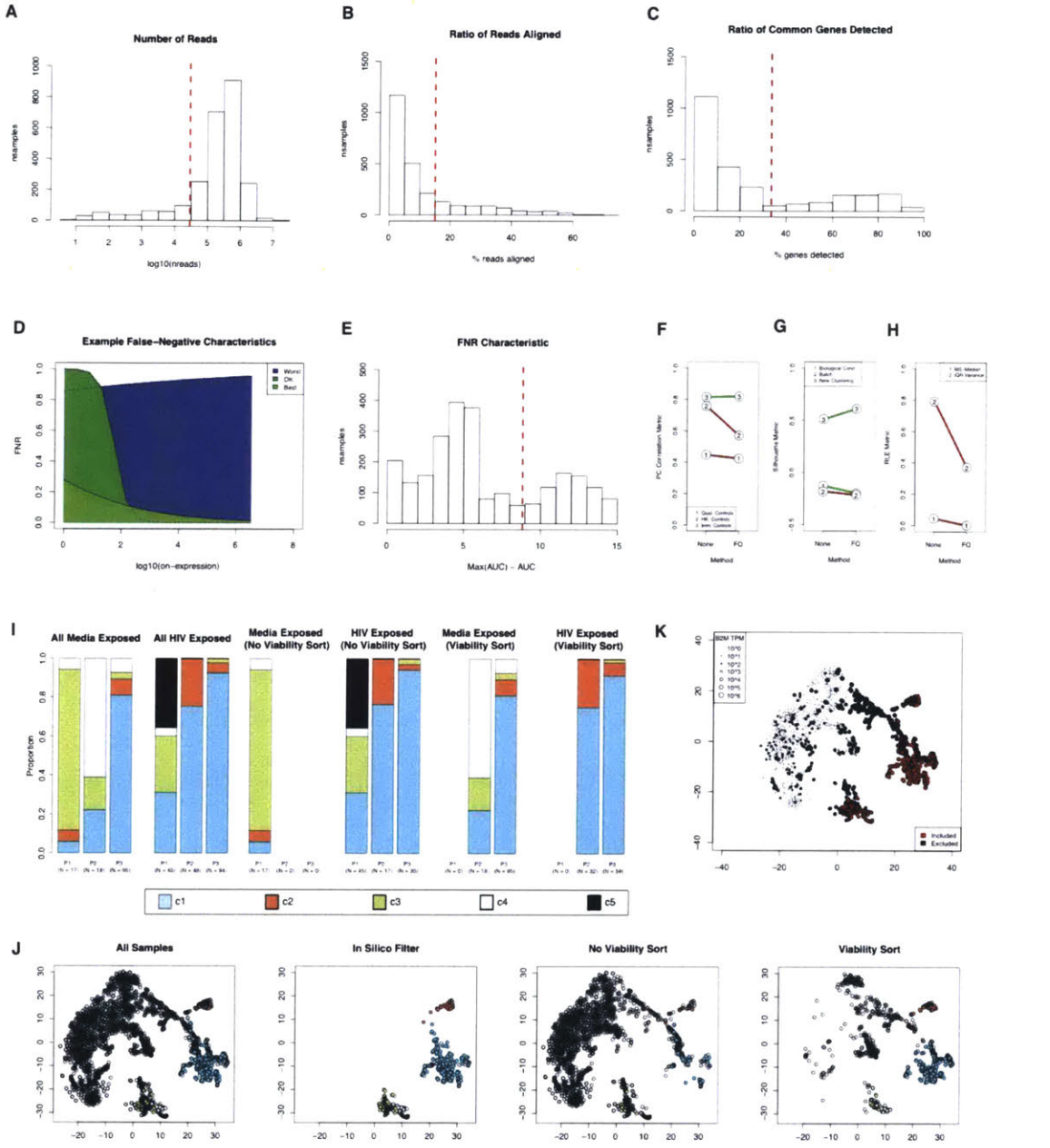

Figure 5-3 | Single-Cell Filtering and Normalization. *Legend next page*

Figure 5-3 | Single-Cell Filtering and Normalization. *Previous page* **(A-C, E)** Distributions of single-cell sample (24 and 48 hours) filtering metrics. Red lines represent adaptive threshold below which all samples (n = 2,489) were removed from further analysis (see **Methods**). **(A)** Distribution of number of paired-end reads per library. **(B)** Distribution of transcriptome read alignment ratio per library. **(C)** Distribution of the fraction of "common genes" detected per library. **(D)** Example false-negative characteristic fits, exhibiting three different relationships between the mean on-expression (TPM) of constitutively expressed genes and their false negative rate (FNR) or "drop-out rate." Colors represent the overall quality of the area under the curve (AUC) (see **Methods**). **(E)** Distribution of fit FNR AUC per library. **(F-H)** Differences in SCONE metrics before and after full-quantile (FQ) normalization (see **Methods**). **(F)** Correlation between the first 3 expression Principal Components (PCs) and the first 3 PCs computed across "negative" controls (Alignment QC Metrics and housekeeping (HK) genes) tend to decrease while correlations with the first 3 PCs across "positive" controls (innate immune system genes) tends to increase. **(G)** The average silhouette width (ASW) of biological condition (patient x exposure x timepoint x viability sort) and the ASW of batch both decrease. However, the ASW of *de-novo* clustering tends to increase. **(H)** The mean sample-median relative log-expression (RLE) decreases, as does the variance of the sample inter-quartile range RLE decrease: both global differential expression and differential expression variability is reduced. (**I**) Stacked bar plots depicting the percentage of total mDCs in each 48h cluster c1-5 (see **Methods**) for each patient under media and viral exposure conditions, stratifying by no viability or viability pre-selection. Both types of single-cell libraries were only obtained from patients p2 and p3. As seen in HIV-1-exposed samples, viability-sorted compositions are comparable to samples without viability sorting. (**J**) tSNE plot of (un-normalized) log(TPM+1) expression, including all cells from 24 hours and 48 hours, HIV and Media exposures, with or without viability gating. Points are colored according to a 48h cell's membership to clusters c1-5. Various subsets are plotted independently, including cells passing in silico cell filter, cells that were not sorted on viability, and cells passing viability sorting. Viability sorting tends to exclude cells from low-quality clusters, enriching the fraction of cells passing quality filtering. (**K**) tSNE plot from (**J**), sizing points according to estimated expression levels of B2M. Red samples passed in silico cell filter. Clusters of cells excluded by filter exhibited very low-levels of the housekeeping transcript.

5.3 Reproducibility-Based Functional Analysis Reveals a Robust Antiviral Signature

To further examine the five EC mDC response states (clusters c 1-5) and their interrelationships, we utilized FastProject [28], a software package for visualization and interpretation of scRNA-seq data with reference to prior biological knowledge (**Methods**). Coherently-varying gene expression signatures identified by FastProject (**Figure 5-4A**) repeatedly implicated c1 and c2, but not c3-c5, as responses associated with elevated DC activation (**Figure 5-4B**). Intriguingly, the transcriptional behavior of c1 mDCs appeared

more consistent with elevated innate antiviral activity, displaying maximal values among signatures for DCs exposed to viruses, such as HIV-1 and Newcastle virus (p-value = $2.5x10^{-9}$, $7.2x10^{-13}$, respectively; two-sided Kolmogorov-Smirnov (K-S) test c1 vs c3; c1, n = 220; c2, n = 26; c3, n = 35; **Figure 5-4B**). In contrast, c2 was specifically distinguished by signatures of DCs stimulated through alternative pathogen associated molecular patterns (PAMPs), such as LPS and R848 (p-value = $8.4x10^{-9}$, $8.6x10^{-11}$, respectively; two-sided K-S test c2 vs c1; c1, n = 220; c2, n = 26; c3, n = 35; **Figure 5-4B**), or by bacteria or parasites. Motivated by the biological relevance of signatures contrasting c1 and c2 against the remaining clusters, we tested for differential expression of each of these two populations against the pool of c3, c4, and c5 cells.

As in most experiments involving non-model organisms, inter-subject biological and technical variability poses a substantial confounding risk by systematically distorting or exaggerating transcriptome-wide differences between groups. To address this, we developed and applied the *differential expression* module of *scRAD*: instead of explicitly modeling donor effects on single-cell expression distributions [35], *scRAD* performs differential expression analysis separately for every donor and then combines the results using IDR meta-analysis [29] (**Methods**). In previous simulation studies, this model-based meta-analysis technique has demonstrated greater discriminative power than other approaches [29]; in our study, it better emphasizes aspects of clustering that are reproduced over multiple donors. In order to partition differentially expressed genes (c1 vs. c3-5, and c2 vs. c3-5) into a common-evidence set from both clusters (c1 and c2) and two cluster-unique sets, we used *scRAD* again, this time performing meta-analysis to aggregate the differential expression results obtained independently for c1 and c2 (**Methods**). This computational approach allows us to identify reproducible differences which are possible

contributors to the EC phenotype on an experiment of this scale, spanning multiple unique individuals with many potential confounding factors.

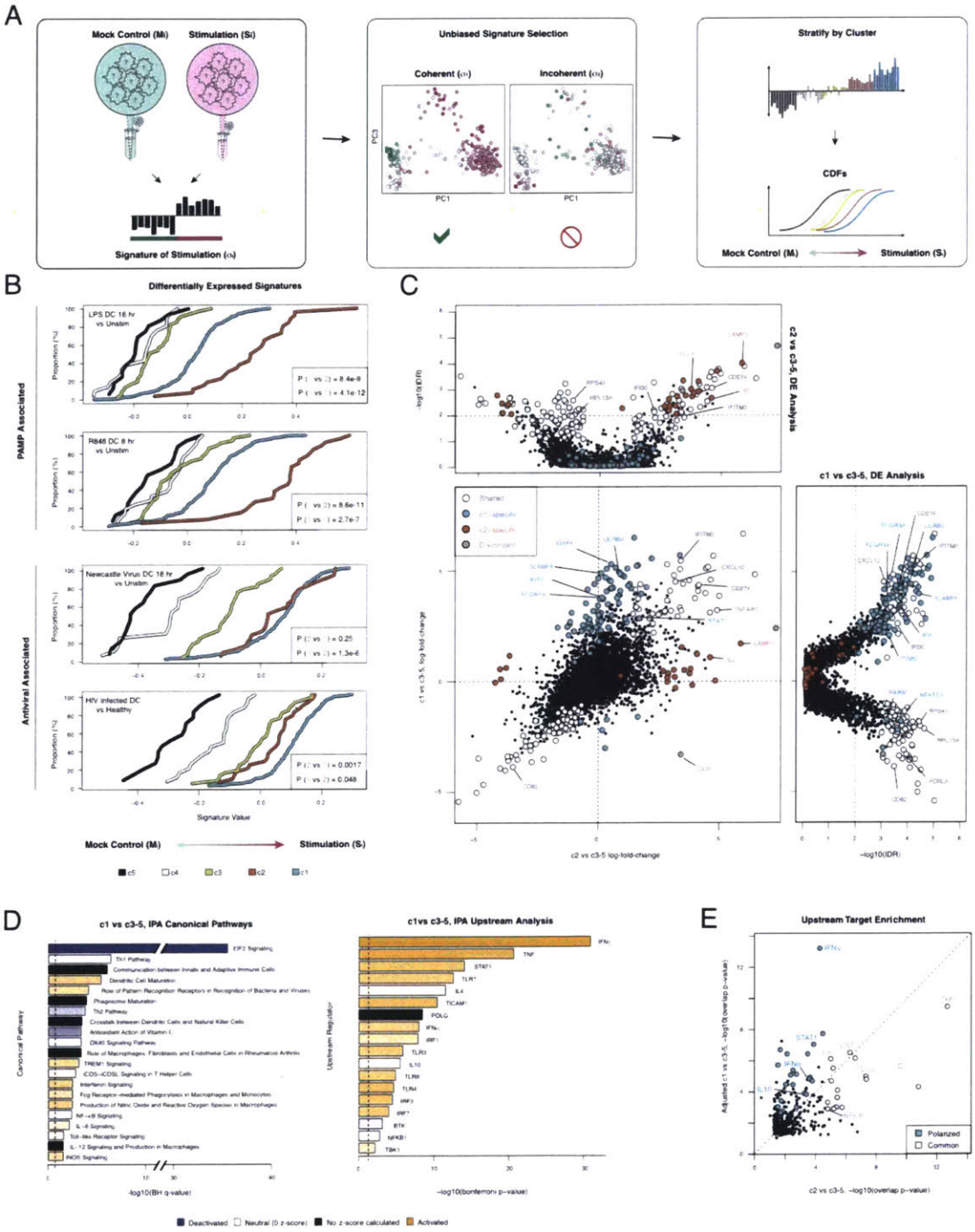

Figure 5-4 | Characterization of transcriptional single-cell response groups.
Legend Next Page

Figure 5-4 | Characterization of transcriptional single-cell response groups.
Previous page
(A) Left: schematic of signature database. The expression of a bulk sample of simulated DCs (S_i) is compared to the expression of a mock control (M_i). Highly ranked up-regulated and down-regulated genes comprise the signature σ_i. Middle: σ_i is applied to all cells in the study, and FastProject identifies pairs of expression data projections and σ_i for which σ_i varies coherently across the projection. Right: Coherent σ_i values are binned by cluster to nominate specific cluster contrasts as biologically meaningful. **(B)** Cumulative distribution function (CDF) comparisons for single cells from each cluster identified with FastProject gene signatures derived from GSE14000 [9], GSE22589 [10], GSE18791 [11], and GSE2706 [17] (see **Methods**). The single-cell signature value quantifies the extent to which each cell is polarized toward a stimulated instead of unstimulated expression state. Clusters with gene expression signatures more closely mapping to the stimulated condition shift right, while clusters characteristic of un-stimulated shift left. Kolmogorov–Smirnov (KS) tests show significant differences in these signatures between the first 3 clusters (c1, n = 220; c2, n = 26; c3, n = 35). **(C)** Potential genes specific for c1 (cyan), c2 (orange), shared between c1 & c2 (white), or inconsistent across individuals (grey). Individual volcano plots of negative log irreproducible discovery rate (IDR) vs mean differential log-expression between clusters c1 and c3-5 (right) and c2 vs c3-5 (left; **Methods**). **(D)** Selected ingenuity Pathway Analysis (IPA) (see **Methods**) results for Canonical Pathways (Benjamini-Hochberg q-value < 0.01) and Upstream Regulators (Bonferroni p-value < 0.05) significantly deactivated (blue), neutral (white: with z score; black: without z score) or activated (orange) in c1 vs c3-5. **(E)** Comparison of putative upstream regulators from IPA for c1 vs c2-5 and c2 vs c3-5 (see **Methods**).

In line with known pathway elements shared between the DC antiviral and bacterial/parasitic response pathways [1,36], we uncovered 121 genes that were commonly up-regulated when comparing either c1 or c2 to c3-5 (**Figure 5-4C**). Additionally, we identified 103 genes that were uniquely called as up- or down-regulated in c1 or c2 relative to the remaining clusters (**Figure 5-4C**). Genes preferentially expressed by c1 over c2 include the interferon-inducible gene *IFIT3*, whereas genes preferentially expressed by c2 encode molecules associated with endocytosis and antigen presentation (e.g., *LAMP3* [37], **Figure 5-4C**), suggesting different levels of activation or polarization between c1 and c2. A targeted analysis of the expression of 28 Interferon-Stimulated Genes (ISGs) regulated by HIV-1 [25,38] suggested that c1 displayed the most potent and coherent interferon-induced

transcriptional signatures (p-value = $2.5x10^{-7}$, two-sided K-S test c1 vs c2; c1, n = 220; c2, n = 26).

Ingenuity Pathway Analysis [15] (IPA) of differentially expressed gene lists revealed that the gene set reproducibly differentiating c1 from c3, 4, 5 is enriched for pathways related to DC maturation (Benjamini-Hochberg (BH) q-value = $4x10^{-6}$), innate recognition of microbes by PRR (q = $8x10^{-5}$), interferon (q = $3x10^{-3}$) and TLR signaling (q = 0.03, **Figure 5-4D)**. These pathway enrichments do not reach significance (q < 0.05) for c2. We partitioned the set of putative upstream regulators predicted by IPA according to "common" or "polarized" activity across c1 and c2 (**Methods)**. Among the polarizing regulators, we observed several molecules associated with antiviral responses uniquely active in c1 (IFNG, IFNA, STAT1). We also saw evidence of specific TLR activation (TLR3, TLR4) for c1 but not c2 (**Figure 5-4D,E**). Overall, these observations suggest that c1 represents a subset of mDCs in an activated viral response state that could potentially drive the effective innate antiviral immune responses observed in bulk mDC from ECs [25].

5.4 Reproducible Biomarker Identification for c1 mDCs

To further study the c1 response state, we sought to identify putative markers for prospectively isolating c1 cells after exposure to HIV-1 across ECs. We developed two reproducibility-based criteria for surface marker candidacy, which have been implemented in the *biomarker selection* module of *scRAD*: 1) the surface marker must be encoded by a transcript that is reproducibly up-regulated in c1 vs c3-5 (IDR < 0.01); and, 2) the

transcript encoding the surface marker should be correlated with sufficiently many genes, in a reproducible manner, across all donors (see **Methods** for additional details).

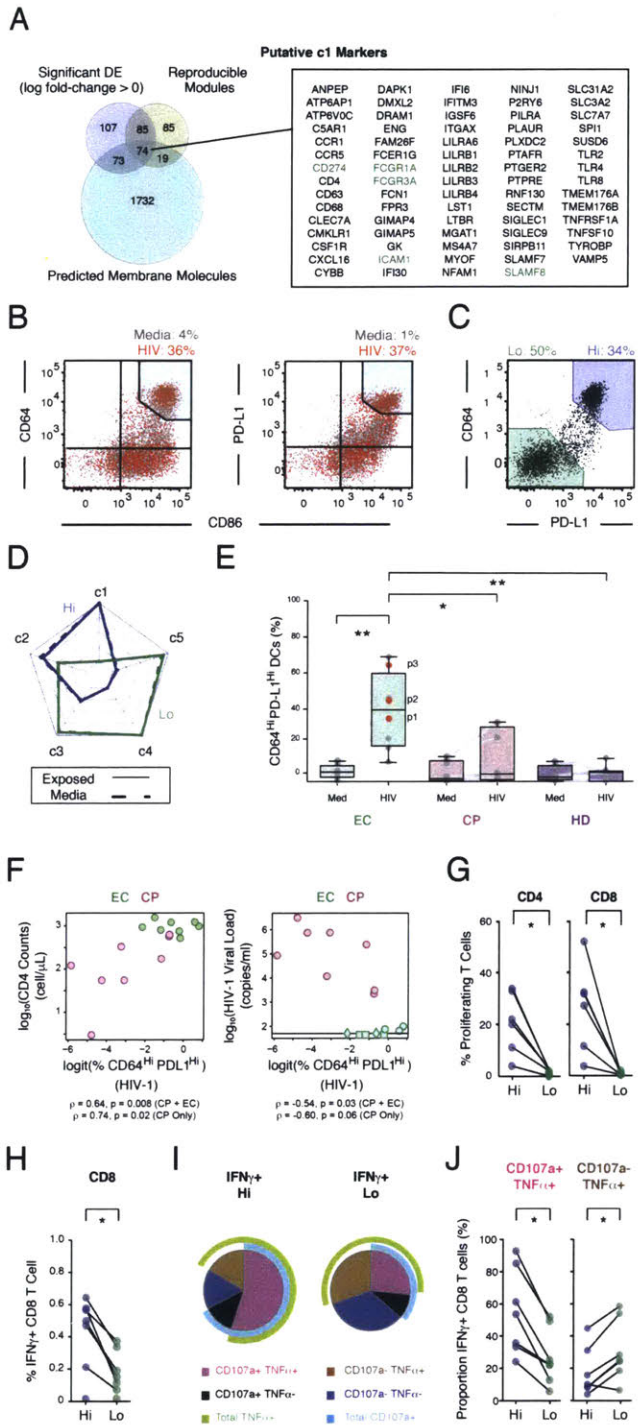

Figure 5-5 | CD64 and PD-L1 enrich in highly functional c1-like mDCs. (A) Selection of c1-specific genes encoding surface proteins for validation as c1 markers. 74 genes (listed in box) were: 1. differentially expressed between c1 and c3-5; 2. reproducibly correlated with other c1 genes across all three ECs profiled; and, 3. predicted membrane proteins (see **Methods**). Candidate markers shown in green were selected for validation by FACS (**Fig 5-4A**). **(B)** Flow cytometry analysis of either CD64 (y-axis, left panel) or PD-L1 (y-axis, right panel) vs CD86 (x-axis) expression in mDCs from EC patient 1 (p1). Numbers above represent the percentage of $CD64^{Hi}$/$PD\text{-}L1^{Hi}$ cells (top right gate; light blue) at 24 hours in media (grey) and VSV-G pseudotyped HIV-1 virus exposure (red) conditions. **(C)** Flow cytometry plots showing analysis of CD64 vs PD-L1 expression on mDCs exposed to VSV-G pseudotyped HIV-1 for 24h, defining 2 populations: $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ (Hi; blue) and $CD64^{Lo}$,$PD\text{-}L1^{Lo}$ (Lo; green). Percentage in each gate is listed above. **(D)** Radar plots (see **Methods**) representing relative similarities of each subset (c1-5) to population-level RNA-Seq data from cells in the Hi and Lo PD-L1,CD64 gates 48h after viral (solid line) or media exposure (dashed line).

Continued next page

Figure 5-5 | CD64 and PD-L1 enrich in highly functional c1-like mDCs. *Continued*
(E) Proportions of $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDCs induced from multiple elite controllers (EC; n = 8), untreated chronic progressors (CP; n = 8) and health donors (HD; n = 7) after 24h of culture in media or VSV-G pseudotyped HIV-1 (*, $p < 0.05$; **, $p < 0.01$; two-tailed Wilcoxon signed-rank test). (**F**) Correlation between the proportions of $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDCs induced in ECs (n = 8) and untreated CPs (n = 8) or just CPs and clinical CD4 T cell count (p-value = $8x10^{-3}$ (two-sided) and $2x10^{-2}$ (one-sided) respectively, Spearman correlation permutation p-value) or between the proportions of $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDCs induced in ECs (n = 8) and untreated CPs (n = 8) or just CPs and HIV-1 viral load (p = $3x10^{-2}$ (two-sided) and $6x10^{-2}$ (one-sided) respectively, Spearman permutation p-value)). Diamond and square points represent indeterminate viral loads of < 20 and < 50 copies/ml respectively. **(G)** Proportion of proliferating $CD4^+$ (left) and $CD8^+$ (right) T cells co-cultured with the Hi and Lo sorted virus-exposed mDCs populations (n = 6 patients). **(H)** Proportion of total $IFN\gamma^+$ $CD8^+$ T cells cultured with the Hi and Lo sorted virus-exposed mDCs populations (n = 7 patients). Statistical significance for **(G,H)** were evaluated using a two-tailed Wilcoxon matched pairs signed-rank test (*, $p < 0.05$). **(I)** Pie chart generated with data from n=7 patients showing CD107a and $TNF\alpha$ expression on $CD8^+$ T cells cultured with Hi (left) or Lo (right) mDCs. **(J)** scatter plots of proportions of $CD107a^+$, $TNF\alpha^+$ (left) and $CD107a^+$, $TNF\alpha^-$ (right) $CD8^+$ T cells cultured with Hi and Lo mDCs. Statistical significance was evaluated using a two-tailed Wilcoxon matched pairs signed-rank test, n = 7 patients (*, $p < 0.05$).

Using this procedure, we obtained a list of 74 candidate c1 mDC markers (**Figure 5-5A**). Based on antibody availability, we selected five proteins (*FCGR3, FCGR1, CD274, ICAM1, SLAMF8*) to profile 24h after infection with pseudotyped HIV-1 by flow cytometry in $CD14^-$ $CD11c^{Hi}$ $HLADR^+$ DCs from our cultures (**Figure 5-5B**). Among these five candidate markers, CD64 (*FCGR1A*) and PD-L1 (*CD274*) exhibited the most dramatic and consistent virus-induced upregulation among $CD14^-$, $CD11c^{Hi}$, $HLA\text{-}DR^+$ mDCs isolated from the PBMCs of the 3 ECs we previously characterized by scRNA-Seq, as well as those mDCs from five additional EC donors (**Figure 5-5B;** p-value = $7.8x10^{-3}$; two-tailed Wilcoxon matched-pairs signed rank test; n = 8). CD64 is an Fc-receptor for IgG [39], while PD-L1 has been implicated in mediating the balance between T cell activation and immunopathology, as well as immediate effector differentiation and long-term memory formation in T cells [40]. Importantly, high expression of PD-L1 has also been found on tolerogenic murine mDCs in chronic LCMV infection [41] and in inflammatory lymph node-resident mDCs from HIV-1 infected individuals [42]. Nevertheless, high expression of IFN

and inflammatory cytokines identified in our pathway analysis of c1 and high CD86 expression levels on CD64^{Hi} and PD-L1^{Hi} cells, indicate that these cells are highly activated inflammatory DCs (**Figure 5-4**).

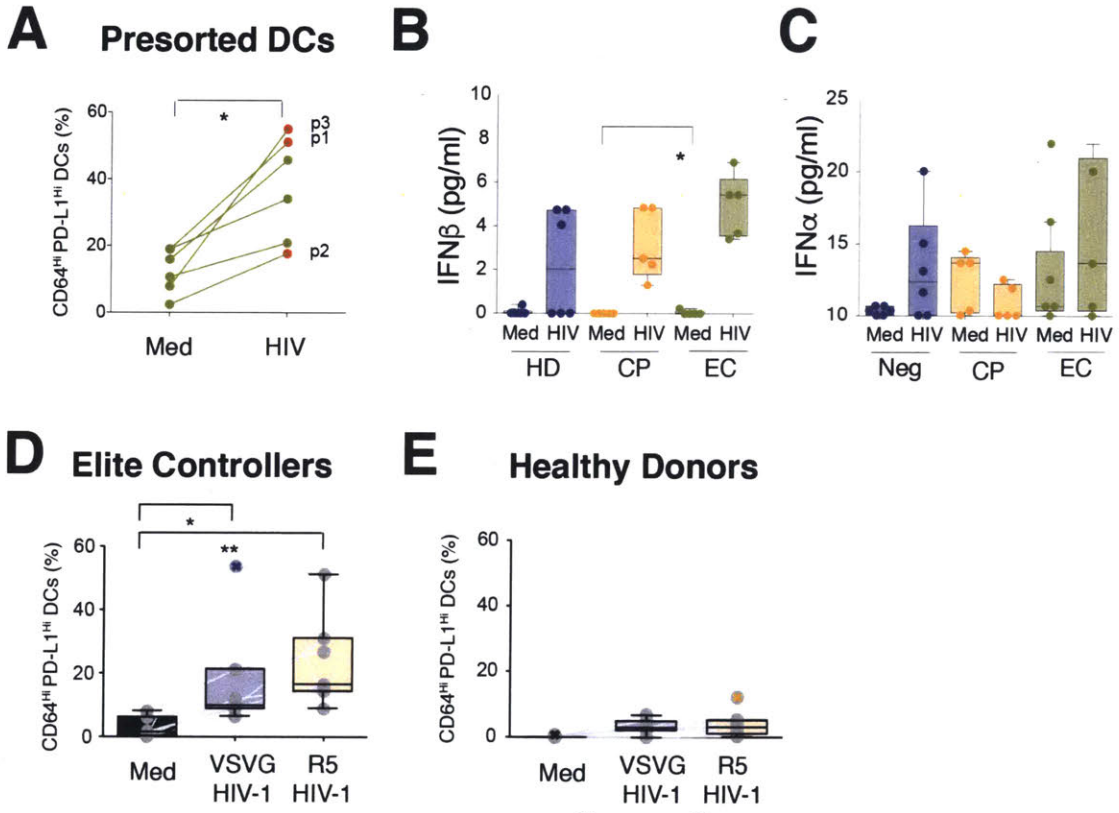

Figure 5-6 | Induction of c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC in response to different HIV-1 strains and culture conditions. **(A)** Proportions of CD64^{Hi},PD-L1^{Hi} cells detected in mDCs sorted prior to culture in the presence of media and VSV-G pseudotyped HIV-1 virus. Statistical significance was calculated using a two-tailed Wilcoxon test (*, $p < 0.05$; $n = 6$). **(B)** ELISA analysis of IFN beta protein levels present in culture supernatants of mDCs from healthy donors (Neg, blue; n=6), chronic progressors (CP, orange; n=5) and elite controllers (EC, green, n=5) exposed for 48h to either media (Med) or VSVG pseudotyped HIV-1 virus (HIV). mDCs were presorted from PBMC prior to in vitro culture. *p=0.0397, One tailed Mann Whitney test. **(C)** Luminex analysis of IFN alpha protein levels present in culture supernatants of mDCs from healthy donors (Neg, blue; n=6), chronic progressors (CP, orange; n=5) and elite

When we analyzed mDCs based on surface expression levels of CD64 and PD-L1, we observed two dominant mDC populations after viral stimulation: one CD64^{Hi},PD-L1^{Hi} and the other CD64^{Lo},PD-L1^{Lo} (**Figure 5-5C**). Population-level transcriptional profiling of mDCs

sorted on CD64^{Hi},PD-L1^{Hi} at both 24 and 48h post-viral stimulation revealed gene expression profiles dominated by the signature of the c1 and, to a lesser extent, c2 response states. In combination with the observation that mDCs sorted on CD64^{Lo},PD-L1^{Lo} matched a mixture of c3-5 (**Figure 5-5D**), we concluded that sorting on CD64 and PD-L1 co-expression enriches for c1 cells. While these two markers are predominantly associated with c1 responses, we note that they are not necessarily causally involved in inducing either phenotype. In line with the single-cell observations above, the c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC phenotype observed in EC could be effectively induced in mDCs alone (without supporting PBMCs) exposed to VSV-G pseudotyped HIV-1 virus, indicating that generation of the CD64^{Hi},PD-L1^{Hi} mDC phenotype does not require paracrine signals from neighboring non-mDCs (**Figure 5-6A**). Collectively, these findings suggest that c1 mDCs might have the potential to drive enhanced antiviral antigen presentation relevant to control of HIV-1 infection.

5.5 Functional Characterization of c1 mDCs

Given the ties between strong antiviral activation and immune control of HIV-1, we naturally wondered whether the CD64^{Hi},PD-L1^{Hi} mDC phenotype, common to ECs, was unique to EC individuals and might be linked to common features of immune control against HIV-1 within the EC phenotype. While the CD64^{Hi},PD-L1^{Hi} mDC phenotype was consistently and efficiently induced in HIV-1 exposed mDCs from ECs, generation of CD64^{Hi},PD-L1^{Hi} mDC was also observed in HIV-1 exposed mDCs from chronic progressors (CP) and healthy donors (HD), although at markedly lower proportions (**Figure 5-5E;** n = 8 per group). Consistent with the more effective induction of this phenotype in ECs, we found higher levels of type-I IFN present in culture supernatants

from pre-isolated DCs exposed to HIV-1 from these patients as compared to alternative cohorts (**Figure 5-6B,C**). Notably, these cohort-intrinsic differences were also observed when mDCs were exposed to a more physiological CCR5-tropic HIV-1 viral strain (**Figure 5-6D,E**), suggesting that this phenomenon is not restricted to VSV-G pseudotyped HIV-1 strains. Correlating the fractional abundance of $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDCs after HIV-1 exposure against clinical phenotypes, we observed a significant positive association with $CD4^+$ T cell count across both CPs (one-sided) and ECs+CPs (two-sided; p-value = $2x10^{-2}$ and $8x10^{-3}$ respectively, Spearman correlation permutation p-value). Plasma HIV-1 viral loads, meanwhile, were negatively associated with percentages of $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDCs across all patients (p-value = $3x10^{-2}$, Spearman correlation two-sided permutation p-value), with borderline-significant association in CPs alone (p-value = $6x10^{-2}$, Spearman correlation permutation one-sided p-value, **Figure 5-5F**). These associations show that a patient's $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDC fraction after viral stimulation tracks traditional biomarkers along a spectrum of HIV-1 control, suggesting that the ability to induce c1-like mDCs might be a useful biomarker of enhanced protective immune responses against HIV-1.

We next sought to directly probe the association between the induction of c1 responses and the enhanced functionality observed in bulk mDCs from EC. We first examined the putative enhanced antigen presentation and T cell activation abilities of the c1-like $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ subset of mDCs by performing mixed leukocyte reactions to compare our CD64,PD-L1 high and low mDC subpopulations (**Methods**). In these experiments, the c1-enriched/$CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDC population demonstrated superior ability to stimulate $CD4^+$ and $CD8^+$ T cell proliferation relative to $CD64^{Lo}$,$PD\text{-}L1^{Lo}$ mDCs across multiple ECs (**Figure 5-5G;** p-value = $1.6x10^{-2}$ and p-value = $3.1x10^{-2}$, respectively; two-tailed Wilcoxon matched-pairs signed rank test; n = 6). Similar results were observed in assays conducted

with T cells from ECs, where CD64^{Hi},PD-L1^{Hi} mDCs were capable of efficiently stimulating the production of IFNγ in a significantly higher proportion of autologous CD8^{+} T cells as compared to CD64^{Lo},PD-L1^{Lo} mDCs (**Figure 5-5H;** p-value = $3x10^{-2}$; two-tailed Wilcoxon matched-pairs signed rank test; n = 5). Further, IFNγ^{+} CD8^{+} T cells primed in the presence of c1-enriched/CD64^{Hi},PD-L1^{Hi} mDCs expressed significantly higher levels of both the degranulation markers CD107a and TNFα (**Figure 5-5I,J;** p-value = $1.5x10^{-2}$; two-tailed Wilcoxon matched-pairs signed rank test; n = 7), mirroring the polyfunctional CTL responses observed in ECs [22,23].

5.6 Signature Meta-Analysis of Candidate Adjuvants for c1 mDCs

Given the possible therapeutic and prophylactic potential of c1-like DCs for studies in non-EC populations with less efficient responses to *in vitro* viral stimulation (**Figure 5-5E**), we next sought to uncover the common signaling pathways involved in the acquisition of the c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC phenotype so that we might engineer its frequency. IPA results for c1 had highlighted several signatures of human DC stimulation, including multiple components of several TLR signaling pathways (**Figure 5-4D**), thus we aimed to compare our single-cell expression profiles to perturbed bulk expression data in order to determine which TLR pathways were most compatible with the c1 signature vs c3-5.

We define, for every cell and every TLR ligand we tested (**Methods**) a "stimulation score", which reflects the similarity between the cell's transcriptional profile and the one induced by the ligand (using weighted correlation; see **Methods**). We then score each ligand by the extent to which its respective stimulation scores in c1 cells are higher than in clusters

c3-5 (using a KW test). Finally, using the *differential signature analysis* module in *scRAD*, we combine the resulting p-values across donors. Notably, for this analysis we used the Stouffer-Z p-value combination method (**Figure 5-7A, Methods**) since the number of hypotheses (i.e., TLR ligands) is small, leading to instabilities in the IDR inference.

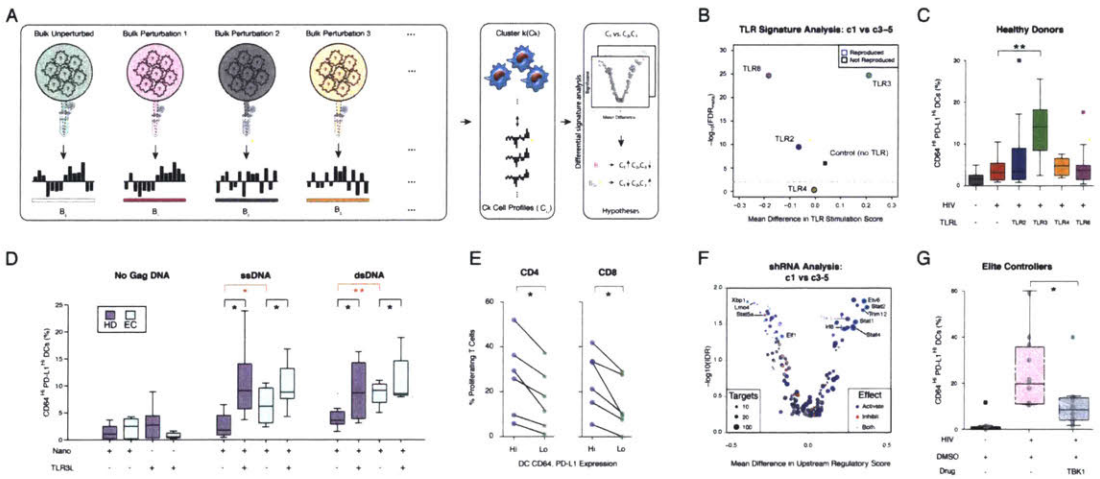

Figure 5-7 | Immunomodulators can alter the fractional abundance of the c1 mDC phenotype. (A) Top: Schematic of bulk expression data (B_i) from publicly available perturbation data. Bottom: Each cell's expression profile (C_{1j}) is correlated with all B_i so as to compare similarities of the single-cell cluster 1 to all bulk expression profiles. **(B)** Volcano plot of negative log meta-analysis false discovery rate (FDR) vs mean difference in "TLR stimulation score" between c1 and c3-5. Scores are computed from weighted correlations between single-cell profiles and transcriptional patterns from human DCs (see **Methods**) after 48h of stimulation with media control (black) or agonists for either TLR2 (PAM3CSK4, dark blue), TLR3 (Poly I:C, green), TLR4 (LPS, orange), TLR7/8 (Gard, purple), or TLR9 (CpG, light blue). Tests reproduced with FDR < 0.01 in both stratified analyses are highlighted in blue. **(C)** Proportion of $CD64^{Hi}$,$PDL1^{Hi}$ cells among mDCs from PBMCs isolated from HIV-negative individuals cultured in the absence or the presence of VSV-G pseudotyped HIV-1, alone or in combination with TLR ligands (TLRL: TLR2L, PGNA, n = 11; TLR3L, Poly I:C, n = 11; TLR4L, LPS, n = 8; TLR8L, CL097, n = 11; **Methods**). Statistical significance was calculated using Kruskal-Wallis and Dunn's tests (**, $p < 0.01$). **(D)** Proportions of $CD64^{Hi}$, $PD\text{-}L1^{Hi}$ cells among mDCs from healthy individuals (indigo) and elite controllers (olive) cultured in the absence or the presence of Poly I:C and polymer nanoparticles loaded with single-stranded (ss) or double stranded (ds) 100 nucleotide HIV-1 DNA (**Methods**; n = 8, HIV negative individuals; n = 7, ECs). Statistical significance was calculated using either two-tailed Wilcoxon signed-rank test (black) or two-tailed Mann-Whiney test (red) to compare differences within or among patient groups, respectively (**, $p < 0.01$; *, $p < 0.05$). **(E)** Proportion of proliferating $CD4^+$ or $CD8^+$ T cells after culture with Hi or Lo mDC from a healthy donor stimulated with TLRL3 and nanoparticles containing gag ssDNA (*, $p < 0.05$; two-tailed Wilcoxon signed-rank test. n = 6).

Continued next page

Figure 5-7 | Immunomodulators can alter the fractional abundance of the c1 mDC phenotype. *Continued* **(F)** Volcano plot of negative log irreproducible discovery rate (IDR) vs mean difference in upstream regulatory score between c1 and c3-5 based on single-cell correlations with shRNA-perturbation profiles from mouse DCs stimulated with LPS for 6h (adapted from Chevrier *et al, Cell*, 2011[1]; see **Methods**). The net effect (activate, inhibit, both) of each perturbation is denoted by color (red, blue, grey; respectively), as is its breadth (size). **(G)** Proportions of CD64^{Hi},PD-L1^{Hi} cells among EC mDCs cultured in the presence or absence of virus and DMSO (control, magenta) or BX795 TBK1 inhibitor (cyan; n = 10; **Methods**). Statistical significance was calculated using a two-tailed Wilcoxon signed-rank test (*, p < 0.05).

Our meta-analysis showed that c1 cells correlated most positively with TLR3 stimulation via Poly I:C compared to the c3-5 (FDR < 0.01; **Figure 5-7B**), generating the actionable hypothesis that triggering the endosomal dsRNA sensor TLR3 might selectively activate downstream pathways that synergize with innate viral sensing mechanisms to increase the fraction of mDCs maturing towards a c1-enriched/CD64^{Hi},PD-L1^{Hi} phenotype (**Figure 5-8**). Analyses of microarray data from mouse DCs stimulated with a comprehensive panel of TLR ligands also suggested that the c1 state most strongly positively correlated with TLR3 activation [36]. To directly test this hypothesis, we incubated PBMCs from several healthy donors (n = 7) – that do not spontaneously generate large proportions of c1-enriched/CD64^{Hi},PD-L1^{Hi} cells *in vitro* in the presence of VSV-G pseudotyped HIV-1 (**Figure 5-5E**) – with virus and different TLR agonists for 24 hours. In contrast to the other TLR ligands tested, we observed that co-incubation of mDCs with virus and Poly I:C led to a significant increase in the proportion of c1-enriched/CD64^{Hi},PD-L1^{Hi} mDCs in PBMCs from healthy individuals (TLR3L: p-value = 0.0091, n = 11; Kruskal-Wallis and post-hoc Dunn's test; TLR2L, TLR4L, and TLR8L, not significant; n = 11, 8, 11, respectively) (**Figure 5-7C**). Meanwhile, in ECs, a TLR3, but not a TLR4, inhibitor had a modest, but significant, effect on the acquisition of the c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC phenotype (p-value = 3.9×10^{-3}; two-tailed Wilcoxon signed-rank test (**, p < 0.01; n = 9).

To explore the generality and therapeutic applicability of our adjuvant strategy, we next examined whether we could couple the same TLR3 activation with direct DNA-based targeting of the cytosolic innate immune recognition machinery that senses viral DNA products [43] rather than use the virus itself. To address this, we incubated PBMCs from healthy donors or ECs simultaneously with a TLR3 agonist (Poly I:C) and single- or double-stranded HIV-1 Gag DNA (ssDNA or dsDNA, respectively) encapsulated in polymeric nanoparticles (**Methods**). A similar delivery vehicle has previously been shown to selectively activate cGAS- and STING-dependent immune recognition pathways, which are involved in innate immune sensing of HIV-1 during natural infection [44]. When we analyzed the fraction of mDCs differentiating into c1-enriched/CD64^{Hi},PD-L1^{Hi} cells, we found that activation with either ss/dsDNA or Poly I:C (TLR3 agonist) alone in PBMCs from healthy donors was less efficient at inducing c1-enriched responses (p-value = $7x10^{-2}$, nano vs Poly I:C alone; p-value = $5x10^{-2}$, nano vs ssDNA; p-value = $1x10^{-2}$, nano vs dsDNA; two-tailed Wilcoxon matched-pairs signed rank test; n = 8; **Figure 5-7D**, comparisons not highlighted). Combining both stimuli, however, significantly increased the proportion of c1-enriched/CD64^{Hi},PD-L1^{Hi} mDCs in PBMCs isolated from healthy donors (p-value = $1.6x10^{-2}$ and p-value = $3.1x10^{-2}$ for ss- and dsDNA, respectively; two-tailed Wilcoxon matched-pairs signed rank test; n = 8; **Figure 5-7D**). Similar results were obtained with cells from ECs (p-value = 0.0469 for both ss- and dsDNA; two-tailed Wilcoxon matched-pairs signed rank test; n = 7; **Figure 5-7D**), with the notable exception that, in ECs, exposure to dsDNA alone led to significantly higher levels of c1-like/CD64^{Hi},PD-L1^{Hi} mDCs relative to cells cultured only in media (p-value = $3x10^{-2}$; Wilcoxon matched-pairs signed rank test; n = 7; **Figure 5-7D**, comparison not highlighted), suggesting a heightened baseline predisposition of EC to respond to intracellular DNA. In

mixed leukocyte reactions, the $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDCs generated from healthy donors incubated with TLRL3 and nanoparticles containing gag dsDNA stimulated greater proliferation in $CD4^+$ and $CD8^+$ T cells compared to the $CD64^{Lo}$,$PD\text{-}L1^{Lo}$ mDCs from the same assay (p-value = $3.5x10^{-2}$ and p-value = $3.1x10^{-2}$, respectively; two-tailed Wilcoxon signed-rank test; n = 6), suggesting that adjuvant induced $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDCs in healthy donors are highly functional antigen presenting cells like their EC counterparts (**Figure 5-7E**).

5.7 Reproducible Differential Signature Analysis Reveals Immunomodulators of c1 mDCs

To identify additional nodes for rationally modulating the acquisition of the c1 functional state, as well as to examine the general applicability of the IDR-framework for uncovering putative regulators of c1's (or any other state's) induction, we again applied the differential signature module of *scRAD* (see **scRAD Vignette**); in this instance, due to limited public availability of human perturbation data, we turned to a published data set of the transcriptional effect of ~200 transcription factor and signaling molecule perturbations in LPS-stimulated mouse DCs that are highly conserved with humans [1,36]. We ranked the perturbations by the degree to which they reproducibly favored the generation of one or more (here, c1) responses over others (here, c3-5; **Methods**). Unlike in the TLR analysis, here we had a sufficient number of hypotheses, and therefore utilized *scRAD*'s core IDR-based functionality [29]. The resulting meta-analysis nominated several putative regulators for modulating the fractional abundance of c1 mDCs in response to a virus or virus-like stimulation (**Figure 5-7F**). Among our top positive regulators of c1 was TBK1, a recognized signal mediator that is activated downstream of multiple innate immune sensing pathways at the convergence of the organelle-associated adaptors MAVS, TRIF

(downstream effector of TLR3, TLR4), and STING (effector of the intracellular DNA sensor cGAS) [45-47], some of which were previously detected in our IPA Upstream Analysis (**Figure 5-4D**). Notably, the cGAS-STING pathway is known to play a key role in the recognition of cytoplasmic HIV-1 DNA in myeloid cells, including those from ECs [25,43], and cGAS itself (*MB21D1*) was up-regulated in c1 cells (LFC = 1.9, IDR < 0.05). To evaluate whether signaling through TBK1 significantly contributes to the maturation of mDCs into the c1-enriched/CD64^{Hi},PD-L1^{Hi} subset in ECs, we added BX795, a TBK-1 antagonist, to PBMCs from ECs at the time of viral addition and examined the impact on mDC responses (**Methods**). As shown in **Figure 5-7G**, inhibition of TBK1 during viral exposure led to a dramatic and significant abrogation of the induction of the c1-enriched/CD64^{Hi},PD-L1^{Hi} mDC population in ECs (p-value = 2.0×10^{-3}; two-tailed Wilcoxon signed-rank test; n = 10), suggesting that TBK1 is a key driver of the acquisition of the c1 phenotype in mDCs and validating the promise of our computational framework.

5.8 Discussion

In summary, by identifying common responses across multiple unique human donors, we can identify shared features which may contribute to the arising or behavior of the uniting phenotype of interest. Here, we studied elite immune control of HIV-1 infection as an example of enhanced immunity phenotype, and develop and apply a reproducibility-based framework that distinguishes gene expression features shared across EC donors. In doing so, we identify a highly functional CD64^{Hi},PD-L1^{Hi} mDC response state that is primed to drive adaptive immunity - a previously unrecognized correlate of effective antiviral response against viral stimuli. Extending and developing computational approaches to hypothesize reproducible biomarkers and upstream regulators, we have realized a

rational, extendable framework for modulating the relative abundance of this state. These tools, provided as part of our R package, *scRAD*, can be applied to a wide variety of common scRNA-Seq analyses and derive robustness from a reliance on multiple donors. An important feature of the IDR framework [29] is that it is based on rank transformed data rather than the original signal (e.g., p-values); this facilitates the statistical analysis of reproducibility in any ranked set of hypotheses, beyond the three analysis modules presented here (differential expression, biomarkers, and upstream regulators).

Importantly, our study demonstrates a clear association between the ability of ECs to efficiently acquire the $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ mDC phenotype *in vitro* and clinical parameters of immunological control of HIV-1 infection. This suggests that an increased ability to induce the c1 transcriptional programs in mDCs might be indicative of beneficial immune responses associated with control of HIV-1 replication in ECs. An important limitation of our study design is that it only establishes associations, rather than causal relationships, between our observations and clinical and immunological parameters—i.e., it does not directly demonstrate a role for mDC in driving or promoting immune control of HIV-1 infection in ECs *in vivo*. Future studies will also be needed to directly examine the role of c1 DCs in other lymphoid tissues, such as lymph nodes, since our current work focused on PBMCs. Were this to prove true, our adjuvanting and perturbation experiments suggest exciting therapeutic possibilities for non-ECs via co-stimulation of TLR and DNA sensor agonists, and perhaps TBK1 directly. Intriguingly, high expression of PD-L1 has also been described on a subset of lymph node-resident mDCs from HIV-1 infected individuals spanning a range of viral loads [42]. While this study proposes that the lymph node resident $PD\text{-}L1^{+}$ DC subset may dampen immune responses based on PD-L1 expression, as CD64 co-expression was not measured, the relationship of this state to c1 remains unknown.

PD-L1 has also been associated with an alternative, tolerogenic IL-10-producing mDC population induced under long term and chronic infection settings in mice [41] – this state is fundamentally distinct from the highly activated CD64Hi PD-L1Hi DC subset identified in our study which is characterized by expression of multiple inflammatory molecules (**Figure 5-4**), high levels of activating costimulatory molecules (**Figure 5-5**), and efficiently induces T cell proliferation and polyfunctionality. In general, the putative functional differences between the two states highlights the importance of surveying complete extra- and intra-cellular states in ascribing function, given potential redundancy. A critical limitation is the lack of an equivalent *in vivo* system where a direct and causal relationship between mDC responses and the induction of protective HIV-1-specific adaptive immunity can be safely and ethically tested; nevertheless, similar principles may inform future experiments performed with other viruses or virus-like elements (e.g., in a vaccine) in both humans and other organisms.

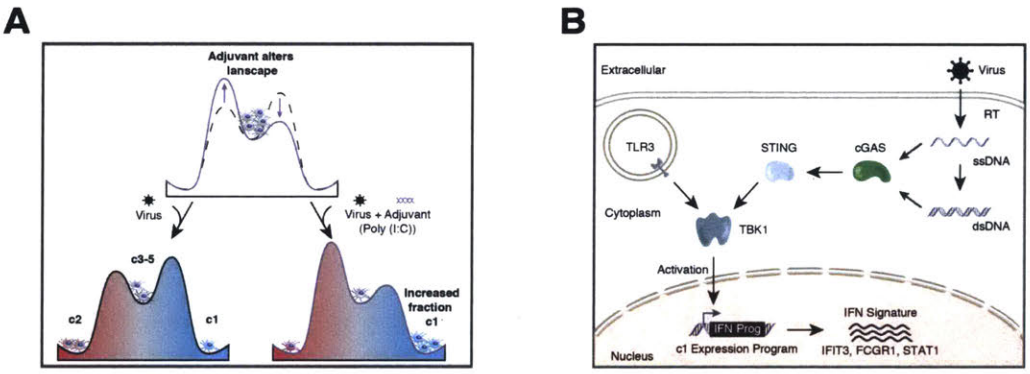

Figure 5-8 | Unifying Model of Results. (A) Potential-energy diagram conceptualizing how adjuvants and other perturbations alter the percentage of mDCs that enter the c1-5 response states upon viral or viral-like exposure. **(B)** Network diagram depicting tested nodes implicated in the c1 mDC response.

Mechanistically, further investigation will be required to identify what biases mDCs from ECs to respond at higher frequency with a c1-enriched phenotype. Given our adjuvanting and perturbation experiments, this enhanced antiviral response capacity could derive from variations in the basal abundance of different DC subsets which, in turn, each have an unequal propensities to generate the c1 responses to nucleic acids; it could similarly derive from dissimilarities in the intrinsic response properties of one or more progenitor or terminally differentiated states, informed by a combination of EC-specific epigenetic modifications and/or complex sets of genetic variants. Since our experimental validations support an inferred role for TLR3 in synergizing with cytosolic viral recognition machinery to induce a TBK1-dependent c1-enriched/CD64^{Hi},PD-L1^{Hi} response, we propose that simultaneous induction of DNA and dsRNA sensing through the cGAS-STING [25,45] and TLR3 pathways might potentiate (**Figure 5-8A**) the maturation (or selective survival) [48] of c1-enriched/CD64^{Hi},PD-L1^{Hi} cells by converging on TBK1 (**Figure 5-8B**), and that these elements might be a natural nexus to explore for EC-specific molecule features. Still, our work demonstrates the potential of scRNA-Seq to discover, genome-wide, functional cellular immune response states, associated markers, and shifts in abundance that may inform the overall efficacy of host immunity.

5.9 Methods

Study Participants

HIV-1 elite controllers (ECs) who had maintained < 2,000 copies/ml HIV-1 Viral Load (VL: 20-98 copies/ml, median 48 copies/ml) for a median of 5 years (range 2-14) in the absence of antiretroviral therapy (EC; $CD4^+$ T cell counts: 515 - 1,543 cells/ml, median 909 cells/ml; n = 8 persons), untreated chronic progressors (CP; VL: 2,190-3,117,608 copies/ml, median 162,807 copies/ml; $CD4^+$ T cell counts: 3 - 623 cells/ml, median 146.5 cells/ml; n = 8 persons), and HIV-1 seronegative healthy donors (Neg, HD; n = 7 persons), were recruited for this study. All subjects gave written informed consent; the Institutional Review Board of Massachusetts General Hospital/Partners Healthcare approved the study protocol.

***In Vitro* Infection with HIV-1 Virus**

Freshly isolated PBMCs were infected with GFP-encoding vesicular stomatitis virus G envelop (VSV-G) pseudotyped or R5-tropic HIV-1 virus (Multiplicity of infection; MOI = 2.4 and 0.4, respectively), kindly provided by Dr. Dan Littman (New York University, New York, New York, USA), for 2 hours at 37°C. 24 and 48h post-infection, $CD14^-$,$CD11c^{Hi}$,$HLADR^+$ mDC were singly sorted (see **Flow Cytometric Analysis and Sorting**) from total PBMC into 96 well plates containing lysis buffer for scRNA-Seq as previously described[49] **(Figure 5-2A)**. In some experiments, sorted $CD14^-$ $CD11c^{Hi}$ $HLADR^+$ mDCs were presorted prior to *ex-vivo* infection with VSV-G pseudotyped HIV-1.

TLR Perturbations

In the TLR agonist experiments, mDCs from PBMCs (see **Flow Cytometric Analysis and Sorting**) were treated with HIV-1 alone or HIV-1 in combination with 2 μg/ml of a TLR2 (PGNA), TLR3 (Poly I:C), TLR4 (LPS), or TLR8 (CL097) ligand (InvivoGen, San Diego, CA) **(Figure 5-7C)**. In the TLR antagonist studies, mDCs from PBMCs were treated with VSV-G pseudotyped HIV-1 (see ***In Vitro* Infection with HIV-1 Virus**) alone or in combination with a TLR3 (CUCPT4A, 60 nM, Tocris), TLR4 (600 ng/ml, LPS-RS, InvivoGen), or TBK-1 inhibitor (BX795, 1μM, InvivoGen) **(Figure 5-7B,G)**.

For our single stranded (ss) and double stranded (ds) DNA stimulation experiments **(Figure 5-7D)**, mDCs from either healthy donors (HD, n = 8) or ECs (n = 7) were cultured for 24h in the presence of Poly I:C and 2 μg/ml of either ss- or ds-Gag DNA[44] that had been encapsulated into polymeric nanoparticles (TransIT-X2, Myrus) following the manufacturer's instructions. Importantly, this approach has been shown to enable intracellular delivery of nucleic acids in primary human innate immune cells, overcoming a critical barrier for nucleic acid delivery and sensing[50].

In our human TLR stimulation experiments **(Figure 5-7B)**, whole PBMCs from a healthy donor were incubated for 48 hours with or without 2 μg/ml of a TLR2 (PGNA), TLR3 (Poly I:C), TLR4 (LPS), or TLR8 (CL097) ligand (InvivoGen, San Diego, CA). Following incubation, mDCs were sorted (see **Flow Cytometric Analysis and Sorting**) into two replicate 5,000-10,000 cell populations and sequenced (see **Single-Cell and Population RNA Samples**).

Flow Cytometric Analysis and Sorting

PBMC were stained with LIVE/DEAD cell blue viability dye (Invitrogen, Carlsbad, CA) and monoclonal antibodies directed against CD11c (BioLegend, San Diego, CA), CD14 (BD Biosciences, San Jose, CA), HLA-DR, CD64, PD-L1, ICAM1, CD16, SLAMF8 (BioLegend) and subsequently analyzed on a Fortessa cytometer (BD Biosciences). Data were analyzed with FlowJo software (Tree Star, Ashland, OR). mDCs were identified from bulk PBMCs as a population of viable $CD14^-$ cells expressing high levels of CD11c and HLA-DR.

For the functional studies on mDC subsets, BVD-negative $CD14^-$ $CD11c^+$ $HLADR^+$ mDCs were sorted into two subpopulations expressing high and low levels of CD64 and PD-L-1 **(Figure 5-5C)**.

In the experiments evaluating polyfunctional $CD8^+$ T cell responses in EC, cultured cells (see **Activation of $CD8^+$ T cells from EC with autologous CD64,PD-L1 mDC**) were first labeled with LIVE/DEAD cell blue viability dye and anti-CD8 and CD-3 monoclonal Abs (BioLegend, San Diego). Subsequently, T cells were fixated and permeabilized and incubated with monoclonal antibodies against TNFα, IFNγ (BioLegend) and CD107a (BD Biosciences).

Mixed Leukocyte Reaction Assays

FACS purified viable $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ and $CD64^{Lo}$,$PD\text{-}L1^{Lo}$ mDC subpopulations, generated after 24h of infection with a VSV-G HIV-1 virus or 24h of incubation with TLR3 ligands (2μg/ml poly I:C and nanoparticle-loaded gag-dsDNA adjuvants), were mixed with allogeneic total peripheral blood T lymphocytes previously stained with 5 µM carboxyfluorescein succinimidyl ester (CFSE, Invitrogen) at a T:DC ratio of 4:1. As a control, T cells were also cultured in the presence of media alone or 2.5 µg/ml PHA

(Sigma) and 50 IU/ml IL-2 (NIH AIDS reagent program). After incubation for 6 days, cells were washed, stained with viability dye and anti-CD4 and anti-CD8 antibodies (BioLegend, San Diego, CA), and CFSE dilution on $CD4^+$ and $CD8^+$ T cell subpopulations was analyzed by flow cytometry using a Fortessa flow cytometer.

Autologous CD64, PD-L1 mDC Subsets

Total $CD8^+$ T cells were isolated by magnetic cell sorting (DynaBeads, Thermo Fisher) from unstimulated PBMC from ECs (n = 5) and cultured in the absence or the presence of autologous $CD64^{Hi}$,PD-$L1^{Hi}$ and $CD64^{Lo}$,PD-$L1^{Lo}$ mDC sorted from an alternative PMBC culture previously infected with a VSV-G pseudotyped HIV-1 virus for 24h, as previously described (see ***In Vitro* Infection with HIV-1 Virus**) at a ratio (T:DC = 4:1). After 2h of incubation, cultures were supplemented with Brefeldin A (BioLegend) and Monensin (BD-Biosciences) and left in culture for 16 additional hours. Phenotypic analysis of Polyfunctional $CD8^+$ T cell responses was determined by flow cytometry of intracellular expression of IFNγ, TNFα and CD107a **(see Flow Cytometric Analysis and Sorting; Figure 5-5H-J)**.

Quantification of HIV-1 by qPCR

HIV-1 reverse transcripts present in sorted mDC populations were amplified from cell lysates at 24h post-infection as previously described [51]. Copy numbers of reverse transcripts were obtained after extrapolation to specific standard curves generated from HIV-1-infected 293T cells [25]. qPCR data were normalized to relative *CCR5* gene copy number.

Statistics of *In Vitro* Functional Assays

The significance of differences in the fractional abundance of sorted mDC subsets across different cohorts and in our functional assays – including the mixed leukocyte reactions, culture of $CD8^+$ T cells from EC with autologous mDC and the TLR ligand and DNA stimulation assays – were determined using two-tailed Wilcoxon matched-pairs signed-rank test. In some experiments, we applied a Kruskal-Wallis test with post-hoc Dunn's test – adjusting for test multiplicity - using GraphPad Prism 6 software. The specific test used for each comparison is noted in the text.

Single-Cell and Population RNA Samples

Following sorting, whole transcriptome amplification (WTA) was performed on 96-well plates of single cells as described previously [49]. Briefly, individual lysed cells were cleaned with 2.2x volume AMPure XP SPRI beads (Beckman Coulter, Danvers, MA), and isolated cellular mRNAs were reverse transcribed and amplified.

For the population samples, total RNA was isolated using a RNeasy plus Micro RNA kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. 2 µL of this isolated RNA was then added to 8 µL of water and cleaned with 2.2x volume beads. Finally, 1 µL of this cleaned RNA was used in a WTA reaction [49].

Following WTA, PCR products were cleaned with 0.9x volume SPRI beads and eluted in water. The concentration of cDNA in the resulting solution was determined using a Qubit

3.0 Fluorimeter (ThermoFischer, Waltham, MA) and analyzed using a high sensitivity DNA chip for BioAnalyzer (Agilent, Santa Clara, CA).

cDNA Library Preparation for RNA-Seq

WTA products were diluted to a concentration of 0.1 to 0.4 ng/μL, tagmented and amplified using Nextera XT DNA Sample preparation reagents (Illumina, San Diego, CA). Tagmentation was performed according to manufacturer's instructions, modified to use ¼ the recommended volume of reagents, extending tagmentation time to 10 minutes and extending PCR time to 60s. PCR primers were ordered from Integrated DNA Technologies (Coralville, IA). Nextera products were then cleaned twice using 0.9x SPRIs and eluted in water. The final library was quantified using Qubit and analyzed using a high sensitivity DNA chip. It was then diluted to 2.2 pM and sequenced on a NextSeq 500 (Illumina).

Single-Cell Expression Quantification

RNA-Seq reads were aligned to the RefSeq hg38 transcriptome (GRCh38.2) using Bowtie2 [52]. The resulting transcriptomic alignments were processed by RSEM to estimate the abundance (expected counts and transcripts per million (TPM)) of RefSeq transcripts [53].

Several genes were quantified multiple times due to alternative isoforms unrelated by RefSeq annotation. Prior to expression data normalization, these TPM estimates were summed to produce a single TPM estimate per RefSeq gene symbol.

Single-Cell Filtering and Gene Filtering

For each single-cell library, we computed transcriptome alignment and quality metrics using FastQC (Babraham Bioinformatics), Picard tools (Broad Institute), and custom scripts. Computed metrics included: 1) number of reads, 2) number of aligned reads, 3) percentage of aligned reads, 4) number of duplicate reads, 5) primer sequence contamination, 6) average insert size 7) variance of insert size, 8) sequence complexity, 9) percentage of unique reads 10) ribosomal read fraction, 11) coding read fraction, 12) UTR read fraction, 13) intronic read fraction, 14) intergenic read fraction, 15) mRNA read fraction, 16) median coefficient of variation of coverage, 17) mean 5' coverage bias, 18) mean 3' coverage bias, and 19) mean 5' to 3' coverage bias.

We used the *metric_sample_filter* function from the SCONE package [27] to flag libraries with low numbers of aligned reads (< 28,840), low percentages of aligned reads (< 15%), and low percentages of detected transcripts (< 33.4% of Ensembl GRCh38.80 protein-coding genes expressed at > 100 TPM in at least 10% of samples – or "common genes") **(Figure 5-3A-C)**. We identified 99 genes of candidate constitutive expression by fitting a population-wide Fano factor as a linear function of mean TPM, selecting the 99 common genes with minimal fit residual. These genes covered a range of 50.0 – 35,000 TPM. For each sample, the relationship between mean detected TPM and detection rate (or "false negative characteristic") was modeled as a logistic function; the area under this fitted curve was utilized to distinguish samples with poor detection properties **(Figure 5-3D,E)**. Out of 2,489 initial samples, only 393 (318 at 48H and 75 at 24H) samples passed this primary filter. We note that some of this loss is due to our decision to exclude viability stain for some single-cell RNA-Seq sorts. Importantly, this viability selection did not appear to skew the sub-composition of cell-states passing our sample filtering criteria (see **Clustering Analysis and Visualization**).

Following cell filtering, genes were retained for downstream analysis if they were annotated as protein-coding, and expressed at levels greater than 100 TPM in at least 5 high-quality cells.

Single-Cell Data Normalization

In order to normalize TPM data, we applied full-quantile normalization method, restoring original zero values to zero following normalization. This restoration step was necessary due to widespread zero-ties. We used normalization metrics of the SCONE [27] package to assess performance of this strategy.

The first three scores measure the maximum absolute correlation between the first 3 PCs of the TPM matrix and the first 3 PCs of: i) the matrix of library-level qc metrics, ii) the un-normalized matrix of TPM estimates for "negative control" genes from the MSigDB[9] "HSIAO_HOUSEKEEPING_GENES" gene set, and iii) the un-normalized matrix of TPM estimates for "positive control" genes from the MSigDB "REACTOME_INNATE_IMMUNE_SYSTEM" gene set. Following normalization, the first two scores decreased while the third increased slightly, suggesting that technical structure has been removed from the data while retaining structure associated with the biological processes at hand **(Figure 5-3F)**.

The next three scores measure the average silhouette width for various classifications across a Euclidean metric defined on the first 3 PCs of the TPM matrix: i) biological class = patient ID x exposure x time point x viability, ii) batch class, and iii) average silhouette width where each stratification of batch and biology has been separately clustered using the Partitioning Around Medoids (PAM) clustering algorithm. Following normalization, the first two scores decrease, suggesting that confounding by biological and batch factors

could not be addressed by this normalization. However, the rise of the third score suggests greater intra-stratum clustering following normalization **(Figure 5-3G)**.

The last two scores: i) the median absolute relative log-expression (RLE), and ii) the variance of the RLE inter-quartile range both decreased, implying reduced global differential expression following normalization **(Figure 5-3H)**.

Clustering Analysis and Visualization

Principal Component Analysis (PCA) was applied to all filtered and normalized single-cell log-TPM data collected at the 48-hour time point; consequent analysis was limited to the first 50 PC values (defined per cell) explaining 32% of expression variance. For each choice of dimension d ranging from 2-50, a Euclidean cell-distance matrix was computed over the first d PCs. The PAM clustering algorithm was used to cluster cells over a range of k = 2 to 10 clusters. Let S(k,d) represent the average silhouette width of a PAM k-clustering on d dimensions. We define k(d) as the unique choice of k that maximizes S(k,d) for any choice of d. We selected d so as to maximize cluster number and tightness:

$$k(d) \geq k(d') \; \forall \; d' \neq d$$

$$S(k(d), d) \geq S(k(d'), d') \; \forall \; \{d' | k(d') = k(d)\}$$

d = 7, and k = 5 were the selected clustering parameters. This method is implemented in the *pamkd* function in the *scRAD* package.

Due to the high-dimension of the underlying expression space, clustering was visualized using a 2D tSNE projection applied to the d = 7 distance metric (5,000 iterations). The 5 clusters were annotated in clockwise order.

After clustering we applied ordinary least-squared regression to model each gene i's expression in cell j as a function of patient, exposure, and cell type:

$$g_{ij} \sim \alpha_i + \beta_i^p * Patient_j + \beta_i^e * Exposure_j + \beta_i^c * Cluster_j$$

Patient contrasts were coded p1_vs_p3 and p1_vs_p2, exposure contrasts coded hiv_vs_media, and cluster contrasts codes c2/3/4/5_vs_c1. Two-sided t-tests identified 131 and 14 genes that were significantly associated with patient and exposure respectively (Bonferroni-adjusted p-value < 0.01), while 1,170 genes were significantly associated with cluster contrasts. These numbers suggest that cluster identity is far more determinant of global gene expression than patient or exposure. Cluster proportions are themselves associated with patient and exposure condition: for c1/2/3/4 we modeled the relative abundance of cluster k as a logistic model of Patient and Exposure:

$$P(c_k) \sim \alpha_k + \beta^p * Patient + \beta^e * Exposure$$

While all 4 clusters exhibited significant association by patient ($p < 0.05$), all but c2 showed significant evidence ($p < 0.05$) of exposure dependence: the c1 proportion was enriched by HIV exposure, while both c3 and c4 were depleted by the exposure.

In patients p2 and p3, for which viability sorts were applied to some batches, we observed similar cluster compositions across both exposure conditions at 48 hours (**Figure 5-3I**), suggesting that pre-selection of viable cells does not affect the distribution of the clusters identified and analyzed in this study. Instead, the effect of viability sorting appears to be the depletion of a large, low-quality cluster exhibiting low B2M expression uncharacteristic of mDCs (**Figure 5-3J-K).**

Reproducible Module Gene Analysis

Our clustering analysis captured the full distribution of cell states seen across the three ECs, but we also attempted to identify clusters of genes – gene modules – that were consistently co-regulated across patients at 48H. Unlike differential expression analysis, this unsupervised approach aims to identify transcripts serving as reliable proxies of reproducible gene expression patterns.

We first pooled the normalized log-TPM data for each patient and separately computed the gene-gene Pearson correlation matrix. Each correlation value was Fisher-transformed and scaled to a z-score with 0 median and a Median Absolute Deviation (MAD) equal to 0.67 (computed over the upper-triangle). Only gene pairs with abs(z) > 2.4 in all three patient matrices were considered "reproducible gene pairs." This step is implemented in the *scRAD::get.repro.thresh.adjacency* function in R.

For each gene, we tallied the number of reproducible gene pairs to which it belongs. We considered whether we could find genes with significantly more pairs than would be expected by chance; these genes could serve as reliable proxies of reproducible correlations. The distribution of pair counts was modeled as a zero-inflated Poisson process, including a randomly-connected Poisson component and an unconnected zero-component. Under this null model, we computed upper-tail p-values using the *scRAD::pzipdegree* function, identifying 263 genes with p-values below 0.01 after Bonferroni adjustment. As these genes are connected to a large number of reproducible gene pairs, we called these proxy genes "reproducible module genes."

Complete clustering of the median gene-gene correlation across the 3 patients (using correlation distance) demonstrates how these genes cluster into 3 specific modules.

Single-Cell Signature Analysis

We searched Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) for all study entries matching the query: "(("homo sapiens"[Organism] NOT "mus musculus"[Organism]) AND ("expression profiling by array"[DataSet Type] OR "expression profiling by high throughput sequencing"[DataSet Type])) AND ("dendritic cell"[Sample Source] or "dendritic cells"[Sample Source])", utilizing the results to identify relevant expression signatures from the MSigDB C7 collection. We then applied FastProject[10] to identify representative expression signatures in our normalized TPM data. Signature inputs include the selected MSigDB signatures, a curated signature of 28 IFN-response genes[25,28], 3 unsigned signatures of our reproducible modules, and a pre-computed cluster signature. Results show that PCA components 1 and 3 represent both biological signatures and reproducible module signatures more faithfully than alternative 2D projections.

We selected a few of the top signatures from our FastProject analysis, considering the cumulative distribution of signatures across each of the 5 clusters **(Figure 5-4A,B)**. Two-sided Kolmogorov–Smirnov (KS) tests were performed between the signature distributions of clusters c1 (n = 20), c2 (n = 26), and c3 (n = 35) in order to monitor the extent to which these signatures demonstrated differential signatures.

Differential Expression Analysis

Based on our signature analysis above, we considered 3 differential expression comparisons: i) c1 vs c3, c4, and c5 (or "c1 vs c3-5"), ii) c2 vs c3,c4, and c5 (or "c2 vs c3-5"), and iii) c1 vs c2. Any differential expression (DE) analysis downstream of *de novo* clustering analysis demands careful consideration. Traditional DE analysis aims at identifying transcripts that vary markedly by sample class; a common goal is to rank the

relative importance of transcripts in characterizing underlying expression states. Within the single-cell context, cell class is frequently defined based on low-dimensional representations of expression data. Therefore, the assumption that most genes are not differentially expressed between classes may not hold. Null models based on this assumption are ill-suited to the data, and will naturally yield uncalibrated probabilistic-based scores, e.g. deflated p-value distributions.

A natural way to calibrate DE scores and monitor batch-specific effects is to consider measures of reproducibility over stratified, replicate experiments: in our case, over multiple patients. We can map clusters from replicate experiments so that cluster contrasts are made comparable. For example, in our analysis we clustered cells from all patients simultaneously - offering a natural mapping between clusters called in the three patients: e.g. c1 cells in patient 1 belongs to the same biological "pseudo-replicate" as c1 cells in patient 2.

Our meta-analytical DE approach, implemented using the *scRAD::kruskalIDRm* tool, relies on the reproducibility metric known as Irreproducible Discovery Rate (IDR)[29]. This metric evaluates a matched set of "signals" measured in two or more replicate experiments. In this analysis, we pooled cells from patients 1 and 2 to define the first study stratum, and considered all cells from patient 3 to be the second replicate stratum. We pooled cells from patients 1 and 2 together because the fewest high-quality cells were sequenced in these patients; pooling them together increased average stratum power. Though we performed a 2-replicate analysis, our scRAD package modifies the Expectation-Maximization algorithm from the *idr* CRAN package to handle 3 or more replicates (see ***scRAD* vignette** on GitHub).

We performed simple DE analysis in each replicate study using Kruskal-Wallis tests; for each comparison, this yielded two lists of log-fold-changes and two lists of p-values. The two-component IDR mixture model was used to fit the joint distribution of p-values obtained from these tests. For each gene, we can estimate a probability that the gene is a member of an "irreproducible component" for which p-values are high and uncorrelated versus a "reproducible component" for which p-values are low and correlated. Sorting genes by increasing probability of irreproducibility, one can compute the cumulative probability of membership for all genes of same or lower rank, defining an IDR. Genes called with an IDR < 0.01 were reported as "differentially expressed."

We compared this approach to DE effects estimated according to a more traditional model of log-expression in gene i in cell j:

$$g_{ij} \sim \alpha_i + \beta_i^p * Patient_j + \beta_i^c * Cluster\_contrast_j$$

The genes that meet the significance criterion but not the reproducibility one may be good candidates for patient-specific differential expression. Patient clustering was tighter for high IDR genes, while cluster contrasts were tighter for lower IDR genes. These results exemplify how our meta-analysis approach targets covariance structures shared across patients.

If we assume the difference between c1 and c2 is small compared to their common differences with clusters c3, c4, and c5 (c3-5), we may claim that the more a gene is reproducibly differentially expressed for one comparison, the more likely that gene should be reproducibly differentially expressed in the other. By this assumption, IDR analysis can be applied to two lists of idr values from separate experiments in order to identify genes for which idrs obtained from both comparisons are themselves correlated vs uncorrelated. Genes passing this threshold and showing common sign of differential expression were

called "Shared" genes in **Figure 5-4C**. Some of the remaining differentially expressed genes from these two comparisons were partitioned into three additional groups: i) "c1-specific" for which a gene is called differentially expressed in both c1 vs c3-5 AND c1 vs c2 comparisons, but not c2 vs c3-5. ii) "c2-specific" which is analogously defined, and iii) "discordant" for which genes are called differentially expressed in all 3 comparisons.

Candidate surface markers for c1 were identified using the *scRAD::getMarkers* tool. This tool reports the intersection of three gene sets: i) genes differentially expressed between c1 and c3-5, ii) Reproducible module genes, and iii) Predicted membrane molecules from the Human Protein Atlas (http://www.proteinatlas.org) **(Figure 5-5A)**.

Ingenuity Pathway Analysis

For each of the main 3 differential expression comparisons, we applied Ingenuity Pathway Analysis[15] (https://www.ingenuity.com) to the list of log-fold-change (mean of log-fold-changes from 2 replicate tests) and IDR, setting a cutoff of IDR < 0.05. The user data set was used as the reference set for p-value calculation, and all experimentally verified mammalian associations were included in the analysis. IPA reported Benjamini-Hochberg q-values for Canonical Pathways enrichments, and we performed our own Bonferroni p-value adjustment for all reported Upstream Analysis p-values.

Validation of c1-enriched $CD64^{Hi}$,$PD\text{-}L1^{Hi}$ Population

As with the scRNA-Seq data, we applied RSEM alignment and sample-filtering procedures to population RNA-Seq samples sorted by c1 candidate markers, leaving 13 samples covering 8 possible conditions (24h/48h, HIV/media, Hi/Lo). Expression values for 6,557

genes were normalized using traditional DESeq scaling normalization [54], followed by gene-level regression on the first PC of QC metrics, retaining the residual for downstream analysis. Duplicate gene symbols were averaged as above. 576 of the differentially expressed gene symbols from the c1 vs c3-5 comparison, passing TPM gene filter, were detected in population experiments. A weighted mean was computed for each of these shared genes, for each single-cell sub-population c1-5, and Pearson correlations were computed between sorted populations and population means after log1p-transforming both data sets. Radar plot cycles representing these correlations are presented on a min-max scale per bulk condition (min: 0.32-0.54, max: 0.81-0.89). Correlation values for replicate sample conditions (n = 2) were averaged prior to plotting.

Prediction of Upstream Regulators of c1

In order to generate hypotheses related to the sensing mechanisms behind c1 response, we performed IPA (as described above) and identified several innate immune pathways – included specific TLR signaling pathways – selectively induced in this population. Due to the limited availability of genome-wide human stimulation data, we opted to compare our single-cell expression profiles to publically available expression profiles of mouse bone-marrow-derived dendritic cell (BMDC) populations exposed to five TLR agonists (lipopolysaccharide (LPS), Pam3CSK4 (PAM), Polyinosinic:polycytidylic acid (Poly I:C), gardiquimod (Gard), and CpG DNA (CpG)) and one control (unstimulated) condition[36]. Replicate microarray samples from each condition were averaged, followed by averaging over probes of a gene symbol. Homologs were mapped using the biomaRt Bioconductor package, and only uniquely mapping genes were considered for further analysis. Normalized single-cell TPM was log1p-transformed, and gene abundances were centered

by weighted means (using the W false-negative weight matrix defined above). Mouse data was log-transformed, and genes were centered by their mean value. We computed a weighted correlation estimate (using the W matrix) for each pair of single-cell and mouse population taken at the 24H time-point of the mouse study. For each bulk sample, we applied two-tailed Wilcoxon rank-sum tests to examine differences in correlation between cells from c1 and c3-5. The correlations were referred to as the "TLR stimulation score," as they measure the extent to which the sub-population-specific response is correlated with the TLR-stimulated profile. Using Stouffer's z-method, we combined p-values collected from the two donor pools used in differential expression (all implemented in *scRAD::kruskalMeta*) reporting a meta-analysis FDR < 0.01 **(Figure 5-7B)**[36]. Weighted correlation of samples i and i' is defined by the equations:

$$\mathrm{wMean}_i\left(X,W\right)=\frac{\sum_j X_{ij}W_{ij}}{\sum_{j'} W_{ij'}}$$

$$\mathrm{wCov}_{ii'}\left(X,W\right)=\frac{\sum_j W_{ij}W_{i'j}\left(X_{ij}-\mathrm{wMean}_i\left(X,W\right)\right)\left(X_{i'j}-\mathrm{wMean}_{i'}\left(X,W\right)\right)}{\sum_{j'} W_{ij'}W_{i'j'}}$$

$$\mathrm{wCor}_{ii'}\left(X,W\right)=\frac{\mathrm{wCov}_{ii'}\left(X,W\right)}{\sqrt{\mathrm{wCov}_{ii}\left(X,W\right)\mathrm{wCov}_{i'i'}\left(X,W\right)}}$$

Where weights of population data are set to unity.

We next sought to generate analogous results using RNA-Seq data collected from human mDCs rather than distant mouse BMDCs. We applied RSEM alignment and sample-filtering procedures to population RNA-Seq data collected from DCs incubated for 48 hours with or without various TLR ligands (see above), leaving 8 samples covering 5 possible conditions (no TLR, TLR2/3/4/8). Expression values for 18,482 genes were normalized using traditional DESeq scaling normalization [54]. Duplicate gene symbols were

averaged as above. We applying the same meta-analysis pipeline as for the mouse array data, ranking various inductions in their relative similarity to c1.

Drawing again on available characterizations of the mouse BMDC system, we chose to correlate single-cell gene expression profiles with shRNA knockdowns of TLR signaling network components)[1] to highlight potential upstream regulators mediating c1 response. Publically available – and normalized – nCounter population data were mapped to unique human homologs, log-scaled and gene-centered as above. Normalized single-cell TPM estimates were similarly log1p-transformed and centered by weighted mean. Weighted correlation estimates were computed as in the TLR analysis above, and for each shRNA experiment, we applied two-tailed Wilcoxon rank-sum tests to examine differences in correlation between c1 and c3-5. The opposite of the correlation was referred to as the upstream regulatory score, as it measures the extent to which the sub- specific response is anti-correlated with the shRNA-knockdown profile. Instead of simple meta-analysis on the donor pools used for the TLR stimulation scores, we applied the *scRAD::kruskalIDRm* analysis as in the differential expression analysis, defining IDR < 0.05 as our threshold for calling differential signatures **(Figure 5-7F)**.

AVAILABILITY OF DATA AND MATERIALS

Single-cell and bulk RNA-seq data is available through the Gene Expression Omnibus (GEO accession GSE80212).

This study utilized two publicly available expression data sets i) Amit et al. 2009 [33], accessible via GEO accession GSE1772 and ii) Chevrier et al. 2011, accessible via

"Supplemental Information" S2 and S7 provided in [32]. Signature analyses relied on expression signatures defined in MSigDB (http://software.broadinstitute.org/gsea/msigdb).

The ***scRAD*** package is available on GitHub (https://github.com/YosefLab/scRAD) under Artistic License 2.0. Normalized scRNA-Seq expression data, meta data, and average bulk expression profiles from the TLR induction study are available as data objects in the package. See the *v0-genome-bio* tag for code at time of manuscript submission.

ETHICS APPROVAL

Samples from patients were obtained after all subjects gave written informed consent; the Institutional Review Board of Massachusetts General Hospital/Partners Healthcare approved the study protocol (IRB approval number 2012P001628). Samples were collected according to the Helsinki Declaration.

5.10 References

1 Chevrier, N. *et al.* Systematic Discovery of TLR Signaling Components Delineates Viral-Sensing Circuits. *Cell* **147**, 853-867, doi:papers3://publication/doi/10.1016/j.cell.2011.10.022 (2011).

2 Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science* **332**, 677-678, doi:10.1126/science.1206351 (2011).

3 Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, 1-16, doi:10.1016/j.cell.2015.11.013 (2015).

4 Allman, D. & Pillai, S. Peripheral B cell subsets. *Curr Opin Immunol* **20**, 149-157, doi:10.1016/j.coi.2008.03.014 (2008).

5 Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nat Immunol* **16**, 343-353, doi:10.1038/ni.3123 (2015).

6 Kanno, Y., Vahedi, G., Hirahara, K., Singleton, K. & O'Shea, J. J. Transcriptional and Epigenetic Control of T Helper Cell Specification: Molecular Mechanisms Underlying Commitment and Plasticity *. *Annu. Rev. Immunol.* **30**, 707-731, doi:10.1146/annurev-immunol-020711-075058 (2012).

7 Merad, M., Sathe, P., Helft, J., Miller, J. & Mortha, A. The Dendritic Cell Lineage: Ontogeny and Function of Dendritic Cells and Their Subsets in the Steady State and the Inflamed Setting. *Annu. Rev. Immunol.* **31**, 563-604, doi:10.1146/annurev-immunol-020711-074950 (2013).

8 Villani, A. C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, doi:10.1126/science.aah4573 (2017).

9 Ceppi, M. *et al.* Ribosomal protein mRNAs are translationally-regulated during human dendritic cells activation by LPS. *Immunome research* **5**, 5, doi:10.1186/1745-7580-5-5 (2009).

10 Manel, N. *et al.* A cryptic sensor for HIV-1 activates antiviral innate immunity in dendritic cells. *Nature* **467**, 214-217, doi:10.1038/nature09337 (2010).

11 Zaslavsky, E. *et al.* Antiviral response dictated by choreographed cascade of transcription factors. *Journal of immunology* **184**, 2908-2917, doi:10.4049/jimmunol.0903453 (2010).

12 Cohen, A. A. *et al.* Dynamic Proteomics of Individual Cancer Cells in Response to a Drug. *Science* **322**, 1511-1516, doi:10.1126/science.1160165 (2008).

13 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

14 Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369, doi:10.1038/nature13437 (2014).

15 Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496**, 461-468, doi:10.1038/nature11981 (2013).

16 Feinerman, O. *et al.* Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response. *Molecular Systems Biology* **6**, 1-16, doi:10.1038/msb.2010.90 (2010).

17 Napolitani, G., Rinaldi, A., Bertoni, F., Sallusto, F. & Lanzavecchia, A. Selected Toll-like receptor agonist combinations synergistically trigger a T helper type 1-polarizing program in dendritic cells. *Nat Immunol* **6**, 769-776, doi:10.1038/ni1223 (2005).

18 Casanova, J. L. Human genetic basis of interindividual variability in the course of infection. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E7118-7127, doi:10.1073/pnas.1521644112 (2015).
19 Liu, R. *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86**, 367-377 (1996).
20 Blankson, J. N. Effector mechanisms in HIV-1 infected elite controllers: highly active immune responses? *Antiviral research* **85**, 295-302, doi:10.1016/j.antiviral.2009.08.007 (2010).
21 Saez-Cirion, A. *et al.* HIV controllers exhibit potent CD8 T cell capacity to suppress HIV infection ex vivo and peculiar cytotoxic T lymphocyte activation phenotype. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 6776-6781, doi:10.1073/pnas.0611244104 (2007).
22 Gao, X. *et al.* AIDS restriction HLA allotypes target distinct intervals of HIV-1 pathogenesis. *Nature medicine* **11**, 1290-1292, doi:10.1038/nm1333 (2005).
23 Kiepiela, P. *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432**, 769-775, doi:10.1038/nature03113 (2004).
24 Huang, J. *et al.* Leukocyte immunoglobulin-like receptors maintain unique antigen-presenting properties of circulating myeloid dendritic cells in HIV-1-infected elite controllers. *Journal of virology* **84**, 9463-9471, doi:10.1128/JVI.01009-10 (2010).
25 Martin-Gayo, E. *et al.* Potent Cell-Intrinsic Immune Responses in Dendritic Cells Facilitate HIV-1-Specific T Cell Immunity in HIV-1 Elite Controllers. *PLoS pathogens* **11**, e1004930, doi:10.1371/journal.ppat.1004930 (2015).
26 Alter, G. *et al.* HIV-1 adaptation to NK-cell-mediated immune pressure. *Nature* **476**, 96-100, doi:10.1038/nature10237 (2011).
27 Cole, M. *et al.* SCONE: Single Cell Overview of Normalized Expression data. R package version 1.0.0. (2016).
28 DeTomaso, D. & Yosef, N. FastProject: A Tool for Low-Dimensional Analysis of Single-Cell RNA-Seq Data. *BMC Bioinformatics*, 1-12, doi:10.1186/s12859-016-1176-5 (2016).
29 Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752-1779, doi:10.1214/11-AOAS466 (2011).
30 Peris-Pertusa, A. *et al.* Evolution of the functional profile of HIV-specific CD8+ T cells in patients with different progression of HIV infection over 4 years. *Journal of acquired immune deficiency syndromes (1999)* **55**, 29-38, doi:10.1097/QAI.0b013e3181e69609 (2010).
31 Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 1-23, doi:papers3://publication/doi/10.1038/nature14966 (2015).
32 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).
33 Chen, H. *et al.* CD4+ T cells from elite controllers resist HIV-1 infection by selective upregulation of p21. *The Journal of clinical investigation* **121**, 1549-1560, doi:10.1172/JCI44539 (2011).
34 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 1-5, doi:10.1038/nmeth.2639 (2013).
35 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*, 1-13, doi:10.1186/s13059-015-0844-5 (2015).

36 Amit, I. *et al.* Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science (New York, NY)* **326**, 257-263, doi:10.1126/science.1179050 (2009).
37 de Saint-Vis, B. *et al.* A novel lysosome-associated membrane glycoprotein, DC-LAMP, induced upon DC maturation, is transiently expressed in MHC class II compartment. *Immunity* **9**, 325-336 (1998).
38 Harman, A. N. *et al.* HIV infection of dendritic cells subverts the IFN induction pathway via IRF-1 and inhibits type 1 IFN production. *Blood* **118**, 298-308, doi:10.1182/blood-2010-07-297721 (2011).
39 van der Poel, C. E., Spaapen, R. M., van de Winkel, J. G. & Leusen, J. H. Functional characteristics of the high affinity IgG receptor, FcgammaRI. *Journal of immunology* **186**, 2699-2704, doi:10.4049/jimmunol.1003526 (2011).
40 Odorizzi, P. M., Pauken, K. E., Paley, M. A., Sharpe, A. & Wherry, E. J. Genetic absence of PD-1 promotes accumulation of terminally differentiated exhausted CD8+ T cells. *The Journal of experimental medicine* **212**, 1125-1137, doi:10.1084/jem.20142237 (2015).
41 Cunningham, C. R. *et al.* Type I and Type II Interferon Coordinately Regulate Suppressive Dendritic Cell Fate and Function during Viral Persistence. *PLoS pathogens* **12**, e1005356-1005326, doi:10.1371/journal.ppat.1005356 (2016).
42 Carranza, P., Del Rio Estrada, P. M., Diaz Rivera, D., Ablanedo-Terrazas, Y. & Reyes-Teran, G. Lymph nodes from HIV-infected individuals harbor mature dendritic cells and increased numbers of PD-L1+ conventional dendritic cells. *Hum Immunol* **77**, 584-593, doi:10.1016/j.humimm.2016.05.019 (2016).
43 Gao, D. *et al.* Cyclic GMP-AMP synthase is an innate immune sensor of HIV and other retroviruses. *Science* **341**, 903-906, doi:10.1126/science.1240933 (2013).
44 Yan, N., Regalado-Magdos, A. D., Stiggelbout, B., Lee-Kirsch, M. A. & Lieberman, J. The cytosolic exonuclease TREX1 inhibits the innate immune response to human immunodeficiency virus type 1. *Nat Immunol* **11**, 1005-1013, doi:10.1038/ni.1941 (2010).
45 Habjan, M. & Pichlmair, A. Cytoplasmic sensing of viral nucleic acids. *Current opinion in virology* **11**, 31-37, doi:10.1016/j.coviro.2015.01.012 (2015).
46 Loo, Y. M. & Gale, M., Jr. Immune signaling by RIG-I-like receptors. *Immunity* **34**, 680-692, doi:10.1016/j.immuni.2011.05.003 (2011).
47 Ma, Z. & Damania, B. The cGAS-STING Defense Pathway and Its Counteraction by Viruses. *Cell host & microbe* **19**, 150-158, doi:10.1016/j.chom.2016.01.010 (2016).
48 Chen, M., Huang, L., Shabier, Z. & Wang, J. Regulation of the lifespan in dendritic cell subsets. *Mol Immunol* **44**, 2558-2565, doi:10.1016/j.molimm.2006.12.020 (2007).
49 Trombetta, J. J. *et al.* Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr Protoc Mol Biol* **107**, 4 22 21-17, doi:10.1002/0471142727.mb0422s107 (2014).
50 Jones, A., Kainz, D., Khan, F., Lee, C. & Carrithers, M. D. Human macrophage SCN5A activates an innate immune signaling pathway for antiviral host defense. *J Biol Chem* **289**, 35326-35340, doi:10.1074/jbc.M114.611962 (2014).
51 Buzon, M. J. *et al.* HIV-1 persistence in CD4+ T cells with stem cell-like properties. *Nature medicine* **20**, 139-142, doi:10.1038/nm.3445 (2014).
52 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

53 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).
54 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

Chapter 6: Conclusions

In this final chapter, we explore remaining questions and possible extensions of this work. While this work has contributed to our understanding of how cells respond to perturbations at varied relevant biological scales, many questions and fascinating avenues of exploration remain. The acute liver injury and high fat diet work presented Chapters 3 and 4 was conducted in exclusively male mice, leaving room for future work to build complementary datasets in females, which are known to be less susceptible to the damaging effects of both acetaminophen-induced acute liver injury and obesity-related liver cancer, to further characterize and mechanistically define these sex-differences. In both males and females, we could build time course trajectories to better understand the long term effects of the high fat diet chronic injury model. Additionally, combining a mouse high fat diet trajectory dataset with human clinical data would serve to help us determine the relevance of the mouse findings to contextualize a single time point human sample, and may point toward biomarkers of disease or therapeutic targets. Computational approaches developed in Chapter 5 can identify reproducible responses across phenotypically similar human donors in future clinical datasets. Many of the approaches and insights presented in the preceding chapters will aid the success of future work.

6.1 Contributions of this work

Here, we have applied scRNA-Seq approaches to understand cellular responses to perturbations at the relevant biological scale, ranging from within a single organ, to across an organ system to between genetically unique individual human donors with a shared phenotype of interest. In Chapter 3, we identify and describe a compensatory phase of liver response to injury, in which the surviving hepatocytes upregulate their expression of critical liver function genes to take over for the work previously accomplished by the hepatocytes lost to injury. Following the functional compensation phase, the liver initiates proliferation, leading to cellular recovery and a return to the pre-injured state. Our single cell data allow us to identify rare cycling hepatocytes and characterize them in greater detail than possible before. Our data suggest that the cycling hepatocytes do not upregulate compensatory liver function to the same degree as non-cycling hepatocytes, at least not for all genes. We also show that macrophage derived Wnts support functional compensation in hepatocytes.

In Chapter 4, we extend our approach from focusing on an acute injury targeted to a single organ, to exploring chronic damage from a long-term high fat diet across multiple gastrointestinal and immune compartments which are affected by this perturbation. In this pilot study, we scRNA-sequenced samples across multiple gastrointestinal and reference immune compartments in mice after six months on a high fat diet (HFD) or control diet (CD) to explore cellular changes and molecular drivers which contribute to obesity-linked inflammation and cancer in the liver and gut. We observe possible immune shifts in both the liver and gut, identify PPAR activation primarily in the HFD proximal regions and especially in the HFD proximal enterocytes, and distinct differences in biology between high fat and control hepatocytes. We nominate pathways possibly contributing to HFD-induced changes in the liver and identify a small group of HFD hepatocytes with an elevated stemness signature which may indicate precancerous changes.

Finally, we characterize shared features across multiple unique human donors with a common phenotype. We explore dendritic cell responses to HIV-1 virus across multiple unique human donors and identify reproducible behaviors which may contribute to the elite controller (EC) phenotype. We develop broadly applicable analytical approaches to identify reproducible responses across donors and to nominate candidate targets for rationally modulating the system. A highly functional subgroup of dendritic cells which is better able to prime T cells for proliferation emerges from the data in all sequenced ECs. We then modulate the response in healthy DCs to induce more of the highly functional DCs observed in ECs by activating TLR3, which was predicted from the sequencing data analysis. Our methods for identifying reproducible responses across donors and nominating targets for functional validation are broadly applicable, and could be of use to future studies.

In the following sections, we suggest extensions of the work presented here to new projects which will explore remaining questions.

6.2 Sex-differences

Sex-differences in liver disease frequencies are discussed in detail in Appendix A. Briefly, even after controlling for behavioral factors, biologically-based differences in susceptibility to disease between males and females remain. The liver in particular is known to be a highly sexually dimorphic organ, with rates for many types of liver disease varying substantially between the sexes[1].

Female mice are known to be much less susceptible to liver damage from acetaminophen (APAP) overdose than males[2,3]. Appendix A describes our work to date comparing male (data from Chapter 3) to female (data acquisition in progress) mice at multiple time points following APAP

exposure. We do observe clear sex-based differences, including less damage to female livers. Additional work to improve female data quality and fully build the female dataset is still required.

Females are less susceptible not only to APAP-induced damage but also have lower rates of hepatocellular carcinoma (HCC), an obesity-linked cancer observed in our HFD mice[1,4]. Similar to what we've begun to do with comparing male and female responses to APAP, we could build a complementary female HFD and CD dataset to compare to the male data presented in Chapter 4. In doing so, we may uncover new pathways and molecular regulators that contribute to female protection.

6.3 High Fat Diet Time Course

All high fat diet data presented in Chapter 4 originates from mice on HFD or CD for six months, but the biological changes in response to HFD occur over many months, progressing to more severe disease, including cancer, by 9-14 months on the diet[5]. Future work may build a time course, profiling mice after 3, 6, 9, and 12 months on the diet to explore changes in cell types and cell behaviors over time. Time course work would improve our understanding of how the cells change though the phases of metabolic syndrome-related disease and possibly pathways activated or deactivated driving development of inflammation and cancer in the liver or intestine. Pseudotime computational approaches could be applied to explore cellular trajectories as the cells respond to HFD. Many of the existing pseudotime packages were built based on differentiation data sets[6] and may require adaptations for applications to a cellular response dataset. Ideally, we will have spontaneous HCC tumors included in the late time points of the dataset. While HCC does spontaneously form in HFD mice in this cohort, additional carcinogen treatment may be desirable to increase the tumor frequency.

6.4 Combining human clinical data with mouse high fat diet data

We have begun to pilot experiments to extend our high fat diet work to clinical human samples: wedge biopsies from livers of patients undergoing weight loss surgery. Approximately 75% of obese patients have nonalcoholic fatty liver disease (NAFLD), which spans a continuum of liver disease including hepatic steatosis, nonalcoholic steatohepatitis (NASH), fibrosis and cirrhosis[7].

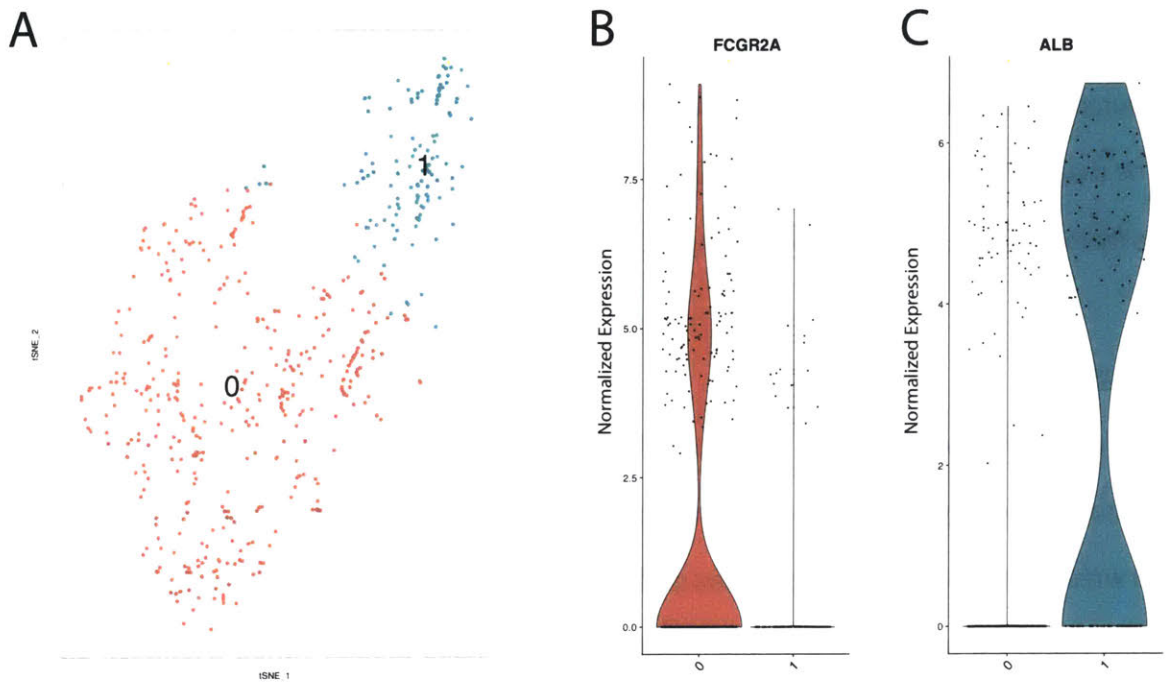

Figure 6-1 | Results from pilot weight loss surgery liver biopsy.
(A) tsne with SNN clustering identifies two clusters. **(B)** Violin plots depicting expression level of FCGR2A (a macrophage/kupffer-expressed gene) and ALB (a hepatocyte-expressed gene).

In comparing clinical human data to a trajectory dataset built in mice, we can confirm the relevance of the mouse data and better contextualize the human data along a larger continuum of progressive liver disease. Application of the computational framework presented in Chapter 5 to control donor-to-donor variability and identify reproducible modules will find phenotypically relevant shared responses across patients with similar stages of liver disease. Analysis of human data along the disease trajectory could be used to identify biomarkers which may be used better

categorize clinical patients according to their risk for development of more advanced problems. Analysis of the combined human and mouse data set may enable identification of activated or repressed pathways and nomination of targets for rational therapeutic modulation.

We have conducted a pilot experiment on a clinical weight loss surgery liver biopsy. Unfortunately, low alignment rate suggested poor sample quality and a need to further optimize our processing protocol for handling these fibrotic samples. Despite poor sample quality, we were able to identify some albumin-expressing cells (likely hepatocytes) and FCGR2A-expressing cells (likely immune kupffer or macrophages) (**Figure 6-1**). Future work will require further optimization to improve data quality to facilitate more detailed analysis.

6.5 Conclusion

The coordinated behaviors of many single cells, the fundamental unit of biology, determine the overall functional or dysfunctional response to a perturbation. This response may be largely localized within a single organ, span across multiple related organs, or phenotypically unite unique individuals. Here, we have characterized single cell responses across varied biological scales, according to the relevant scale of the responses profiled. We characterized acute liver injury within the murine liver organ; effects of six months on high fat diet in multiple gastrointestinal organs in mice; and dendritic cell response to HIV-1 virus across multiple human elite controllers. Our work has contributed to our knowledge of liver regeneration, obesity-related pathology, and the highly functional immune responses which contribute to elite control of HIV. We have applied recently developed scRNA-Seq technologies to diverse problems in order to explore these areas in a new way, by examining responses to perturbations in each of many single cells, identifying previously unappreciated response groups and activated pathways. Further, we have developed new methods for controlling donor-to-donor variability and nominating targets for rational modulation

of system responses for potentially therapeutic purposes. These approaches could be applied to future studies facing similar challenges.

Rapid development of scRNA-Seq methods over the last several years has made these, and many other exciting studies, possible. An interdisciplinary approach, bringing together the collective expertise in computational biology, technology and many areas of biology has greatly benefited the scRNA-Seq field, including the work presented here. Modern high-throughput techniques have enabled researchers to profile hundreds of thousands of cells in a single project, and the resulting enormous datasets present many analytical challenges and opportunities. New computational approaches and adequate resources are needed to surmount challenges associated with data noise and increasing scale [8]. Additionally, improved analytical methods will need to be developed for other applications, such as making cross-species comparisons and building pseudotime trajectories for non-developmental/non-differentiation time course datasets, such as a type of cell responding to a perturbation then returning to baseline.

With efficient scRNA-Seq techniques now available, future studies will be able to more cheaply and easily explore more cells than ever before, which will make many exciting new discoveries and insights possible in the wide array of biological fields taking advantage of these powerful techniques. Undoubtedly, scRNA-Seq technique and data analysis approaches will continue to evolve over the coming year to address new and ongoing challenges in this exciting field.

6.4 References

1 Biswas, S. & Ghose, S. Divergent impact of gender in advancement of liver injuries, diseases, and carcinogenesis. *Front Biosci (Schol Ed)* **10**, 65-100 (2018).

2 Waxman, D. J. & O'Connor, C. Growth hormone regulation of sex-dependent liver gene expression. *Mol Endocrinol* **20**, 2613-2629, doi:10.1210/me.2006-0007 (2006).

3 Masubuchi, Y., Nakayama, J. & Watanabe, Y. Sex difference in susceptibility to acetaminophen hepatotoxicity is reversed by buthionine sulfoximine. *Toxicology* **287**, 54-60, doi:10.1016/j.tox.2011.05.018 (2011).

4 Durazzo, M. *et al.* Gender specific medicine in liver diseases: a point of view. *World J Gastroenterol* **20**, 2127-2135, doi:10.3748/wjg.v20.i9.2127 (2014).

5 Beyaz, S. *et al.* High-fat diet enhances stemness and tumorigenicity of intestinal progenitors. *Nature* **531**, 53-58, doi:10.1038/nature17173 (2016).

6 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386, doi:10.1038/nbt.2859 (2014).

7 Nostedt, J. J. *et al.* The Effect of Bariatric Surgery on the Spectrum of Fatty Liver Disease. *Can J Gastroenterol Hepatol* **2016**, 2059245, doi:10.1155/2016/2059245 (2016).

8 Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**, 1145-1160, doi:10.1038/nbt.3711 (2016).

Appendix A: Sex-differences in the Murine Liver Response to Acetaminophen

This appendix describes a pilot study conducted in collaboration with Chad Walesky, Carolyn Winston, Wolfram Goessling, and Alex Shalek.

The liver is a highly sexually-dimorphic organ, with varied expression of many genes, proteins and hormones between males and females, and, consequently, disparities in susceptibility to many types of liver damage and disease. For example, females are known to be less sensitive than males to acetaminophen overdose liver damage. In this Appendix, we describe some preliminary work to replicate in female mice the APAP-response work described in Chapter 3, which was done exclusively in male mice. In doing so, we identify PPAR-α target activation that may be involved in protecting females from APAP damage. Interestingly, in females we identify upregulation of many of the functional compensation genes observed in male APAP response in Chapter 3, even though females sustain less damage and have a lesser need for compensation, suggesting the female liver may activate these compensatory pathways earlier and even under less damaging conditions. Additionally, the immune response to APAP varies between males and females. Most notably, we detect differences in the fractional abundance of immune cells between the male and female samples. However, low cell numbers in female samples limits our interpretation of what these shifts may mean. Future work is needed to fully build this dataset to gain clearer insight into sex-differences in APAP toxicity response.

Appendix A.1 Background

The liver is among the most sexually dimorphic organs, with disparities in susceptibility to many liver-related diseases observed between males and females[1]. Males are two to three times more likely to develop hepatocellular carcinoma, more susceptible to viral hepatitis and progress to cirrhosis more quickly with HCV infection, while females are more susceptible to alcoholic liver damage, and ten times as likely to develop primary biliary cirrhosis and four times as likely to develop autoimmune hepatitis[1-5]. Estrogen and other female sex hormones can act as an antioxidant which may help protect the liver from progression of some disease[1]. Additionally, expression of many cytochrome P450s and other metabolic genes – and consequently drug and steroid metabolism kinetics – vary between males and females[6]. For example, females are known to be less susceptible to acetaminophen (APAP) overdose than males. Previous work has found that more rapid recovery of hepatic glutathione levels, a compound consumed in non-toxic APAP metabolism, in females following depletion in APAP-metabolism helps attenuate liver damage [7]. In Chapter 3, we profiled murine livers following APAP or partial hepatectomy injury in exclusively male mice. Here, we extend our exploration of the liver's response to APAP-overdose to male-female differences.

Appendix A.2 Sample and data quality

To explore differences between male and female responses to APAP, we profiled livers from mice, both male and female, at 6 and 24 hours following APAP exposure and untreated using techniques as described in Chapter 3. Histology confirms that female mouse livers sustain less damage (smaller tunnel positive area) than males following equal APAP dosage, as has been reported previously[7] (**Fig A-1**).

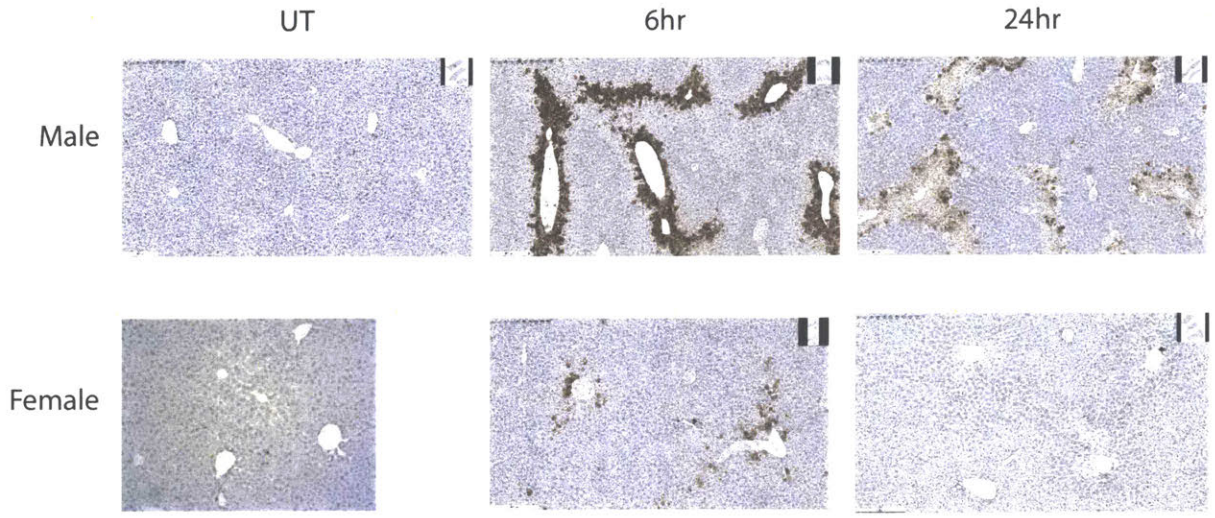

FIGURE A-1 | TUNEL staining in APAP-treated male and female livers.
TUNEL staining on mouse livers untreated, 6 hours and 24 hours post-APAP treatment for male (top) and female (bottom). TUNEL-positive (dark brown) cells are apoptotic.

To gain further insight into sex-differences in APAP toxicity response, we perform Seq-well[8] to generate a pilot dataset containing two untreated samples, one APAP 6hr sample and one APAP 24hr sample from males (subset of data presented in Chapter 3) and females. After sequencing, we aligned data to the mm10 transcriptome to obtain genes by cells matrixes. We then filtered data to include only cells with >500 transcripts read, and >300 genes expressed to eliminate non-cell events and low-quality cells. Overall the female data, particularly the female APAP data, were of much lower quality than the male data, with very few cells recovered from the female samples (**Table A-1**). The sequence alignment rate on the female samples was extremely low (~7% for APAP-treated female, ~15% for untreated female). Low alignment rates often indicate that low input sample quality may have been a problem. The alignment rate for the male samples presented, though better than female, was still lower than what we typically expect for Seq-well alignments (~25-40% for these male samples vs >50% for typical Seq-well alignment). When

processing these single cell suspensions, we first spin slowly to preferentially collect the large hepatocytes (data here and Chapter 3 are all from hepatocyte-enrichment preps). We then take the supernatant and spin it again at faster speed to collect non-parenchymal cells (analyzed in Appendix B). Alignment rates for non-parenchymal samples from untreated males are around 50%, suggesting that hepatocytes are particularly sensitive to processing and likely to have lower alignment rates than typical cells. The high sensitivity of hepatocytes to damage in processing to single cell suspensions has been noted before[9], and is appears to be a feature of this cell type.

The profusion protocol used to process the whole liver to a single cell suspension was developed and optimized with male mice; therefore, further adjustments may be required for optimal use with female animals. The female mice and livers are physically smaller and may benefit from a shorter or gentler profusion. Additionally, we have made some changes to the profusion and processing protocol in our more recent work with male mice (increasing number of cells loaded on array, non-parenchymal cell sample, changed profusion entry point) which improved sample and data quality for males and would likely improve for females as well. Female mice are less susceptible to damage from APAP, so their livers should be healthier than the males when we begin processing them. Since it is possible to obtain quality data from the more damaged male livers, it should be possible to obtain quality data from females as well. Future work will be needed to optimize processing for female livers and expand the female data set to match the male dataset described in Chapter 3.

	Cells passing filter	Alignment Rate
A24F1	29	6.75%
A24M1	859	37.50%
A6F2	121	7.05%
A6M2	685	24.40%
UTF1	700	n.c.
UTF2	183	14.60%
UTM1	1213	37.30%
UTM5	1446	n.c.

Table A-1 | Cell passing filter and alignment rate by sample
n.c. = not calculated

Appendix A.3 Cell type fractional abundance varies by sample type

To visualize data structure, we performed dimensional reduction by principal components analysis (PCA) and t-stochastic neighbor embedding (tsne), followed by clustering using shared nearest neighbors (SNN) (**Fig A-2A-C**). We identify multiple hepatocyte clusters and several non-parenchymal cell (NPC) clusters including Kupffer cells, Neutrophils, B cells, macrophage/monocytes and T cells **(Fig A-2B,C)**. The hepatocyte clusters separate by treatment condition, with separate hepatocyte clusters for untreated male 1 (UTM1); UTM5; 6 hours post APAP treatment male (A24M); 24 hours post APAP treatment male (A24M); and combined all female samples (UTF1, UTF2, A6F, A24F).

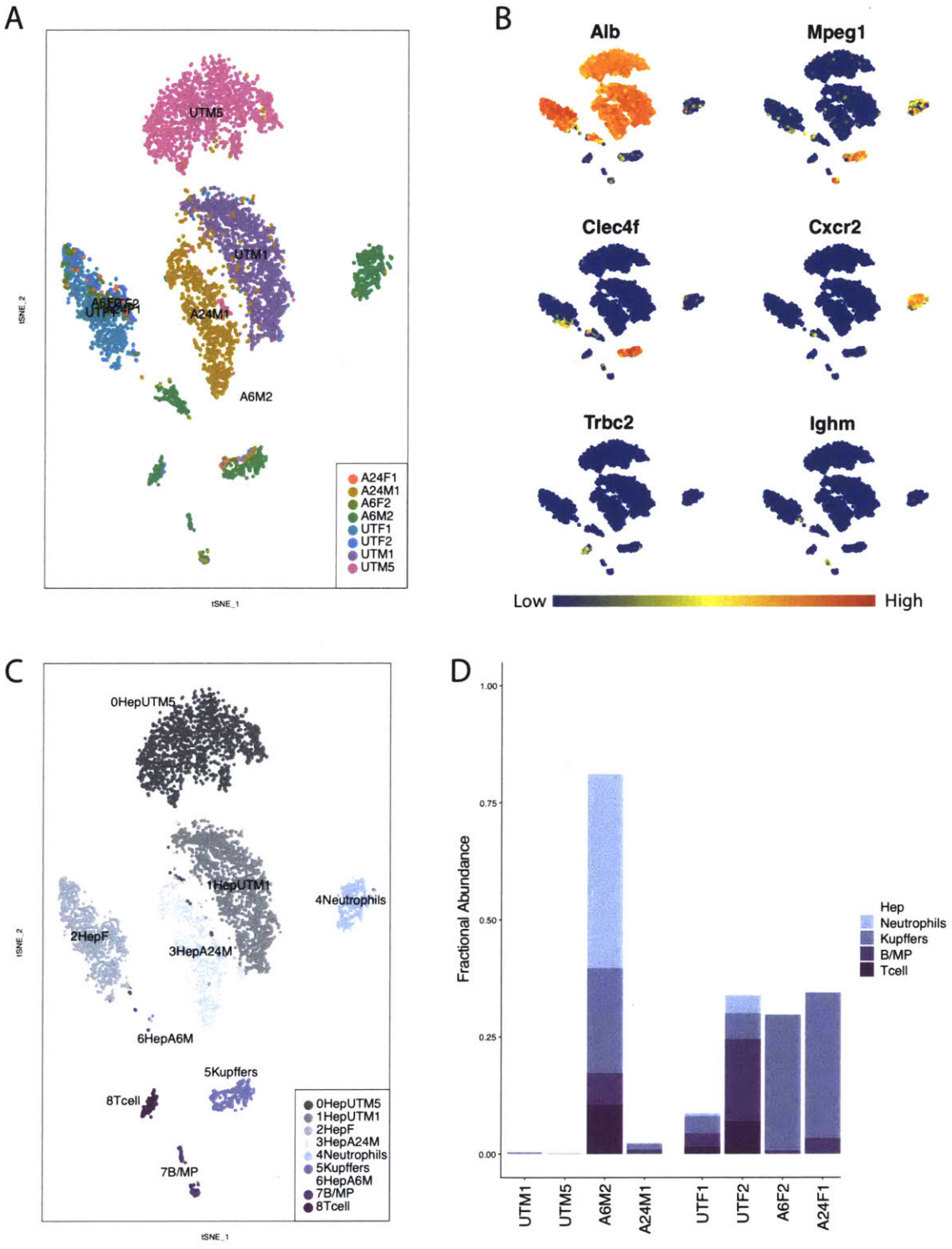

Figure A-2 | Cell types present in male and female APAP-treated liver samples. (A) tsne of all data, colored by sample of origin. **(B)** tsnes colored by expression of marker genes for liver cell types. **(C)** tsne colored by SNN clustering. Each cluster is annotated by cell type. Hepatocyte clusters additionally annotated by main sample(s) of origin. **(D)** Stacked barplot of fractional abundance of each cell type in each sample.

While the spin speed used to prepare these samples enriches for hepatocytes, other cell types are still present in most of the samples. The fractional abundance of each cell type varies greatly from one sample to the next **(FigA-2D)**. In both UTM1 and UTM5, NPCs are virtually non-existent. In contrast, A6M contains the largest share of NPCs and nearly all of the neutrophils. Six hours after APAP treatment we expect immune infiltration to the liver, especially neutrophils, in response to the toxicity-induced tissue damage. Additionally, APAP-toxicity is damaging to the hepatocytes, killing many of them and possibly making others less fit and unable to survive the additional stress of processing, further biasing the sample toward NPCs. By 24 hours post-APAP treatment, the hepatocyte-NPC ratio has nearly returned to normal in males, as immune infiltration decreases and hepatocytes begin to recover. In female samples many more NPCs in untreated samples than in males (8.71% UTF1; 33.9% UTF2; vs <1% UTM1 and UTM5). These differences may stem from suboptimal processing of female samples, biasing data collection away from more sensitive hepatocytes, or it may reflect real biological differences. Interestingly, we observe little change in the NPC fractional abundance from UTF (8.71% and 33.9%) to A6F (29.8%) and A24F (34.5%), although the specific cell types making up the NPC fraction do shift. The female liver is much less susceptible to damage from APAP overdose and may not sustain enough damage to trigger major immune infiltration or significantly compromise hepatocyte quality. Alternatively, immune infiltration into the female liver may already begin to dissipate by 6 hours. Given the low cell numbers available in the female APAP data, it is difficult to interpret possible shifts in NPC cell type abundance and not possible to perform further analysis on these cell types. The immune cell abundance varies considerably between UT female samples (8.71% and 33.9%) which may be related to varied sample quality (700 and 183 cells passing filter). Higher quality female APAP data, more replicates and NPC-enriched samples may be able to address these questions and inconsistencies.

Appendix A.4 Hepatocyte variation across sample conditions

To more closely focus on hepatocyte responses, we subsetted our data to include only hepatocyte clusters, then filtered out all cells with a Hepatocyte Signature score three standard deviations or more below the mean (see Chapter 3 for more detailed explanation). We then further subsetted the dataset to allow a maximum of 85 cells per sample to prevent the much larger male dataset from dominating the analysis (85 hepatocytes filtered in for all samples except samples where fewer cells available: A24F, 19 cells; A6F, 82 cells) for a total balanced hepatocyte dataset of 611 cells.

Dimensional reduction and SNN clustering identified unique individual clusters for each of the male samples, a cluster of untreated female hepatocytes containing both UTF1 and UTF2, and a cluster APAP-treated female hepatocytes containing both A6F and A24F **(Fig A-3A)**. The shared UTF cluster, in contrast to the separate UTM1 and UTM5 clusters, suggests that the female livers may be more similar to one another at baseline. Separation by animal in male untreated samples holds up in the analysis in Chapter 3 which included more cells and more mice, demonstrating that this is a consistent pattern in males. Additional experiments with more female mice and higher quality female data are needed to determine whether untreated female hepatocytes will continue to appear more similar from one animal to one another in transcriptional space than the males. Many of the genes separating untreated male mice are pheromone-related (e.g. major urinary proteins, Mups) **(Fig A-3B)**. We hypothesize that dominance ranking in the cage (mice are house five to a cage) may influence expression of these genes. Possibly, in female mice, dominance rankings may have less of an effect on the liver's transcriptional profile. We have also considered possible effects of circadian rhythm or feeding times on transcriptional liver data, but additional experiments are needed to explore these possibilities.

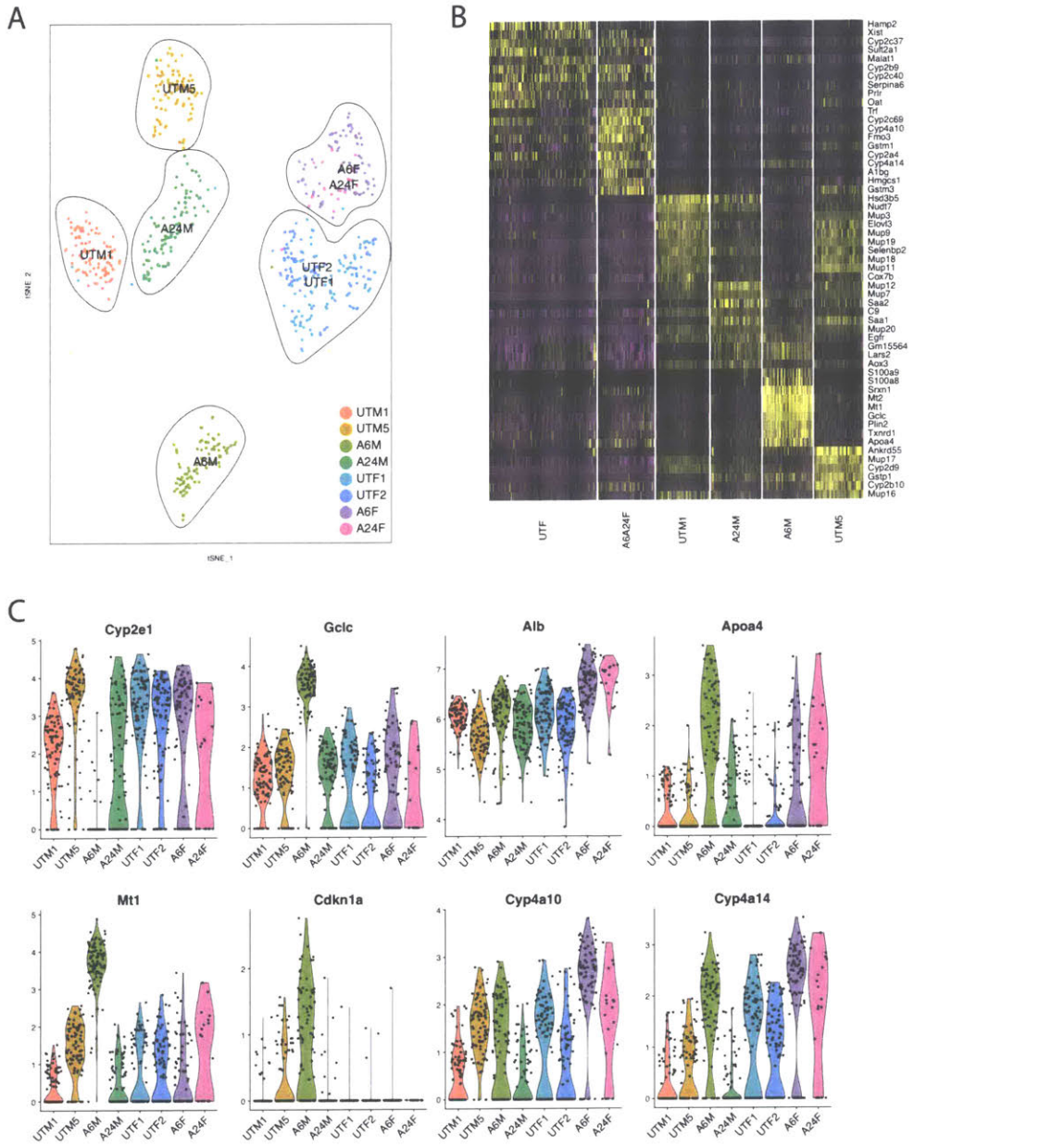

Figure A-3 | Differences in transcriptional expression between APAP-treated males and females.

Legend next page.

Figure A-3 | Differences in transcriptional expression between APAP-treated males and females.
Previous page.
(A) tsne of subsetted hepatocytes, colored by sample condition. SNN clusters outlined in black. **(B)** Heatmap of marker genes for each cluster. **(C)** Violin plots of selected genes.

The APAP-treated samples also cluster differently in the males than in the females, with separate clusters for A6M and A24 M, and a shared cluster for A6F and A24F **(Fig A-3A)**. This may be because the liver damage is more severe in the male, making the A6 sample very distinct as the liver responds to the extensive damage. In contrast, female liver damage is less severe, so there may be a lesser response. The female liver is known to upregulate glutathione synthesis to replenish stores more quickly than males, thus limiting liver damage following APAP overdose[7]. Because the female liver responds more quickly, it is possible that an earlier time point (e.g. 3 hours) is needed to capture the peak female hepatocyte response. Additionally, the very small number of A24F hepatocytes (19 cells) limits the amount of variation they contribute to the total dataset, possibly preventing them from grouping out into their own cluster. Additional experiments to add more cells, especially A24F, are needed to more clearly map female response patterns.

To identify transcript expression unique to each cluster, we ran differential expression and visualized the results as a heatmap. The heatmap reveals sex-specific expression, with expression of many genes shared across multiple female samples (e.g. *Xist*) and multiple male samples (e.g. *Mup17*), but few genes highly expressed in both male and females **(Fig A-3B)**. The A6 male cluster displays a distinct expression pattern, marked by many response genes described in Chapter 3 (*Mt1, Txnrd1, Gclc*) **(Fig A-3B,C)**. The APAP-treated female cluster is responding with several cytochrome P450 genes, but not elevated expression of many of the genes found in A6M. This is likely because the female liver either responds earlier or does not

respond in the same way since it does not sustain as much damage. Remarkably, *Cyp2e1*, the gene that encodes the protein responsible for toxic APAP metabolism, expression is zero in nearly all cells in A6M, but normal in A6F **(Fig A-3C)**. In male mice, the APAP toxicity kills the *Cyp2e1*-expressing cells at six hours (Chapter 3). Strikingly, unlike A6M hepatocytes, A6F hepatocytes do not express elevated levels of *Gclc*, the gene that encodes the protein that synthesizes glutathione, despite literature reports that this is the main mechanism by which females are protected from APAP[7]. These differences could be the result of earlier response and lesser liver damage in the female. Histology shows some cell death in the pericentral regions in A6F, but not nearly as much as in A6M **(FigA-1)**; suggesting that likely some of the *Cyp2e1*-expressing cells survive in the female. Additionally, since the female responds more quickly, it is possible that she has already compensated for the loss of damaged cells by 6 hours, as we have seen in males at 24 hours (Chapter 3). This faster response also means that we may find elevated *Gclc* expression at earlier time points in females rather than at 6 hours.

Although the female does not express elevated levels of response genes (e.g. *Mt1, Gclc, Txnrd1*) we do observe elevated expression of some liver function genes (*Alb, Apoa4*) which we described as part of a functional compensation phase in the males in Chapter 3 **(Fig A-3C)**. This suggests that the females may be functionally compensating, even though they do not sustain as much damage to their livers as the males. The normal level of *Cyp2e1* expression in the A6F hepatocytes may be due, in part, to functional compensation that occurred earlier than 6 hours, in line with the fact that females are known to respond more quickly. Other genes, like *Cdkn1a* (p21), are high only in the male at 6 hours, but may be elevated at earlier time points or not at all in the female **(Fig A-3C)**. p21 prevents cells from entering the cell cycle, preventing hepatocytes from replicating in the A6M condition where generation of reactive oxygen species (ROS) may damage DNA. In females at 6 hours the ROS have likely already dissipated, eliminating the need to halt proliferation.

Finally, we find marked upregulation of *Cyp4a10* and *Cyp4a14* in the APAP-treated females and modest upregulation in the APAP-treated males. Both of these genes are downstream targets of PPAR-α. Activation of PPAR-α is protective against APAP damage through induction of mitochondrial uncoupling protein-2 which decreases phosphorylation of JNK in the mitochondria, decreasing generation of ROS[10,11]. Protection of mice from APAP-induced hepatotoxicity through PPAR-α activation protects mice has been previously reported, however potential sex-differences in activation or this pathway are, to our knowledge, previously unknown.

Appendix A.4 Conclusions and future directions

Here, we have transcriptionally profiled and analyzed male and female livers following APAP treatment. Females sustained less liver damage and did not respond to damage with upregulation of redox, and damage response genes as the males did at 6 hours post-exposure. Interestingly, the females did not upregulate glutathione synthesis (*Gclc*) in our data, despite literature reports that swift upregulation of this pathway is what protects females from APAP-induced liver damage. Females do upregulate some liver function genes and may have initiated functional compensation and other damage responses, as described for males in Chapter 3, earlier than 6 hours. Finally, we uncovered enhanced expression of PPAR-α activation targets in females and nominate this pathway as a possible contributor to female protection from APAP-toxicity.

While our existing dataset provides some insight into sex-differences in APAP-induced hepatotoxicity, additional experiments are needed to optimize female processing and build a high quality female dataset as a companion to the male dataset in Chapter 3. With high quality female

data, we will be powered with more cells and more genes captured to make comparisons between the sexes. The addition of earlier time points for females will address whether they respond with increase glutathione production and functional compensation at an earlier time point. While females have less liver damage than males, and therefore less need for compensation, our existing data suggests that they do compensate to some degree. Additional experiments with a higher dose of APAP for females so as to more closely recapitulate the level of liver damage observed in males will also be informative. Analysis of the full dataset will reveal how male and female functional compensation compare. Molecular pathways which protect females from liver damage from APAP may be explored as therapeutic targets for treatment of APAP-overdose or other types of acute liver failure.

Beyond hepatocyte responses, our limited dataset hints at possible variation in baseline immune populations and immune responses. Some shifts in immune cell populations between samples in males and females are observed here, with more immune infiltration into the male liver at 6 hours post-APAP treatment. All of the data presented here originated from hepatocyte-enriched samples, which contain some immune cells but mostly hepatocytes. We have also collected NPC-enriched samples, in which the majority of cells are immune, from some of our male conditions. Future experiments should collect and profile NPC-enriched samples from these livers to increase NPC numbers in the dataset and explore how these cells may participate in liver damage response. Analysis of existing NPC data from male mice is described in Appendix B.

Extensions of this work will broaden our knowledge of sex-differences in the liver. New experiments may explore sex-differences in response to and metabolism of drugs other than APAP in the male and female liver. Understanding how drugs are differently metabolized in males in females is critical for proper dosing. Further extensions of this work may explore how sex-

differences contribute to disparities in frequencies of many types of liver disease between males and females.

Appendix A.4 References

1 Biswas, S. & Ghose, S. Divergent impact of gender in advancement of liver injuries, diseases, and carcinogenesis. *Front Biosci (Schol Ed)* **10**, 65-100 (2018).

2 Wasley, A. *et al.* The prevalence of hepatitis B virus infection in the United States in the era of vaccination. *J Infect Dis* **202**, 192-201, doi:10.1086/653622 (2010).

3 Durazzo, M. *et al.* Gender specific medicine in liver diseases: a point of view. *World J Gastroenterol* **20**, 2127-2135, doi:10.3748/wjg.v20.i9.2127 (2014).

4 Parikh-Patel, A., Gold, E. B., Worman, H., Krivy, K. E. & Gershwin, M. E. Risk factors for primary biliary cirrhosis in a cohort of patients from the united states. *Hepatology* **33**, 16-21, doi:10.1053/jhep.2001.21165 (2001).

5 Manns, M. P. *et al.* Diagnosis and management of autoimmune hepatitis. *Hepatology* **51**, 2193-2213, doi:10.1002/hep.23584 (2010).

6 Waxman, D. J. & O'Connor, C. Growth hormone regulation of sex-dependent liver gene expression. *Mol Endocrinol* **20**, 2613-2629, doi:10.1210/me.2006-0007 (2006).

7 Masubuchi, Y., Nakayama, J. & Watanabe, Y. Sex difference in susceptibility to acetaminophen hepatotoxicity is reversed by buthionine sulfoximine. *Toxicology* **287**, 54-60, doi:10.1016/j.tox.2011.05.018 (2011).

8 Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* **14**, 395-398, doi:10.1038/nmeth.4179 (2017).

9 MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383, doi:10.1038/s41467-018-06318-7 (2018).

10 Patterson, A. D., Shah, Y. M., Matsubara, T., Krausz, K. W. & Gonzalez, F. J. Peroxisome proliferator-activated receptor alpha induction of uncoupling protein 2 protects against acetaminophen-induced liver toxicity. *Hepatology* **56**, 281-290, doi:10.1002/hep.25645 (2012).

11 Manautou, J. E. *et al.* Clofibrate pretreatment diminishes acetaminophen's selective covalent binding and hepatotoxicity. *Toxicol Appl Pharmacol* **129**, 252-263, doi:10.1006/taap.1994.1250 (1994).

Appendix B: Non-Parenchymal Cell Response to Acute Injury

In Chapter 3 we profile hepatocyte responses to acute liver injuries; here we examine transcriptional responses of non-parenchymal cells. We identify several cell types and examine Wnt expression. Additional samples – particularly untreated and acetaminophen-treated – are needed to complete this dataset.

Appendix B.1 Background

The liver possesses remarkable regenerative capacity, with the ability to return to its original size and maintain function even following major injury from toxic metabolites or surgical resection[1]. In Chapter 3 we profile hepatocyte responses to two acute injury models: a zone-dependent toxic injury, acetaminophen (APAP); and a zone-independent surgical resection, partial hepatectomy (PH). We identify a macrophage derived Wnt-dependent functional compensation phase of liver injury response, division of labor between cycling and non-cycling hepatocytes. Using Wntless-KO models, we demonstrate that macrophage-derived Wnts are required for the functional compensation response in hepatocytes. Previous work has shown that Wnt secretion from endothelial cells within the liver is required for pericentral gene expression and cellular proliferation following injury[2]. Using Wntless-KO models, we demonstrate that macrophage-derived Wnts are required for the functional compensation response in hepatocytes (Chapter 3).

The sequencing analysis in Chapter 3 focuses only on hepatocyte responses, but we have obtained preliminary data for other cell types as well. In processing the liver samples, we first spin the single cell suspensions slowly, preferentially pelleting the large hepatocyte cells. We then spin the supernatant at higher speed to pellet the remaining non-parenchymal cells (NPCs). This varied spin speed technique generates hepatocyte-enriched and NPC-enriched fractions for each sample. Here, we examine the NPC fraction from samples in the acute injury dataset for which this fraction was collected. We identify several cell types and identify preliminary Wnt expression across different groups on NPCs consistent with previous reports.

Appendix B.2 Cell Types

To profile NPC responses to acute injury, we sequenced NPC-enriched samples from mouse livers following no treatment (UT), APAP- treatment or PH. We performed dimensional reduction by t-Stochastic Neighbor Embedding (t-SNE), and identified clusters by shared nearest neighbors (SNN) (**Figure A-1A,B**). The majority of the cells originate from PH samples. In our chronologically earliest experiments, the APAP-treated and some untreated, we did not include the NPC-enriched fraction in the sequencing workflow and did not originally generate this data. The untreated and APAP 96hour samples contained mostly hepatocytes and may be repeated to obtain more NPCs. Experiments to obtain NPC data for additional samples are ongoing.

Calculated module scores for expected cell type are represented in each cluster(**Figure A-1-C**). Liver endothelial cells (LEC) make up cluster 8; Kupffer cells clusters 0, 1, 2, 6, 11; hepatocytes clusters 5, 7, 9, 15, 16, and mixed immune cells cluster 12. Clusters 3, 10, 12, and 13 score highly for both Kupffer and LEC signatures; Kupffer and LECs associate together *in vivo*, generating these doublet clusters. Very few of the UT and A96 cells are NPCs (N.B. the spin protocol enriches for NPCs, but is not a perfect separation. We expect lesser immune infiltration in UT samples than in injury response samples, meaning the UT likely have lower overall numbers of NPCs in the

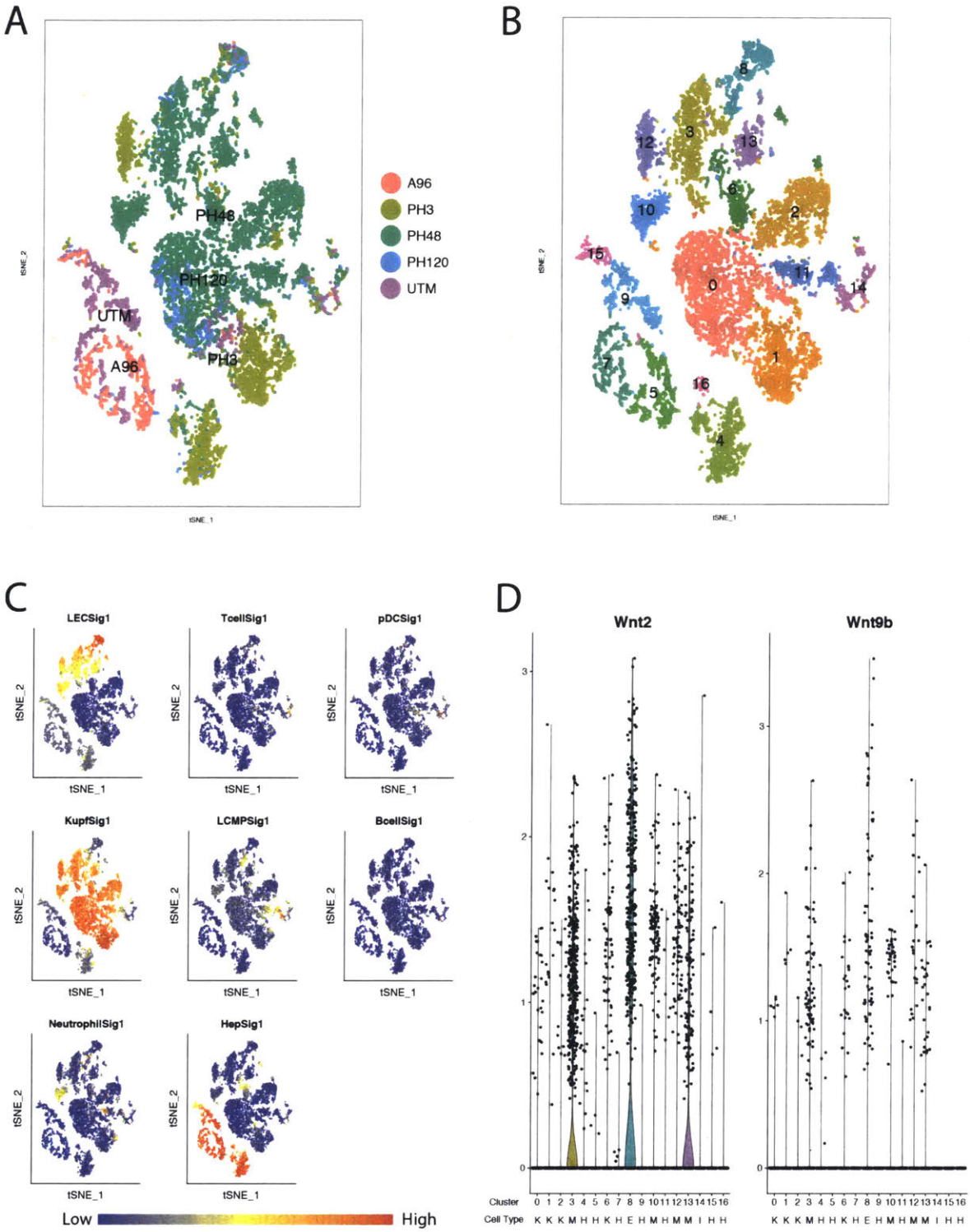

Figure B-1 | Cell types and Wnt expression in the NPC
Legend Next Page.

Figure B-1 | Cell types and Wnt expression in the NPC
Previous Page.
(A) t-SNE or all NPC data. Colored by treatment condition. **(B)** t-SNE colored by shared nearest neighbors (SNN) clustering. **(C)** t-SNE colored by cell type signature score. Blue low, yellow median, red high.**(D)** Violin plots of Wnt2 (left) and Wnt9b (right) expression in each cluster. Cluster number and cell type identity noted below each violin. K, Kupffer; M, mixed Kupffer and liver endothelial cell; H, hepatocyete; I, immune cells other than Kupffer (T cell, B cell, neutrophil, macrophage, pDC).

liver to begin). The mixed immune cluster contains liver capsule macrophages (LCMP), B cells, T cells, Neutrophils and pDCs. These cell types are not abundant enough to drive enough of the total variation to form their own clusters. When the dataset is expanded to include more time points, we may have enough of these cells to generate separate clusters by cell type, whether in the full dataset, or, if needed, by subclustering the "mixed immune" cluster.

Appendix B.3 Wnt Expression

Wnt expression and secretion by NPCs supports pericentral gene expression, proliferation, and functional compensation in hepatocytes[2] (Chapter 3). LEC cells produce primarily Wnt2 and Wnt9b, and these Wnts are secreted and travel to nearby hepatocytes where they support pericentral gene expression, or, following injury, proliferation[2,3]. Macrophages have also been shown to produce primarily Wnt2 and Wnt9b, though at much lower levels than the LECs[2,3]. Indeed, we identify expression of these Wnts in the LEC, Kupffer and mixed LEC/Kupffer clusters in our data (**Figure B-1D**). Wnts other than Wnt2 and Wnt9b were expressed at very low levels or not at all. Unfortunately, we do not have enough LEC or Kupffer cells from untreated samples to determine whether these Wnts are induced in PH. Pervious work has demonstrated the importance of these Wnts originating in LECs for proliferation following PH, but the upregulation of these Wnts in Kupffers/Macrophages and their contribution to functional compensation has not yet been explored.

Appendix B.4 Methods

Samples were prepared as described in Chapter 3. Only nonparenchymal-enriched (faster spin) samples are included here. Data was filtered as described in Chapter 3. The cell type signature scores were calculated using the AddModuleScore in Seurat. The cell type gene lists were from Halpern *et. al*[4].

Appendix B.5 References

1 Michalopoulos, G. K. Hepatostat: Liver regeneration and normal liver tissue maintenance. *Hepatology* **65**, 1384-1392, doi:10.1002/hep.28988 (2017).

2 Yang, J. *et al.* beta-catenin signaling in murine liver zonation and regeneration: a Wnt-Wnt situation! *Hepatology* **60**, 964-976, doi:10.1002/hep.27082 (2014).

3 Preziosi, M., Okabe, H., Poddar, M., Singh, S. & Monga, S. P. Endothelial Wnts regulate beta-catenin signaling in murine liver zonation and regeneration: A sequel to the Wnt-Wnt situation. *Hepatol Commun* **2**, 845-860, doi:10.1002/hep4.1196 (2018).

4 Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat Biotechnol* **36**, 962-970, doi:10.1038/nbt.4231 (2018).