Genomic diversity of tumor and normal tissues revealed by single-nuclei RNAseq computational analysis

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Anugraha Mahalakshmi Raman

Committee in charge:

Professor Bing Ren, Chair
Professor Kun Zhang, Co-Chair
Professor Joseph Ecker
Professor Kelly Frazer
Professor Sheng Zhong

2019
The Dissertation of Anugraha Mahalakshmi Raman is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

Co-Chair

Chair

University of California San Diego

2019
# TABLE OF CONTENTS

SIGNATURE PAGE.................................................................................................................. iii

TABLE OF CONTENTS ............................................................................................................. iv

LIST OF FIGURES .................................................................................................................... vi

LIST OF SUPPLEMENTAL FILES ........................................................................................... vii

ACKNOWLEDGMENTS ............................................................................................................. viii

VITA ........................................................................................................................................... x

ABSTRACT OF THE DISSERTATION ....................................................................................... xi

CHAPTER 1: Introduction ......................................................................................................... 1

  1.1 Single Nuclear RNA-seq............................................................................................... 1

  1.2 Glioblastoma Heterogeneity ....................................................................................... 3

  1.3 References .................................................................................................................... 6

CHAPTER 2: Using intronic reads to deconvolve frozen tissue samples from human heart, and human and mouse brain by single nucleus RNA-seq ......................................... 10

  2.1 Abstract ......................................................................................................................... 10

  2.2 Introduction ................................................................................................................... 11

  2.3 Results .......................................................................................................................... 14

  2.4 Discussion ..................................................................................................................... 20

  2.5 Conclusions .................................................................................................................. 22

  2.6 Figures .......................................................................................................................... 26

  2.7 Supplemental Figures ................................................................................................. 32

  2.8 Methods ......................................................................................................................... 40

    2.8.1 Data generation ....................................................................................................... 40

    2.8.2 Data Analysis ......................................................................................................... 42

  2.9 Acknowledgments ......................................................................................................... 44

  2.10 References .................................................................................................................. 45

CHAPTER 3: Inter and Intra tumor transcriptomic heterogeneity in Glioblastoma ........ 49

  3.1 Abstract ........................................................................................................................ 49

  3.2 Introduction .................................................................................................................. 50
3.5 Experimental Procedures ................................................................. 63
3.5.1 Data generation ........................................................................... 63
3.5.2 Data Analysis ............................................................................... 64
3.6 Figures ............................................................................................ 68
3.7 Supplemental Figures ................................................................. 75
3.8 Acknowledgements ........................................................................... 81
3.9 References ......................................................................................... 82

Appendix: Hydroxymethylation Landscapes in Human Embryonic Cell Lineages and Adult Tissues ................................................................. 86
A.1 Summary ......................................................................................... 86
A.2 Introduction ................................................................................... 86
A.3 Results ........................................................................................... 88
A.4 Discussion ...................................................................................... 94
A.5 Appendix Figures ............................................................................ 98
A.6 Methods .......................................................................................... 115
A.7 Acknowledgements .......................................................................... 120
A.8 References ....................................................................................... 121
# LIST OF FIGURES

Figure 2.1. snRNA-seq analysis of adult mouse forebrain recapitulates spatial neuronal heterogeneity and reveals neural stem cell like cell populations .........................................................26

Figure 2.2 snRNA seq analysis of adult human frontal cortex and hippocampus illustrates glia cell heterogeneity .........................................................................................................................28

Figure 2.3 Intrinsic read based clustering shows increased specificity for cell clustering of single nuclei RNA-seq datasets in human heart ventricles ....................................................30

Supplemental Figure 2.1 Experimental and computational workflow ........................................32

Supplemental Figure 2.2 Clustering comparison using intrinsic and exonic reads vs. intrinsic reads .................................................................................................................................34

Supplemental Figure 2.3 snRNA-seq analysis of adult mouse forebrain recapitulates spatial neuronal heterogeneity and reveals neural stem cell like cell populations ....35

Supplemental Figure 2.4 snRNA seq analysis of adult human frontal cortex and hippocampus illustrates glia cell heterogeneity .........................................................................................................................36

Supplemental Figure 2.5 Intrinsic read based clustering shows increased specificity for cell clustering of single nuclei RNA-seq datasets in human heart ventricles .............38

Figure 3.1: snRNA-seq and snATAC-seq profiles of GBM Patients ........................................68

Figure 3.2: CNV Identification using snATAC-seq profiles ..........................................................70

Figure 3.3: snRNA-seq and snATAC-seq profiles of GBM Patients ........................................72

Figure 3.4 Inter and intra-tumoral transcriptional heterogeneity ..................................................74

Supplemental Figure 3.1 snRNA-seq GBM patient clustering ....................................................75

Supplemental Figure 3.2 Characterization of snATAC-seq identified CNVs ...............................78

Supplemental Figure 3.3 snRNA-seq clustering of tumor cells from all patients ......................79

Supplemental Figure 3.4 Pseudotime Analysis of ASC-Like cells in GBM11 &12 .............80
LIST OF SUPPLEMENTAL FILES

Chapter 2:

Supplementary Table 1 (comparison of snRNA-seq studies)

Supplementary Table 2 (gene signatures for adult mouse forebrain clustering)

Supplementary Table 3 (gene signatures for human frontal cortex/hippocampus clustering)

Supplementary Table 4 (gene signatures for human heart ventricle clustering)

Chapter 3:

Supplementary Table 1 (List of amplified regions identified in GBM patients)

Supplementary Table 2 (List of DE genes between OPC/OPC-Like cells and ASC/ASC-Like cells)

Appendix:

Supplementary Table 1 (GO enrichment for hmC profiles in ES cells and Adult Tissues)
ACKNOWLEDGMENTS

I would like to thank my advisor Bing Ren for his mentorship and encouragement over the years. In his lab, and through this program, I was given the resources to transition from a background in wet-lab synthetic biology to computational biology. I would also like to thank my committee members.

I would like to thank my friends and colleagues, Ramya Raviram and Sebastian Preissl, for their close collaborations and deep insights across both single cell projects I worked on during my PhD. I would also like to thank previous members of the Ren lab, specifically Gary Hon who mentored me during my first few years in the lab, and kept a collaborative, learning environment alive.

Finally I would like to thank my friends and family for their support. Boyko not only provided me with technical and career advice, but also offered a great deal of emotional support over the last 5 years. I had engaging scientific discussions with my friends away from and at UCSD, including Michael Yu, Akshay Balasubramani, Anjana Srivatsan, Leen Jamal, Justin Mih, Ameya Velingker, and Tarun Grover. Dorothy Sears, Michelle Dow, and Margaret Donovan made it possible for me to create and sustain a graduate-women-in science-support group (Scientista) at UCSD, and Dorothy imparted valuable advice on being a woman in STEM. My mentors in the arts, Aler Krishnan and Tim Miller, allowed me to find balance and appreciate the practice of mindfulness.

Chapter 2, in full, is a manuscript in preparation High-throughput single nuclei RNA-seq from limited amounts of frozen human and mouse tissues. Anugraha M Raman, Sebastian Preissl, Zhang Cheng, Song Chen, Zhen Ye, Samantha Kuan, Kun
Zhang, Cathy Barr, Bing Ren. The dissertation author is the first author of this manuscript.

Chapter 3, in full, is also in preparation as *Integrative single cell characterization of tumor heterogeneity in flash frozen GBM tumor sections*. Ramya Raviram*, Anugraha M Raman*, Sebastian Preissl, Clark Chen, Bing Ren. The dissertation author will be the co-first author of this paper. She analyzed the single nuclear RNA-seq data, and Dr. Raviram analyzed the single cell ATAC-seq data.
VITA

2012 Bachelors of Arts, Harvard College  
*Biomedical Engineering; Minor: Global Health & Health Policy*

2019 Doctor of Philosophy, University of California, San Diego  
*Bioinformatics and Systems Biology*

PUBLICATIONS

In Preparation:

*High-throughput single nuclei RNA-seq from limited amounts of frozen human and mouse tissues*  
Anugraha M Raman, Sebastian Preissl, Zhang Cheng, Song Chen, Zhen Ye, Samantha Kuan, Kun Zhang, Cathy Barr, Bing Ren

*Integrative single cell characterization of tumor heterogeneity in flash frozen GBM tumor sections*  
Ramya Raviram*, Anugraha M Raman*, Sebastian Preissl, Clark Chen, Bing Ren
Humans are complex organisms, comprised of many tissues, which are comprised of different cell types. Modern single cell sequencing technologies allow scientists to interrogate epigenetic, genetic and transcriptional profiles of single cells, and challenged them to re-evaluate the meaning of cell type. In this thesis, I define cell type based on clustering of single cell transcriptional and epigenetic signatures-- cells with similar molecular signatures share a common cell type, while cells with different signatures have different cell types. I use this definition to characterize multiple distinct cell types in normal and tumor tissues.
In the first half of this thesis, I describe an analysis workflow that I optimized for single nuclear RNA-seq. I assess its biological interpretability by analyzing deeply characterized human and mouse brains, and then apply it to a less characterized tissue—human heart ventricles. I demonstrate my method’s ability to capture rare and meaningful biological cell subtypes, and describe differences between the right and left ventricles of human hearts at the single cell level.

Next I use this method to analyze flash frozen tumor sections from five glioblastoma patients. In this study, I combine single nucleus RNA sequencing (snRNA-seq) with single cell open chromatin profiling data (snATAC-seq). By combining these two single cell technologies, we can confidently identify cells carrying copy number variation in tumor samples, ask which of these copy number changes are represented at the transcriptomic level, and ultimately characterize the observed inter- and intra-patient heterogeneity.
CHAPTER 1: Introduction

Over the course of my PhD I have transitioned from analyzing populations of cells (bulk data) to analyzing single cells. I began my PhD research analyzing epigenetic modifications, specifically base-resolution hydroxymethylation data, with the goal of understanding this modification’s role in gene regulation. Attempting to integrate such epigenomic data with other high-throughput molecular assays: RNA-seq, whole genome bisulfite sequencing (WGBS), ChIP-seq and ATAC-seq data in adult tissues, I realized that my observations likely represented the biology of the predominant cell type in each tissue. Therefore, I focused my efforts on learning how to analyze single nuclear RNA-seq (snRNA-seq) data. After analyzing snRNA-seq profiles in normal human brain, normal mouse brain, and normal human heart samples, I had the opportunity to apply my skills to a more translational project—characterizing flash frozen glioblastoma sections for which open chromatin and transcriptional signatures were profiled at the single cell level.

1.1 Single Nuclear RNA-seq

Single cell sequencing (DNA and RNA) was chosen as Nature Method’s “method of the year” in 2013. At the time, this choice was motivated by the potential of this technology to uncover tissue heterogeneity, understand the differences between healthy and disease tissues, and to profile rare cell types\textsuperscript{1}. Since 2013, many better single cell techniques have been created to profile cells at the transcriptional level, and large scale efforts are underway to essentially create a “reference human cell atlas ” for different cell types\textsuperscript{2,3}.

The characterization of so many transcriptomes called for a redefinition of \textit{cell type}. Several scientists in the field were asked to define this seemingly common notion,
and their answers varied from redefining cells based on cellular states, to cellular lineages (“cell history”), to grouping cells based on their function within tissues\textsuperscript{4}. While there is no agreement on how exactly cell type should be defined, researchers in the field do agree that defining a cell’s type by its unique transcriptional profile is only one component of that cell’s identity.

Single nuclear methods were developed shortly after the now ubiquitous whole-cell methods to perform transcriptomic sequencing at the single cell level\textsuperscript{5–10}. Single nuclear methods isolate and sequence only nuclear RNA, thereby capturing nascent transcription. While the nucleus contains only 10% of the cell’s mRNA, making snRNA-seq data sparser than scRNA-seq data, single nuclear methods can be more readily performed on frozen tissues, and can capture cell types that are rare or difficult to dissociate from a solid tissue.

Due to this sparsity, a few studies have focused on demonstrating that the transcriptional diversity captured in snRNA-seq data is sufficient, and in some cases even more comprehensive in capturing the heterogeneity observed when performing scRNA-seq\textsuperscript{11–13}. Bakken et al compared 463 cells to an equal number of nuclei from the adult mouse brain (primary visual cortex) to show that they could capture the same eleven cell types in both assays, and that gene expression levels derived from reads mapping to introns from snRNA-seq data shared a strong correlation with gene expression levels in single cells\textsuperscript{11}. Lake et al also showed that snRNA-seq results were consistent with scRNA-seq results of mouse neurons; however, gene expression was concordant for cell type specific and metabolic genes, but not genes enriched for mitochondrial respiration in these cells\textsuperscript{12}. Wu et al showed that snRNA-seq could be
used to capture rare cell types in non-neuronal cells (adult kidney), and did not capture transcriptional heterogeneity associated with stress, which is likely believed to be an experimental artifact introduced by the tissue dissociation process\textsuperscript{13}.

Over the years, single cell transcriptomic assays have scaled exponentially in throughput, while the cost of these technologies has significantly dropped\textsuperscript{14}. While single cell RNA-seq may be preferred to capture certain cell types, such as those with a large cytoplasmic fraction, or to perform certain computational techniques to answer questions about unbiased developmental trajectories,\textsuperscript{15} single nuclear RNA-seq may be necessary to capture rare cell types or to work with frozen tissue. This makes snRNA-seq an excellent candidate to study normal and tumor frozen tissues. In my thesis I use snRNA-seq to profile flash frozen tissues from normal tissues (adult mouse forebrain, human frontal cortex, hippocampus and heart ventricles) and tumor sections (glioblastomas).

1.2 Glioblastoma Heterogeneity

Glioblastoma (GBM) is the most common and aggressive brain tumors with median survival of 12 to 15 months. Common interventions such as chemo-, and radio-therapies have immunosuppressive effects and even surgical excision of the tumor has not shown prolonged success\textsuperscript{16}. Furthermore, patients who have shown recurrence of glioblastoma following treatment usually only live for 6-9 more months, and most do not receive a secondary treatment\textsuperscript{17}. The lack of success in GBM treatment can be attributed to the physical and molecular heterogeneity of GBM.

The most comprehensive attempt to characterize the molecular heterogeneity in GBM was undertaken by The Cancer Genome Atlas\textsuperscript{18,19}. Their pilot study identified three pathways (receptor tyrosine kinase signaling, p53 signaling, and RB signaling) as
core pathways mutated in GBM\textsuperscript{18}. Soon after, four molecular subtypes of GBM (proneural, neural, mesenchymal and classical) were defined by a set of 840 genes\textsuperscript{19}. GBMs of the classical subtype commonly show amplification of $EGFR$ and express neural precursor and stem cell markers. The mesenchymal subtype highly expresses both mesenchymal and astrocytic markers, has the worst survival of the four subtypes, and its microenvironment has been shown to be enriched for tumor promoting M2 macrophages\textsuperscript{20}. The proneural subtype carries mutations in $PDGFRA$ and $IDH1$, highly expresses oligodendrocyte development genes, and is associated with the best patient outcomes due to frequent mutation of $IDH1$\textsuperscript{21}.

With the increasing affordability and throughput of single cell technologies, several studies have used the single cell lens the characterize GBM heterogeneity. In one of the first single cell transcriptomic profiling studies of GBM, cells within five GBM tumors carried a mixtures of signatures from the four molecular subtypes previously defined\textsuperscript{22}. Additionally, among these five patients, an increase in molecular heterogeneity within a given tumor was associated with worse survival. Single cell DNA profiling in GBM has also been used to reveal clonal sub populations and to characterize differences between cells with varying proximity to the tumor core\textsuperscript{23,24}. More recently, a study attempting to investigate the glioblastoma cell of origin used single cell sequencing to show that cells carrying mutations in the tumor free subventricular zone (SVZ) of a patient’s brain can give rise to the glioblastoma tumor\textsuperscript{25}.

Our current understanding of GBM inter/intra-tumoral heterogeneity is focused on the transcriptomic and genomic aspects, and is insufficient for creating successful therapies\textsuperscript{26}. To address this gap, the epigenetic intra-tumoral heterogeneity of GBM,
specifically the open chromatin profile, needs to be characterized. ATAC-seq is an epigenetic assay that can reveal active regulatory elements by profiling open chromatin\textsuperscript{27}. The utility of this assay has increased after several approaches computationally linked active regulatory elements to their target genes at the single cell level\textsuperscript{28,29}. While ATAC-seq has been used to characterize GBM at the bulk level, it has yet to be applied to GBM at the single cell level\textsuperscript{27,30}. In my thesis, I combine single nuclear profiling of RNA and open chromatin in flash frozen tumor sections, from five GBM patients, to characterize inter and intra patient heterogeneity.
1.3 References


CHAPTER 2: Using intronic reads to deconvolve frozen tissue samples from human heart, and human and mouse brain by single nucleus RNA-seq

2.1 Abstract

Single cell (scRNA-seq) and single nuclei RNA-seq (snRNA-seq) are powerful tools to dissect tissue heterogeneity. Isolation of intact cells, like neurons and cardiomyocytes, from frozen tissues is very challenging; however, tissues are amenable to isolation of intact nuclei. In this study, we generated snRNA-seq profiles for ~24,000 single nuclei from flash frozen brain and heart tissues. SnRNA-seq profiles are significantly sparser than those generated from scRNA-seq, and nuclear transcriptomes can be biased by spliced cytoplasmic RNA molecules originating from highly expressed genes, which stick to nuclei of other cells. By using a customized computational workflow and focusing on intronic reads for nuclear clustering analysis, we overcome these limitations, and digitally dissect adult mouse forebrain, adult human frontal cortex and hippocampus, and human cardiac left and right ventricles. Interestingly, our intronic read based clustering approach increases sensitivity and sub-clustering for both brain and heart tissues compared to clustering approaches that use exonic reads or a combination of exonic and intronic reads. Moving forward, this approach could be useful for unbiased characterization of a variety of frozen biobanked clinical samples from control and diseased patients.
2.2 Introduction

Single cell RNA-seq (scRNA-seq) is a powerful tool to dissect tissue heterogeneity\textsuperscript{1–5}. However, it is challenging to perform scRNA-seq on primary human tissues since these samples are typically preserved by flash freezing, making it either impossible or extremely difficult to isolate intact cells, like neurons or cardiomyocytes, from frozen tissue specimens. Additionally, isolation of cells often requires enzymatic digestion of tissue that can alter the transcriptome. Nuclei, on the other hand, are more stable, and amenable to isolation from frozen tissues using mechanical tissue homogenization.

Recently, multiple single nuclei RNA-seq (snRNA-seq) experimental protocols have overcome these hurdles in a variety of tissues (Supplementary Table 1)\textsuperscript{2,4–13} Each of these protocols has its own advantages; of note, droplet based protocols from 10x genomics are easy to implement and demonstrate high cell recovery, and protocols such as SPLiT-seq are able to sequence up to a million cells in one experiment using combinatorial barcoding based approaches\textsuperscript{2,14}. Furthermore, 3’ counting based approaches have improved transcript quantification by implementing unique molecular identifiers (UMIs), which can differentiate between pcr duplicates and true transcript copies. Full-length transcript profiling protocols have also been created, and they have tackled hurdles related to genome coverage. These protocols allow for deep characterization of each cell with the caveat of lower cell numbers that can be processed in parallel, and the inability to incorporate UMIs (Supplementary Table 1)\textsuperscript{15}.

For some experiments snRNA-seq is necessary to profile specific cell types in a tissue; however, it is also important to know if these profiles are comparable to scRNA-
seq profiles. Recent studies have not only shown comparable transcriptomic profiles between cells and nuclei isolated from frozen biopsies, but have also suggested that snRNA-seq can sometimes be a less biased approach when trying to profile cells that are easily ruptured in the single cell isolation process\textsuperscript{16–18}. When used to profile the adult kidney, for instance, snRNA-seq not only captured a greater number of cell types than scRNA-seq, but also, lacked transcriptional signatures associated with stress, which are likely the byproduct of the tissue dissociation process\textsuperscript{18}.

While many challenges in the experimental design of isolating and sequencing transcriptomes of nuclei have been addressed, computational challenges that arise from sequencing the nucleus, rather than the whole cell, are only beginning to be addressed\textsuperscript{12,16,17}. Since nuclei contain about 10\% of a cell’s total RNA, data generated from single nuclei versus single cells are sparser. Additionally, the number of reads mapping to intronic regions can be higher than those mapping to exonic regions, due to the abundance of unspliced nascent RNA in the nucleus. In order to address the sparsity of the data and increase the information content that each cell can provide, several recent studies have incorporated both intronic and exonic reads when analyzing single nuclear transcriptomic data\textsuperscript{12,16,17}. Apart from capturing an unspliced nascent RNA in the nucleus, snRNA-seq also captures free floating, cytoplasmic, highly expressed, spliced RNA transcripts. Current studies interpreting snRNA-seq data do not discuss how to deal with the erroneous signal from these spliced RNA transcripts.

Computational analysis of single cell or single nuclear transcriptomic data typically involves a workflow that starts with quality control, and ends with unsupervised clustering of individual cells in order to resolve cell types\textsuperscript{19}. In this study we optimize the
computational workflow for snRNA-seq data, and profile frozen human cardiac left and right ventricular tissue, adult human frontal cortex and hippocampal tissue, and adult mouse forebrain tissue. Specifically, we address the issue of free floating, cytoplasmic, highly expressed, spliced RNA transcripts by using only intronic reads to inform the unsupervised clustering of individual cells; due to the aforementioned data sparsity of snRNA-seq data, we quantify gene expression in ~24000 single nuclei using intronic and exonic reads. Using this approach we digitally dissect the cellular composition of brain and heart tissues.
2.3 Results

snRNA-seq analysis of adult mouse forebrain recapitulates spatial neuronal heterogeneity and reveals neural stem cell like cell populations

A well established but labor-intensive method to isolate nuclei from frozen brain tissue is to use dounce homogenization in combination with density gradient centrifugation. Therefore, we asked if isolating nuclei by grinding brain tissue using a mortar and pestle, followed by permeabilization, would provide a straightforward alternative to nuclei isolation. Using this strategy, we performed single nucleus RNA-seq analysis of frozen forebrain tissue isolated from 8 week old mice for two biological replicates (P56, Fig. 1, Supplementary Fig 1a-c).

In total, 9,137 nuclei from both replicates (3,719 and 5,418 nuclei respectively) passed quality filtering (see methods). Since the number of transcripts detected in single nuclei is lower than that of single cells, we modified the computational analysis workflow for scRNA-seq to account for the data sparsity (see methods). We used sparse singular value decomposition to reduce the dimensions of the expression matrix (Supplementary Fig. 1d). Due to the abundance of unspliced transcripts, more than half of sequenced nuclear RNA is comprised of intronic RNA molecules. Furthermore, free-floating cytoplasmic mRNA molecules of highly transcribed genes are potential sources of noise and background signal in clustering analyses of snRNA-seq data. In order to account for this, we performed clustering using only intronic reads (see methods, Supplementary Fig 2a-c). Due to data sparsity we combined both intronic and exonic reads to recover more transcripts and include single exon genes when performing differential expression analysis. As expected, genes which are recovered when adding back in exonic reads are primarily those with a larger 3’ end or those with a
single exon (Supplementary Fig 2d). Using this workflow, we detected on average 620 genes and 1,074 transcripts per nucleus for the two replicates.

We identified 14 distinct clusters representing major cell classes including excitatory neurons, inhibitory neurons, oligodendrocytes (ODG), oligodendrocyte progenitor cells (OPC), astrocytes (AC), and microglia based on marker gene expression and gene ontology analysis (Fig. 1a,b,c, Supplementary Table 2). In addition, the identified cell clusters and their transcriptional signatures from our analysis were comparable to data generated by DroNc-seq in the mouse hippocampus and cortex (Fig. 1d)11.

To understand the depth of neuronal heterogeneity we could capture, we first annotated 6 excitatory and 3 inhibitory neuron clusters based on expression of Slc17a7, encoding the vesicular glutamate transporter 1, and Gad1, encoding glutamate decarboxylase 1 (Fig 1a,b). Since the forebrain encompasses multiple functionally distinct brain regions like the cerebral cortex, hippocampus, striatum, olfactory bulb and thalamus, we asked if the neuronal clusters represent different brain regions. Further analysis of these clusters readily revealed cluster specific expression of genes for each of the excitatory neuron clusters, enabling us to assign them to cortical layers 2-4 (cluster-specific gene: Enpp2), layers 4-5 (Rorb), layer 6 (Ctgf, Foxp2), hippocampal regions CA1/3 (Neurod6), the dentate gyrus (DG) (C1ql2), and the thalamus (Prkcd) (Fig. 1a-d, Supplementary Table 2)3,21,22.

We then identified four main clusters of inhibitory neurons (MSN, IN1, IN2, INOB) (Fig. 1a, b)3,21,22. One of these clusters, showed high expression of Ppp1r1b, encoding the Protein Phosphatase 1 Regulatory Inhibitor Subunit 1B, a marker for medium spiny
neurons in the striatum (MSN, Fig. 1a, Supplementary Fig. 1b). IN1, however, showed high expression for genes associated with inhibitory neurons of the olfactory bulb (Meis2, Cpa6) (Fig 1a)\textsuperscript{3,21}. To delineate the heterogeneity of the largest inhibitory neuron cluster, IN2, we performed another round of clustering focusing on these nuclei. This second round revealed eight sub-clusters of which four showed specific expression of canonical marker genes of inhibitory neuron subtypes like Vip, Calb2 and a mixture of Pvalb, Sst and Npy\textsuperscript{11}(Fig. 1e). A third round of clustering clearly separated the later into three distinct groups. Of note, the Npy expressing sub-cluster was enriched for additional markers of CGE-derived neurogilaform cells like Pnoc and Kit in contrast to the other two clusters\textsuperscript{3,21} (Supplementary Fig. 3a). A fourth cluster, IN OB, with very few cells expressing the inhibitory neuronal marker Gad1, was enriched for the bHLH-transcription factor encoding gene, Id2, the Activin A encoding gene, Inhba, Trdn, Ndnf and the biological process of neurogenesis (p=3.16e-2). These genes are markers for inhibitory neurons located in the olfactory bulbs of mice\textsuperscript{3,21}.

Remarkably, we also identified one cluster that was enriched for transcription factors Pax6 and the T-box factor Eomes, respectively, which are involved in neurogenesis (Fig. 1e). In order to further characterize the Eomes+ population, we compared these nuclei to gene expression data from adult mouse neural progenitor cells (NPC)\textsuperscript{24}. The Eomes+ nuclei clustered predominantly with NPCs and activated neural stem cells (aNSCs) (Supplementary Fig. 3c). Taken together, our analysis of snRNA-seq of the adult forebrain, which uses intronic read counts and iterative clustering, identified 23 transcriptionally distinct clusters that reflect different cell types and region specific transcriptional profiles.
snRNA seq analysis of adult human frontal cortex and hippocampus illustrates glia cell heterogeneity

We next sought to profile archived frozen human brain samples. In total, we profiled 10,609 nuclei from human frontal cortex (6,478 nuclei) and hippocampus (4,131 nuclei). On average we detected 761 genes and 960 transcripts per nucleus. To perform downstream analysis, we combined data from both regions and identified 14 clusters representing major classes of neuronal and non-neuronal cell types (Fig. 2a). Some of the populations were more prevalent either in the hippocampus or frontal cortex (Fig. 2b, c). We annotated clusters as described above based on expression of canonical marker genes (Supplementary Fig. 4a). Furthermore, we identified sets of genes enriched for each of the clusters (Fig 2d, Supplementary Table 3). The clusters and their transcriptional profiles were in overall concordance with single nuclei RNA-seq data from human hippocampus and frontal cortex using the DroNc-seq platform (Supplementary Fig. 4b)\textsuperscript{11}.

We identified two inhibitory and two excitatory neuron clusters in the merged data. Of note, the fraction of neurons was very low in the profiled hippocampus sample (Fig 2b). Therefore, the transcriptional profiles of the two excitatory neuron clusters predominantly resembled those of cortical projection neurons and subcortical projection neurons, as previously described\textsuperscript{12} (Supplementary Fig. 4c). The inhibitory neurons exhibited similar transcriptional profiles to VIP+ and SST+ interneuron subtypes described previously \textsuperscript{8} (Supplementary Fig. 4c).

The analysis also revealed two distinct clusters of astrocytes (ACI and II, Fig. 2a,e). In order to understand what could be separating these two groups of astrocytes, we asked if they could come from different regions of the brain. Astrocytes can be
distinguished by their location in grey and white matter as well as their function and morphology into protoplasmic and fibrous astrocyte populations\textsuperscript{25,26}. Protoplasmic astrocytes, found in grey matter, have many processes and are connected by gap junctions; however, fibrous astrocytes, found in white matter, have fewer processes. Expression of \textit{GFAP}, which encodes the glial fibrillary acidic protein, can be used as a molecular marker to distinguish astrocytes from grey and white matter, since it is higher expressed in white matter\textsuperscript{26}. In line with this observation, \textit{GFAP} was the top differentially expressed gene between AC I and AC II, with higher expression in AC II (Fig. 2e, Supplementary Fig. 4a). Visualization of top differentially expressed genes, including previously described grey and white matter markers also suggested that AC I and AC II populations are likely representative of cells derived from grey and white matter, respectively (Fig. 2e). Next, we intersected the list of differentially expressed genes between AC I and II with bulk rna-seq data of white and grey matter astrocytes from the human frontal lobe\textsuperscript{27}. This showed significant enrichment of the bulk grey matter transcriptional profile for AC I ($p = 7.74e^{-06}$), and enrichment of the bulk white matter transcriptional profile for AC II ($p = 5.32e^{-4}$). Genes overexpressed in AC I exclusively overlapped with grey matter specific genes from bulk RNA-seq; whereas AC II exclusively overlapped with white matter specific genes from bulk RNA-seq (Supplementary Fig. 4e). Previously, two astrocyte subtypes have been described in mice: a \textit{Gfap}+ population originating from the glial limitans region of the brain, and a \textit{Mgfe8}+ population consisting primarily of astrocytes evenly distributed around the cortex\textsuperscript{23}. Based on the set of marker genes described, the two astrocyte subtypes
identified in this study might resemble transcriptional signatures of these populations (Supplementary Fig. 4d).

We also revealed four clusters of oligodendrocytes (ODG I-IV) and two clusters of oligodendrocyte precursor cells (OPC I, II) (Fig. 2a). Three of the ODG clusters were shared between hippocampus and frontal cortex whereas ODG II was only present in the hippocampus (Fig. 2b, c). In addition, cortical and hippocampal OPCs were also strikingly different (Fig. 2b,c).

Oligodendrocyte heterogeneity in juvenile and adult mouse brains has been attributed to maturation of OPC to ODG \(^28\). Therefore, we asked if the differences between ODG clusters could be explained by ODG maturity in the human samples. Differential expression analysis between the four ODG clusters (Supplementary Table 3) revealed significant enrichment for genes associated with the most mature oligodendrocytes found in the mouse brain (p = 0.02) in the hippocampus-specific ODG II. Next, we performed pseudotime ordering of cells from the ODG and OPC clusters. The end points of the trajectory were occupied by OPC I, II and ODG II nuclei. Nuclei from ODG I, III and IV were located on the trajectory in between these two extremes. The four ODG clusters also showed significant differences in the median pseudo-times, a proxy for ODG maturity level (p < 0.005). Based on this pseudotime analysis oligodendrocyte maturation seems to be reflected from young to mature in the order ODG III, ODG IV, ODG I and finally ODG II (Fig 2f).

Next, we performed a comparison between the human and mouse data. We focused our analysis on cell clusters represented in both the merged data from human frontal cortex and hippocampus as well as the mouse forebrain. We excluded excitatory
neurons from our analysis since the majority of these cells in the human dataset originated from the frontal cortex, and did not represent the diversity of excitatory neurons identified in the mouse forebrain. Using nuclei from MG, AC, ODG, OPC, IN2 in the mouse data set for the comparison, we first identified cluster-specific genes and then analyzed expression of orthologous genes in the human data. The cell type specific expression of these genes was concordant between the two species (Fig 2g). We also further characterized AC and ODG clusters from the mouse forebrain. The transcriptional profile of mouse astrocytes was more similar to human white matter astrocytes ($p = 0.03$, fisher's-exact), and the transcriptional profile of mouse oligodendrocyte was similar to the more mature oligodendrocyte clusters by visual inspection (Fig. 2g).

**Intronic read based clustering shows increased specificity for cell clustering of single nuclei RNA-seq datasets in human heart ventricles**

Finally, we performed snRNA-seq on human cardiac ventricles. While single cell RNA-seq has been performed on mouse ventricles for the non-cardiomyocyte fraction and limited number of cardiomyocytes from human and mouse hearts using a plate based approach $^{29,30}$, there are no reports on high-throughput droplet-based single nuclear RNA-seq from frozen human cardiac ventricles. Several challenges exist in scRNA-seq profiling of cardiac tissue: it is difficult to encapsulate cardiomyocytes into droplets due to their size, and it is extremely challenging to isolate intact cardiomyocytes from frozen specimens. One solution is to purify cardiac nuclei prior to processing with a droplet based microfluidics platform. Based on our previous experience working with cardiac tissues we found that tissue grinding does not
efficiently release cardiomyocyte nuclei. Thus, we have optimized nuclei isolation for snRNA-seq using the gentleMACS platform (see methods).

We applied our technique to 4720 nuclei from human right (1927 nuclei) and left cardiac ventricles (2793 nuclei). On average we detected 384 genes and 502 transcripts per nucleus, allowing us to reveal 9 cell types (Fig 3a). By focusing our clustering analysis on intronic reads, we observed significantly better cell-type separation than we would have using only exonic, or exonic and intronic reads (Fig 3a). Notably, in the heart tissue, an intronic read driven clustering approach also allowed us to observe cluster specific expression of cardiomyocyte specific markers *MYH6* and *TTN* (Supplementary Fig 5a). This is in line with our initial hypothesis, that highly abundant RNA sticking to the outsides of isolated nuclei might confound clustering analysis.

Based on analysis of marker genes we annotated all major ventricular cell types including (cardiomyocytes, adipocytes, endothelial cells, macrophages, fibroblasts, schwann cells, immune cells (T/NK cells), and a mixture of smooth muscle/pericytes) (Fig 3a,b). Cells from each ventricle clustered together; however, adipocytes were primarily derived from the right ventricular sample (Fig. 3c). Cell type specific gene expression of differentially expressed genes identified from our transcriptomic analysis, coupled with GO enrichment analysis, further validated cell type classifications for these nuclei populations (Fig 3b,d, Supplementary 5b, Supplementary Table 4). We also performed gene set enrichment analysis on adipocytes, smooth muscle cells, macrophages  and NK/T cells,  all of which had broad GO enrichment terms, and found that these clusters were enriched for the hallmark adipogenesis genes (*p*=4.35 e-17), vascular smooth muscle contraction (*p*=6.75 e-9), hallmark inflammatory response
genes ($p=7.47 \times 10^{-9}$), and hallmark interferon gamma response genes ($p=2.54 \times 10^{-3}$), respectively.

Next, we asked if there were differences between left and right ventricular cardiomyocytes. Subclustering of the cardiomyocyte cluster showed that the primary difference between these cells was their origin (left or right ventricle) (Supplementary Fig. 5c). Gene set enrichment analysis on genes differentially expressed between these two clusters, indicated that left ventricular cardiomyocytes were enriched for heart contraction, actin mediated cell contraction, and hallmark myogenesis genes; whereas, right ventricular cardiomyocytes were enriched for regulation of blood and circulation; regulation of blood pressure, and circulatory processes (Supplementary Fig. 5d, Supplementary Table 4). In addition, we observed that the number of transcripts and genes captured in left ventricular cardiomyocytes was significantly higher than the number of transcripts and genes captured in right ventricular cardiomyocytes, despite similar sequencing depth (Supplementary Fig. 5e).

### 2.4 Discussion

Single nuclei RNA-seq is a valid alternative to single cell RNA-seq\textsuperscript{6,16–18}. However, sparsity of nuclear RNA-seq data and free-floating mature cytoplasmic RNA molecules pose challenges for analysis. Compared to previous reports, which use exonic or a combination of intronic and exonic reads for clustering, we show that clustering based on intronic reads is advantageous when profiling tissues with high expression of cell-type specific RNAs. In a recent comprehensive comparison of mouse brain single nuclei to matched single cells, it has been reported that the majority of intronic reads from single nuclei map to large genes with tissue specific expression,
implying that these reads likely contribute to biological cellular heterogeneity \cite{16}. In line with this observation, we observe a much better separation of cell types when performing unsupervised clustering on gene expression data generated from intronic reads alone, versus both intronic and exonic reads. Ideally, by using intronic reads alone, we would be able to perform differential expression analysis between our clusters, and annotate our cell types using canonical marker genes. However, many genes, such as those with a large 3’ end, or single-exon genes, are only captured when using reads mapping to exonic regions. Therefore, due to data sparsity, we bring back exonic reads for downstream analysis in order to annotate cell types with more confidence. We observe that relative gene expression differences are still preserved between cell clusters, when including exonic reads which makes this approach viable for accurate cell type classification.

The presented snRNA-seq approach allows us to capture diversity in the mouse and human brain and human heart. In particular, using an intronic-read-based clustering approach compared to an approach using both intronic and exonic reads allows us to computationally reveal the cellular heterogeneity in these samples. For the adult mouse forebrain, this allows to capture previously described spatially diverse neuronal subtypes. In the human frontal cortex and hippocampus, we are able to tease apart excitatory neurons into subcortical and cortical projection neurons, inhibitory neurons into those that are expressing SST and VIP, and separate out oligodendrocytes based on their maturity level assessed by pseudotime analysis. The detected neuronal subtypes are just a few of many subtypes that have been previously characterized in the human brain using snRNA-seq\textsuperscript{12}. One of the reasons for the observed lower
complexity might be the limited cell number profiled as compared to this previous study. However, the type of heterogeneity that we observe in maturing oligodendrocytes has been described only in the mouse brain\textsuperscript{28}. In a study of the human brain, Lake et al are able to place oligodendrocyte progenitor cells and oligodendrocytes into a developmental trajectory; however, they were unable to separate the mature Oligodendrocyte (mOli) population into the mature Oligodendrocyte (mOL) populations observed in the mouse brain\textsuperscript{10}. We do not capture all of these mOL populations either; however, our analysis suggest that our hippocampal specific ODG cluster is similar to the most mature mOL population previously described in mouse, and the other ODG sub clusters we observe come from less mature mOL populations. Finally, in the human heart we identified the major cardiac cell types. The non-cardiomyocyte cell types that we can characterize in human heart sample like immune, schwann, and endothelial subtypes, have been previously reported in a study of the non-cardiomyocyte cells of the mouse heart\textsuperscript{30}.

2.5 Conclusions

In summary, we present an intron read based approach that is more sensitive to deconvolve frozen tissue samples like brain and heart samples into their cellular constituents as compared to approaches including exonic reads. Not only can these single nuclei RNA-seq data be used to infer cell types, but they can also be used to track cell trajectories, best demonstrated in our analysis of human oligodendrocyte differentiation and maturation. As comprehensive references to characterize cell types form different tissues grow, data sparsity will become less of a challenge for cell type
annotation, and this snRNA-seq technique will be useful in the characterization of a variety of frozen biobanked clinical samples from control and diseased patients.
2.6 Figures

Figure 2.1. snRNA-seq analysis of adult mouse forebrain recapitulates spatial neuronal heterogeneity and reveals neural stem cell like cell populations

a T-SNE visualization of mouse adult forebrain annotated with cell type and number of cells (left); b Scaled UMI expression of inhibitory and excitatory neuron markers (Gad1, Slc17a7); c Heatmap plotting row based z-score of average UMI expression in each cluster. Enriched GO terms for cluster specific genes are on the right hand side; d Spearman correlation to previously published data (see methods); e Sub-clustering of cortical and hippocampal interneurons into 8 clusters including an Eomes+ cluster; f Sub clustering of Pvalb+/Sst+/Npy
Figure 2.2 snRNA seq analysis of adult human frontal cortex and hippocampus illustrates glia cell heterogeneity

a T-SNE visualization of human adult hippocampus and frontal cortex annotated with cell type and number of cells; b T-SNE visualization showing cells that come from the hippocampus. Circled clusters predominantly contain cells from the hippocampus sample; c T-SNE visualization showing cells that come from the frontal cortex. Circled clusters predominantly contain cells from the frontal cortex sample; d Heatmap plotting row based z-score of average UMI expression in each cluster. Enriched GO terms for cluster specific genes are on the right hand side; e Scaled UMI gene expression of grey and white matter astrocyte specific markers projected onto identified astrocyte clusters; f Predicted single-cell trajectory rooted at the OPC I/II cells (top); Violinplot of pseudotime estimates for cells in ODGI-IV and OPC I, II. Dashed lines represent different quartiles, and the differences between the median estimated pseudotime for ODGI-IV are all statistically significant (bottom); g Heatmap displaying expression of cell type-specific genes in the mouse (left). Heatmap displaying expression of orthologous human genes in corresponding cell types (right). Values plotted in the heatmap are the row-based z-score of average UMI expression in each cluster.
Figure 2.3 Intrinsic read based clustering shows increased specificity for cell clustering of single nuclei RNA-seq datasets in human heart ventricles

a T-SNE visualization of human heart ventricles annotated with cell type and number of cells (left); T-SNE visualization of clustering results of human heart ventricles when using exonic and exonic+intronic reads (right). b Heatmap plotting row based z-score of average UMI expression in each cluster. Enriched GO terms for cluster specific genes are on the right hand side; c Cells in each cluster plotted by sample of origin (top). Tissue specific clusters are circled; Stacked bar plot showing the percentage of cells from each ventricle in each identified cell type (bottom); d Violin plots of UMI for cell type specific genes using gene expression values from all reads.
a) Human Heart Ventricles

Cluster Specific Gene Expression

b) Cluster Composition

c) Marker Gene Expression

Abbreviation | Description
---|---
NK/T | Natural Killer/T Cells
MC | Macrophage
END | Endothelial Cells
IM | Smooth Muscle
CM | Cardiomyocyte

<table>
<thead>
<tr>
<th>Gene</th>
<th>Immune system process</th>
<th>Immune system process</th>
<th>Biological adhesion</th>
<th>Vasculature Development</th>
<th>Schwann cell differentiation</th>
<th>Extracellular matrix</th>
<th>Circulatory system development</th>
<th>Small molecule metabolic process</th>
<th>Cardiac muscle tissue development</th>
</tr>
</thead>
<tbody>
<tr>
<td>NK/T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>END</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

d) Marker Gene Expression

<table>
<thead>
<tr>
<th>Gene</th>
<th>NK/T</th>
<th>MC</th>
<th>END</th>
<th>IM</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTA2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AHAPOQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCDC888</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Figure 2.1 Experimental and computational workflow

a. Brain tissues were homogenized in liquid nitrogen using mortar and pestle followed by nuclei permeabilization and library preparation using a commercial droplet based scRNA-seq procedure (10x Genomics). b. Cardiac ventricles were homogenized using gentleMACS (Miltenyi), since grinding was not sufficient to release cardiac nuclei. After nuclei permeabilization, library preparation was conducted using a commercial droplet based scRNA-seq procedure (10x Genomics). c. Assessment of barcode collision for a mixture of human and mouse nuclei. For ~2,000 nuclei the barcode collision rate was about 2.0 % which is within the specification provided by 10x Genomics for single cells. d. Computational analysis workflow flowchart.
Supplemental Figure 2.1 Experimental and computational workflow continued
Supplemental Figure 2.2 Clustering comparison using intronic and exonic reads vs. intronic reads

a) T-SNE plots of the adult mouse forebrain. Right shows clustering results using reads that align to both exons and introns; left shows clustering results using only intronic reads.

b) T-SNE plots of the human frontal cortex and hippocampus. Right shows clustering results using reads that align to both exons and introns; left shows clustering results using only intronic reads.

c) T-SNE plots of human heart ventricles. Right shows clustering results using reads that align to both exons and introns; left shows clustering results using only intronic reads.

d) Top row are histograms of the length in base pairs (bp) of the last exon for genes used when you use only exonic reads, only intronic reads, or both exonic and intronic reads; bottom row are histograms of the number of exons for genes used when you use only exonic reads, only intronic reads, or both exonic and intronic reads.
Supplemental Figure 2.3 snRNA-seq analysis of adult mouse forebrain recapitulates spatial neuronal heterogeneity and reveals neural stem cell like cell populations

a Heatmap visualizing row-based z-score of average UMI for marker genes of previously defined interneuron subtypes (left); heatmap visualization row-based z-score of average UMI for marker genes of CGE-derived neurogliaform cells (middle); Scaled UMI expression of PValb, Npy, Sst in sub-clustered T-SNE plot (right); b Violin plots of UMI for cell type specific genes; c Co-clustering of previously published NPC data with our Eomes+ population. Plot showing the percentage of cells of each cell type found in each cluster (top). The size of the circle is proportional to the percentage of cells; cell types in grey font are cell types identified in previously published data; cell types in black font are cell types identified in our data. The three bottom plots show a T-SNE visualization of the co-clustering and are colored by cell type or cluster ID.
Supplemental Figure 2.4 snRNA seq analysis of adult human frontal cortex and hippocampus illustrates glia cell heterogeneity

a Violin plots of UMI for cell type specific genes; b Spearman correlation to previously published data \(^{11}\) (see methods); c Heatmap showing average scaled UMI of genes shown to be markers of SCPN & CPN Ex Neurons \(^{12}\) (top left); Violinplot of UMI of genes specific to each excitatory neuron type (bottom left); Heatmap showing average scaled UMI of genes shown to be markers of SST+ & VIP+ Inhibitory Neurons \(^{11}\) (top right); Violinplot of UMI of genes specific to each inhibitory neuronal type (bottom right); d Boxplot of genes that are markers of astrocyte subtypes in \(^{27}\); e Venn diagram showing intersect between genes differentially expressed between bulk white and grey matter and genes differentially expressed between AC I and AC II.
Supplemental Figure 2.5 Intronic read based clustering shows increased specificity for cell clustering of single nuclei RNA-seq datasets in human heart ventricles

a T-SNE visualization of clustering using intronic reads vs all reads; expression (scaled UMI) of cardiomyocyte specific markers MYH6 & TTN are highlighted in each plot. Expression is calculated using only intronic reads (top), and using intronic & exonic reads (middle, bottom); b Violin plots of ventricular UMI data based on expression calculated from intronic reads alone. These genes are markers of T/NK; c Sub-clustering of ventricular cardiomyocyte cluster separating out the cells by tissue of origin; d Bar plot of GO/GSEA terms enriched in cardiomyocytes of left and right ventricles; e Boxplots of number of genes in each cell type from each sample (p<.05 marked by *, p<.005 marked by **, p<.0005 marked by ***, NS (not significant) (right); Boxplots of number of transcripts in each cell type from each sample (p<.05 marked by *, p<.005 marked by **, p<.0005 marked by ***, NS (not significant) (left); Sup 1d: Gene set enrichment analysis of genes differentially expressed between left and right cardiomyocytes
2.8 Methods
2.8.1 Data generation
1.1 Tissue Collection

Human dorsolateral prefrontal cortex and hippocampus tissue of a healthy 31 year old male were obtained from the National Institute of Child Health and Human Development (NICHD) Brain Bank for Developmental Disorders. Ethics approval was obtained from the University Health Network and The Hospital for Sick Children for use of the tissues. Cardiac ventricles were from a 3 year healthy male described before as part of the Epigenome Roadmap Consortium collection (The Roadmap Epigenomics Consortium, 2015; Biomaterial provided by Shin Lin, Stanford University). Adult mouse forebrain was harvested from 8-week old male C57BL/6NCrl mice (Charles River Laboratories) and flash frozen in liquid nitrogen separately.

1.2 Nuclei isolation
1.2.1 Brain tissue

Brain tissues were ground in liquid nitrogen using mortar and pestle. Next, 10-30 mg of ground tissues were incubated with 500 µl nuclei isolation buffer (2% BSA, 0.2% Triton-X cOmplete (Roche), 1 mM DTT, 0.2 U/µl RNAsin (Promega) in PBS) for 10 min at 4°C. After centrifugation with 500 x g for 5 min nuclei were resuspended in sort buffer (2 % BSA, 1 mM EDTA, 0.2 U/µl RNAsin (Promega) in PBS). Nuclei suspension was filtered before sorting.

1.2.2 Cardiac ventricle tissue

Nuclei from cardiac ventricles was isolated as described previously with minor modifications (cite). Frozen cardiac tissue was thawed on ice and submerged in 1 ml lysis buffer (5 mM CaCl2, 3 mM MgAc, 2 mM EDTA, 0.5 mM EGTA, 10 mM Tris-HCL,
pH 8). Next, ventricles were homogenized using a gentleMACS dissociator (Miltenyi) with gentleMACS M tubes (Miltenyi) using the protocol “protein_01”. Next, 1 ml of lysis buffer supplemented with 0.4 % triton-X was added and the combined cell suspension was filtered. The filter was washed with 1 ml lysis buffer including 0.2 % triton-X to increase nuclei yield. The flow-through was centrifuged (500 x g, 5min) and resuspended in sort buffer.

1.3 Nuclei sorting and single nuclei RNA-seq library preparation

Nuclei were sorted using a SH800 sorter (Sony) after staining with the DNA-dye DRAQ7 (3 µM, Cell Signalling Technologies). Sorted nuclei were pelleted (15 min, 1000 x g) and resuspended in PBS containing 1 % BSA and 0.2 U/µl RNAsin (Promega), at a concentration of 1000 nuclei/µl. Single nuclei RNA-seq libraries were constructed using the Chromium™ Single Cell 3’ v2 Library kit (10x Genomics) according to manufacturer descriptions. Reverse transcription and other amplification steps were carried out on a T100 thermal cycler (Bio-Rad). After reverse transcription, GEMs (Gel beads in emulsion) were lysed and cDNA was cleaned up with MyOne Silane Beads (Thermo Fisher Scientific). Next, single stranded cDNA was PCR-amplified for 14 cycles and purified using SPRIselect Reagent Kit (Beckman Coulter). Next, cDNA was enzymatically fragmented followed by double size selection with SPRIselect Reagent Kit (200-700bp, 0.6x and 0.8x, Beckman Coulter). Subsequently, adaptors were ligated and libraries were constructed by PCR. Another round of double size selection was performed using SPRIselect Reagent Kit (200-700bp, Beckman Coulter). Final libraries were quantified using Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific) and size distribution was measured using Tapestation (High Sensitivity D1000, Agilent). Average
fragment size of successful libraries was 500 bp. The libraries were loaded at a concentration of 13 pM and sequenced on a HiSeq 2500 sequencer (Illumina) with the following parameters (Read1 26 cycles; Index 1 8 cycles; Read 2 98 cycles).

2.8.2 Data Analysis

2.1 snRNA-seq alignment, clustering & marker gene identification

Samples were mapped using the STAR aligner to the gencode version 27 (GRCh38.p10) genome. In order to retain intronic reads, we used the mapping pipeline used in \textsuperscript{12} modifying it slightly to give mapping preference to exons, if these exons overlapped with intronic regions of other genes oriented in the same direction. For downstream analysis, to deconvolve cell types and merge samples, we used the Seurat package in R \textsuperscript{31,32}. The following analysis was performed on a gene by cell matrix of UMIs obtained from reads mapping to intronic regions alone. We first filtered out cells that had less than 200 expressed genes, and all genes that were expressed in ten or fewer cells. In order to filter out potential doublets, we used a filter on the ratio of UMI/genes in all cells. This ratio showed a bimodal distribution. Cells with a higher ratio of UMI/gene were filtered out. We scaled each gene umi count by 1e4, normalized and log transformed the expression values in each cell \textsuperscript{33}, and then scaled to z scores for each gene across all cells. We then used a sparse svd algorithm \textsuperscript{34} to decompose the scaled expression matrix, and then performed the the Barnes-hut implementation of TSNE. For cluster identification we used the FindClusters function in the Seurat package which first identifies the k nearest-neighbors to calculate a shared nearest neighbor graph, and then uses the original Louvain algorithm for modularity based community detection. K parameter for k nearest neighbors was chosen to be k that best optimized the following validity indices (Dunn Index, Davies-Bouldin Index, Silhouette
Coefficient). After cluster identification, we disregarded clusters with less than 10 cells, and mapped back gene expression onto cells, as if we had generated a gene by cell UMI matrix using intronic and exonic reads. We then used Wilcoxon-rank-sum test to identify genes that were significantly upregulated in the cells in one cluster vs. all other cells. Clusters were kept if their cells showed at least 2 fold upregulation of a gene, and merged with another cluster if they did not.

2.11 snRNA-seq alignment and quantification

First we aligned our sequencing data using only reads that mapped to introns (reads representative of transcripts found within the nucleus), and clustered the resulting gene expression matrix to identify cell types. Even though intronic reads alone are sufficient to accurately separate cell types, it was difficult to confidently annotate human and mouse brain cell types without including transcription factors, or genes that lack introns in our analysis. Therefore, we projected back the gene expression values that would be obtained from aligning and clustering all reads onto our cells in order to perform downstream analysis.

2.2 Comparison of single cell data to previously published data

We first created a list of cell-type specific marker genes identified from our differential gene expression analysis on each of the human and mouse brain clusters. For these genes we calculated the pairwise spearman correlations between cell types identified in our data set and in the data set from 6. The cell-type’s averaged scaled UMI counts were used to calculate the correlation.

2.3 Pseudotime Analysis
Pseudotime analysis was performed on cells from the two OPC and four ODG clusters using the Monocle package in R\textsuperscript{35-37}.

2.4 Gene Ontology Analysis

We used the cited publicly available databases\textsuperscript{38-40} for GO Analysis and Gene Set Enrichment Analysis.

2.9 Acknowledgments

Chapter 2, in full, is in preparation to be submitted as “Using intronic reads to deconvolve frozen tissue samples from human heart, and human and mouse brain by single nucleus RNA-seq”. Anugraha Raman, Sebastian Preissl, Zhang Cheng, Song Chen, Zhen Ye, Samantha Kuan, Kun Zhang, Cathy Barr, Bing Ren: The dissertation author is the first author of this manuscript in preparation.
2.10 References


CHAPTER 3: Inter and Intra tumor transcriptomic heterogeneity in Glioblastoma

3.1 Abstract

Glioblastoma is the most common form of primary brain tumor in adults and remains one of the deadliest of human cancers. Intra-tumoral heterogeneity in glioblastomas remains a major challenge in therapeutic development. While single cell based RNA profiling has yielded insights into this heterogeneity, the resolution of these techniques is limited and the pertinence of RNA-based profiling to epigenetic landscape remains unclear. To better address these issues, we have developed and applied single nuclei ATAC-Seq (snATAC-Seq) methods to clinical glioblastoma specimens secured from distinct regions of a tumor and compared these profiles to those derived using single nuclei RNA-Seq (snRNA-Seq). We found that snATAC-seq could identify CNVs at the single cell level with high resolution (20KB), and these CNVs define distinct tumor sub-populations in different geographic locations of the tumor. Using aberrant CNVs as a marker for cells derived from the glioblastoma, we found significant representation of cells with aberrant CNVs that harbor transcriptomic profiles suggestive of normal oligodendrocyte progenitor cell and astrocytes, suggesting potential for these cells to be tumor cells of origin. Our results indicate a combination of snRNA-seq and snATAC-seq as a valuable molecular platform for characterization of glioblastoma heterogeneity.
3.2 Introduction

Gliomas refer to primary brain tumors that arise from neural stem cells or progenitor cells, and account for 28% of all brain tumors and 80% of all malignant tumors\(^1\). These tumors are classified in terms of their severity (grade) and histology (astrocytic, oligodendrogial, ependymal). Of these gliomas, the most severe (grade IV), which are found predominantly in older individuals are classified as glioblastomas, and account for the majority of gliomas\(^2\). Individuals with Glioblastoma (GBM) have an extremely poor prognosis. Median survival after diagnosis and with treatment is in the range of 12 to 15 months\(^1\). Not only can GBM occur in many different regions of the brain, but different tumors can contain different molecular signatures.

The cancer genome atlas (TCGA) made one of the largest efforts to characterize the molecular heterogeneity in GBM\(^3,4\). Using 206 glioblastoma samples, TCGA’s pilot study uncovered three pathways commonly mutated in GBM (receptor tyrosine kinase signaling, p53 signaling, and RB signaling)\(^3\). Shortly after, using the same GBM data, another study revealed four molecular subtypes defined by 840 genes, more specifically by mutations or copy number variants found in \textit{EGFR}, \textit{NF1}, \textit{PDGFRA}, \textit{IDH1}\(^4\). These molecular classifications helped describe inter-tumoral heterogeneity; however, due to the failure of many targeted therapies, which either inhibit \textit{EGFR} or target the commonly mutated pathways in GBM, it is increasingly evident that we must deepen our understanding of intra-tumoral heterogeneity\(^5\).

Single cell technologies are well suited to characterizing intra-tumoral heterogeneity, and have been used to study glioblastoma since 2014. In one of the first studies to perform scRNA-seq on GBM tumors, Patel et al looked at \~400 cells across 5 different patients and showed that individual tumors, characterized as expressing the
molecular signature of only one of the four TCGA subtypes, actually were comprised of cells that expressed signatures of more than one TCGA subtype. Furthermore, an increase in TCGA subtype heterogeneity was associated with decreased survival\(^6\).

Francis et al used a different single cell technology (scDNA-sequencing) to identify and characterize clonal sub populations with different kinds of \(\text{EGFR}\) mutations\(^7\). Several years later, another scRNA-seq study profiled \(\sim\)3500 cells in 4 patients with the goal of characterizing inter and intra patient heterogeneity between neoplastic tumor cells at the periphery of a tumor versus cells at the glioblastoma tumor core. Their study surprisingly revealed some commonalities between genes signatures of invading neoplastic cells\(^8\). More recently, a study attempting to investigate the glioblastoma cell of origin used single cell sequencing to show that cells carrying mutations in the tumor free SVZ region of a patient’s brain can evolve and turn into the glioblastoma tumor\(^9\).

While scRNA and scDNA sequencing technologies, which profile transcriptomic and genomic heterogeneity, have been used to study GBM, nobody has profiled another potentially therapeutically rewarding component of intra-tumoral heterogeneity—single cell epigenomic heterogeneity. ATAC-seq, an epigenomic assay for profiling regions of open chromatin can be used to reveal active regulatory elements genomewide\(^10\). Recently, a large effort to characterize open chromatin profiles in TCGA samples was made using this technology.\(^{11}\) This study described tissue and cancer specific drivers, and linked open chromatin peaks (intergenic regions of high chromatin accessibility) to genes that may be regulated by the accessibility of these peaks.

Integrating scRNA-seq and scATAC-seq profiles of matched glioblastoma samples would allow for thorough classification of this disease since scRNA-seq profiles
would serve as a functional readout of the scATAC-seq profiles. However, in order to achieve this goal, several computational challenges exist. The biggest challenge stems from the cellular composition of glioblastoma tumor sections. In any given frozen tumor section, some fraction of the cells are actually cell types observed in the normal human brain, while others are the tumor cells, which often harbor mutations. Previous studies have used the presence of profilable mutations to computationally separate out these cell types by either using single-cell RNA-seq to identify chromosomal arm level amplifications or deletions, or by identifying point-mutations using reads that span the mutated region\(^\text{6,12}\). Experimentally cell-type specific markers have been used to differentiate non-neoplastic cells from other cells\(^\text{8}\). Other major challenges are related to: combining single cell data across multiple patients, and teasing apart potential batch effects from true genetic variation between patients, and linking regions of high accessibility (ATAC-seq peaks) with genes that these peaks could be associated with\(^\text{13,14}\).

In this study we performed snATAC-seq and snRNA-seq (single nuclear RNA-sequencing) on flash frozen tissue from five different glioblastoma patients with the goal of characterizing transcriptional and chromatin accessibility profiles in subtypes of glioblastoma tumor cells. Two patients provided sectional information representing proximity to the tumor core. Using the generated single cell chromatin accessibility profiles, we first identified CNVs distinct to tumor cell populations and used these regions to identify tumor cells in our snRNA-seq profiles. We then characterized transcriptional differences between sub-populations of tumor cells, and identified two
populations of astrocyte (ASC) and oligo-progenitor (OPC) like cells, which could be potential “cells of origin.”

**Results**

Single nuclei were isolated from flash frozen tumor sections obtained from five different GBM patients (GBM1, GBM4, GBM9, GBM11, GBM12). Five sections representative of distance away from the tumor core, were resected from GBM1, and four sections, also representative of this distance, were resected from GBM4. These nuclei were used to perform single nuclear RNA-seq (snRNA-seq) and single nuclear ATAC-seq (snATAC-seq) (Fig 1a,b). We expected these tumor sections to contain tumor cells, as well as glial and neuronal cells typically found in the human brain. To computationally distinguish between these cell types, we co-clustered cells from each patient , for each assay, with normal cells from the human brain (frontal cortex and hippocampus) that were analyzed in chapter 2 (Fig 1c).

Using the UMAP\textsuperscript{15} algorithm to visualize our clustering results, we identified 25 clusters from our snATAC-seq profiles, and 18 clusters from our snRNA-seq profiles (Fig 1c). These clusters represented normal glial and neuronal cells (astrocytes, microglia, oligodendrocytes, oligodendrocyte progenitor cells, inhibitory and excitatory neurons), as well as tumor sub-clusters. We found there to be 16 tumor sub-clusters in snATAC-seq and 8 tumor clusters in snRNA-seq (T1-T8). In order to gain confidence in our snRNA-seq clustering results, we visualized and annotated cluster specific genes (Sup Fig 1a), and plotted signatures of normal cell types onto our single nuclear RNA-seq profiles (Sup Fig 1b). Enrichment of top differentially expressed genes between clusters corresponded well with our annotations; for example, microglia were enriched
for immune systems processes, and astrocytes and neurons were enriched for genes known to be upregulated in astrocytes and neurons (cahoy astrocytic; cahoy neuronal). Interestingly we also identified two populations (ASC-like and OPC-like) that were enriched for cell cycle and gliogenesis genes. We also tried several approaches to merging and clustering snRNA-seq data across all patient samples (Sup Fig 1c, methods) to see if we could achieve more biologically meaningful cluster separation. When co-clustering each patient sample with cells from the normal human brain, we observed a considerable number of cells that were “non-tumor” cell types (Sup Fig 1d); therefore, we evaluated the validity of different clustering methods used to merge patient samples, by looking for clusters that contained cells from the normal human brain sample, as well as from the tumor sections (see methods). The only approach that resulted in clusters containing cells from our GBM patients, as well as from our normal human brain sample, in similar proportions to what was observed in individual patient clustering, was an approach that used canonical correlation analysis. It is still possible that this algorithm tried to over-fit the similarity between cells across multiple patients, but we were content with the cluster specific transcriptional signatures.

From this analysis, we asked if we could identify clusters that were specific to tumor cells within the tumor sections. By plotting the percentage of cells from each patient in each cluster (Fig 1d), we observed that the majority of tumor cell clusters from snATAC-seq were patient specific; however, the majority of tumor cell clusters from snRNA-seq were mixed across patients. To better understand if the patient specific clustering results from snATAC-seq profiles were truly patient specific or the result of
inter-sample variation, we also characterized differences in chromatin accessibility profiles between tumor sub-clusters.

Each tumor section sampled contained both cells that are normally present in the human brain, as well as tumor cells. To confidently differentiate tumor from normal cells in our tumor sections, we co-clustered our tumor sections with normal cells, and plotted signatures of known cell types; however, we asked if we could also identify aberrant CNVs at the single cell level. scRNA-seq has been used to previously identify CNVs; however, the resolution of this method is limited to the number of genes captured\(^6\). Since snRNA-seq is much sparser than scRNA-seq data, we asked if we could use chromatin accessibility data that profiles DNA instead of RNA, to identify CNVs at high resolution. We were able to identify CNVs at high resolution (at least 30kb) in these patients (Fig 2a, 2b, see methods). For instance, in GBM1, consistent with the patient’s medical report and whole-genome sequencing (WGS) data, we were able to identify amplifications of regions containing *CCND1*, *EGFR*, and *LINC00676*, to name a few (Fig 2b, Supplementary Table 1). Using the snATAC-seq data we identified 26 regions that were amplified across in at least one of our five patient samples. Some regions, such as those containing *EGFR* and *PDGFRA* were amplified in multiple patients (Fig 2c). To understand if these CNVs were transcribed at higher levels, we looked at the cumulative expression of genes in these regions in our snRNA-seq clusters (Fig 2c). While the majority of amplified regions contained genes that were highly expressed in tumor clusters versus normal clusters, there were some exceptions. We see that the amplified region containing *EGFR* is strongly expressed in two clusters containing cells primarily from GBM12, and is moderately expressed in two clusters containing cells...
primarily from GBM4 and GBM1. All three of these patients show amplification of \textit{EGFR} in their medical reports. We also observe high expression of \textit{CCND1} in clusters predominantly comprised of cells from patient 1 (Fig 2b). However, \textit{FGF19}, which is co-amplified with \textit{CCND1} in the scATAC-seq data is not highly expressed in the snRNA-seq data (Sup Fig 2a). To better understand how frequently the CNVs we observed were represented in other patients, we also looked across GBM patients from the cancer genome atlas (TCGA) to characterize the frequency of mutations found in the genes contained in our amplified regions (Sup Fig 2b).

After establishing that snATAC-seq profiles could be used to identify CNVs, we asked for the patients with sectional data (GBM1, GBM4), if the amplifications we observed were specific to a section of the tumor. GBM4 was reported to carry both \textit{EGFR} and \textit{PDGFRA} amplifications, and from the chromatin accessibility profile, it became clear that these two amplifications were mutually exclusive, and largely section specific (Fig 3a). To understand if these amplifications resulted in high expression of these genes, we first computationally isolated all the tumor cells from all patients, sub-clustered these cells and then plotted expression of these genes in the tumor cells of GBM4 (Fig 3b, Sup Fig 3a). The mutual exclusivity of these two amplifications was largely represented in our snRNA-seq data, and has also been previously reported\textsuperscript{6} (Fig 3a,b). In GBM 1, the chromatin accessibility profiles showed amplification of \textit{CCND1} in all tumor cells; however, there also appeared to be clonal subpopulations that were section specific (Fig 3a). When comparing our accessibility profiles to gene expression, we found that \textit{CCND1} was highly expressed in all tumor cells of GBM1; however, transcriptional (snRNA-seq) sub clustering of tumor cells in GBM1 was not
driven by CNV differences present in the four clones observed in the snATAC-seq clustering (Fig 3c, Sup 3a).

Since clonal CNVs captured by open chromatin signatures were not driving the transcriptional heterogeneity of tumor cells, we asked if pre-defined molecular signatures could describe the observed transcriptional heterogeneity. One such molecular signature is the TCGA derived GBM subtypes (neural, mesenchymal, pro-neural and classical) identified from bulk RNA-sequencing\textsuperscript{17}. For each tumor cluster we calculated gene expression signals for all four TCGA defined molecular subtypes, using the snRNA-seq data, and clustered these results. The majority of cells carried a proneural or classical signature, one tumor sub-cluster, T5, which is also specific to GBM4, carried a strong mesenchymal signature, and T1, which is specific to GBM12, carried a strong neural signature. Overall, however, all tumor clusters were a mixture of these four molecular subtypes (Fig 4a). To understand if the snRNA-seq tumor cells contained enrichment for any known pathways or gene-sets, we took all the tumor cells (T1-T8), sub-clustered them, and performed gene set enrichment analysis (Fig 4b, Sup 3a). Some pathways found to be enriched across patients included (MYC targets, G2M checkpoint, and mTORC1 signaling).

Tumors are comprised of cells that initiate them (cell of origin) and cells that help them propagate (cancer stem cells). Glioma cells of origin are believed to be cells that were originally normal glial cells, which acquired mutations that either prevented differentiation or promoted self-renewal\textsuperscript{18,19}. In order to better understand glioma development, studies have looked towards normal gliogenesis to draw parallels; specifically, they have looked at understanding parallels between the gliogenic switch
and glioma-genesis, since glioma stem cells are more similar to glial precursor cells than neural stem cells\textsuperscript{19}. The gliogenic switch refers to the point in neural stem cell (NSC) development, during which expression of several transcription factors pushes the cell towards turning into an \textit{(NFIA, SOX9)} astrocyte precursor (APC) or \textit{(OLIG2, SOX10)} oligodendrocyte pre-cursor cell (OPC)\textsuperscript{19}. Therefore, in order to understand which glial precursor populations were represented in each of our individual tumor sections, we characterized expression of the transcription factors involved in the gliogenic switch (Fig 4c). GBM1 and 9 carried signatures of both OPC and APCs, and GBM 4, 11 and 12 carried a signature of cells on an astrocytic trajectory.

In our snRNA-seq clustering results of all patient sections, we observed two populations consisting of astrocyte (ASC) like and oligodendrocyte progenitor cell (OPC) like cells (Fig 1d, Fig 4a). We observed that GBM4 and 11, which carried a strong signal for APCs, also represented the majority of cells found in the ASC-like cluster. Furthermore, GBM1 and GBM9, which carried a strong OPC signal, represented the majority of cells found in the OPC-like cluster. In order to understand if these OPC and ASC like cells could be potential cells of origin in these patients, we performed pseudotime analysis using the ASC-like, OPC-like, ASC, oligodendrocyte (ODG), OPC and tumor cells in GBM1,4 and 9, and the (ASC-like, ASC and tumor cells) in GBM 11,12 (Fig 4d, Sup Fig 4). In pseudotime trajectory plots, any end of the trajectory can be interpreted as the starting point “time 0.” However, since we know that normal ODG cells must be older than normal OPC cells in our tumor sections, we can use this to interpret these trajectories. Using this interpretation, it appeared that the ASC-like and OPC-like populations could have given rise to the observed tumor
subpopulations. In GBM 1, and 9 it was clear that the OPC-like cells could have given rise to the predominant tumor subtypes in those patients (T3,T4,T6,T8; T2). In GBM 11 we observed that the ASC-like population likely gave rise to the predominant tumor subpopulation (T4); however, in GBM 12, the majority of T7 cells appeared to be the same age as the ASC-like cells. Finally, in GBM 4 the ASC or OPC-like cells could have both resulted in the observed tumor subpopulations.

While GBM 4 didn’t show strong markers of normal gliogenesis towards the OPC and ODG lineages, it did contain a significant portion of cells that were classified to be OPC-like (22%). Since it is believed that changes to the typical gliogenesis pathway lead to glioma formation, and that normal cell types, which are more susceptible to mutations, turn into tumor cells of origin, we asked what the differences were between OPC-like cells in GBM 4 and normal OPCs\textsuperscript{18,19}. We also looked at these differences in GBM 1,4 and 9, as well as the differences between ASC-like and normal astrocytes in GBM 1,4,11,12. These transcriptional differences included increased expression of previously identified CNV genes (GBM1: \textit{CCND1, ORAOV1, ETV1} GBM4: \textit{EGFR} GBM9: \textit{PDGFRA}, GBM12: \textit{EGFR}). (Sup Table 2).

**Discussion:**

The true transcriptional signature of a tumor cell varies at the inter and intra tumoral level; therefore, any method that studies the heterogeneity of these signatures must confidently distinguish tumor from other cells found in each patient. Identification of known genetic abnormalities found in the tumor, such as CNVs, deletions or point-mutations, at the single cell level, provides confidence in cell characterization. In our study, snATAC-seq allows for the identification of CNVs at high resolution. The
expression of genes found in these CNVs can further be investigated, at the single cell level, by using our snRNA-seq profiles. In GBM1, not only were we able to identify clonal tumor sub-clusters, but we were also able to investigate the transcriptional significance of the co-amplification of a locus containing CCND1 and FGF19. FGF19 is amplified in several other cancers including head and neck squamous cell carcinoma and breast $^{20,21}$. In breast cancer, this gene has been commonly co-amplified with CCND1, and associated with poor patient prognosis. Furthermore, expression of FGFR4, FGF19’s co-expressed receptor, has been correlated with resistance to the EGFR inhibitor erlotinib in glioblastoma cell lines$^{22}$. Taken together, these observations suggest that it would be important for patient treatment to know if a patient carried both an EGFR and FGF19 amplification. In GBM 1, it appears that the amplification of FGF19 does not result in high levels of gene expression, and is likely amplified due to its proximity to the CCND1 locus.

The presence of sectional data from GBM1 and 4 allowed us to further investigate our high-resolution CNV calls, and ask if we observed sectional differences associated with section specific CNVs. Interestingly, in GBM 4, we find mutually exclusive amplifications of PDGFRA and EGFR. GBM is marked by mutations in RTKs, and one of the most commonly mutated pathways in GBM is the PI3K pathway$^3$. PDGFRA and EGFR, two genes found to be amplified in GBM4 are both part of this core pathway that regulates cellular proliferation$^{23}$. While these amplifications may appear to be distinct, they both lead to constitutive activation of the PI3K pathway. In patients like GBM4, patients with multiple RTK mutations, anti-PDGFR and anti-EGFR
therapies have either failed or done poorly, suggesting that these patients would be prime candidates for combination therapies aimed at multiple RTKs. While we were unable to identify cancer stem cells (CSC) by looking for known CSC markers, we were able to identify two cell types (ASC-like and OPC-like), represented in all patients, which could be potential cells of origin for the tumor cells analyzed in our tumor sections. Compared to ASCs and OPCs found in the normal human brain, these cells exhibited higher expression of genes amplified in GBM 1, 4, 9 and 12. Both cancer stem cells and tumor cells of origin contribute to tumor progression, and are important to identify. OPCs could serve as a cell of origin for glioma, due to their demonstrated plasticity—their ability to differentiate into astrocytes and oligodendrocytes based on external signaling. A recent study, however, showed that NSCs found in the SVZ are likely tumor cells of origin for glioblastomas. The ASC-like and OPC-like cell types that we identified as potential cells of origin could have differentiated from NSCs found in regions of the brain not sampled in this study.

We have combined snATAC-seq and snRNA-seq to investigate inter and intra-tumor heterogeneity in five patients. Our snATAC-seq analysis has outlined a method for identifying CNVs at high resolution, as well as identifying clonal tumor subpopulations based on CNV signatures. By combining this approach with snRNA-seq to better understand which amplified genes are actually over expressed in sub clonal populations, we provide an analysis framework that has implications for therapy. Furthermore, we provide insight into the transcriptional signatures, shared across five patients, of potential cells of origin in these tumor sections. Future studies designed to give us a better understanding of these cells of origin and their open chromatin
landscape, as well as studies that investigate the open chromatin and transcriptomic landscapes of tumor sub clonal populations before and after treatment will be important to advance current therapies.
3.5 Experimental Procedures
3.5.1 Data generation

1.1 Tissue Collection

Post-surgery flash frozen glioblastoma samples were obtained from Dr. Clark Chen. The normal samples used in this study (Human dorsolateral prefrontal cortex and hippocampus tissue of a healthy 31 year old male) were obtained from the National Institute of Child Health and Human Development (NICHD) Brain Bank for Developmental Disorders. Ethics approval was obtained from the University Health Network and The Hospital for Sick Children for use of the tissues.

1.2 Nuclei isolation

Brain tissues were ground in liquid nitrogen using mortar and pestle. Next, 10-30 mg of ground tissues were incubated with 500 µl nuclei isolation buffer (2% BSA, 0.2% Triton-X cOmplete (Roche), 1 mM DTT, 0.2 U/µl RNAsin (Promega) in PBS) for 10 min at 4°C. After centrifugation with 500 x g for 5 min nuclei were resuspended in sort buffer (2 % BSA, 1 mM EDTA, 0.2 U/µl RNAsin (Promega) in PBS). Nuclei suspension was filtered before sorting.

1.3 Nuclei sorting and single nuclei RNA-seq library preparation

Nuclei were sorted using a SH800 sorter (Sony) after staining with the DNA-dye DRAQ7 (3 µM, Cell Signalling Technologies). Sorted nuclei were pelleted (15 min, 1000 x g) and resuspended in PBS containing 1 % BSA and 0.2 U/µl RNAsin (Promega), at a concentration of 1000 nuclei/µl. Single nuclei RNA-seq libraries were constructed using the Chromium™ Single Cell 3’ v2 Library kit (10x Genomics) according to manufacturer descriptions. Reverse transcription and other amplification steps were carried out on a T100 thermal cycler (Bio-Rad). After reverse transcription, GEMs (Gel beads in
emulsion) were lysed and cDNA was cleaned up with MyOne Silane Beads (Thermo Fisher Scientific). Next, single stranded cDNA was PCR-amplified for 14 cycles and purified using SPRIselct Reagent Kit (Beckman Coulter). Next, cDNA was enzymatically fragmented followed by double size selection with SPRIselct Reagent Kit (200-700bp, 0.6x and 0.8x, Beckman Coulter). Subsequently, adaptors were ligated and libraries were constructed by PCR. Another round of double size selection was performed using SPRIselct Reagent Kit (200-700bp, Beckman Coulter). Final libraries were quantified using Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific) and size distribution was measured using Tapestation (High Sensitivity D1000, Agilent). Average fragment size of successful libraries was 500 bp. The libraries were loaded at a concentration of 13 pM and sequenced on a Hiseq 2500 sequencer (Illumina) with the following parameters (Read1 26 cycles; Index 1 8 cycles; Read 2 98 cycles).

Single nuclear ATAC-seq was performed using the protocol outlined in Preissl et al 2018\(^{38}\).

### 3.5.2 Data Analysis

#### 2.1 snRNA-seq alignment, clustering & marker gene identification

Samples were mapped using the STAR aligner to the gencode version 27 (GRCh38.p10) genome. In order to retain intronic reads, we used the mapping pipeline used in\(^7\), modifying it slightly to give mapping preference to exons, if these exons overlapped with intronic regions of other genes oriented in the same direction. For downstream analysis, to deconvolve cell types and merge samples, we used the Seurat package in R\(^{33,34}\). The following analysis was performed on a gene by cell matrix of UMIs obtained from reads mapping to intronic regions alone. We first filtered out cells
that had less than 200 expressed genes, and all genes that were expressed in ten or fewer cells. In order to filter out potential doublets, we used a filter on the ratio of UMI/genes in all cells. This ratio showed a bimodal distribution. Cells with a higher ratio of UMI/gene were filtered out. We scaled each gene umi count by 1e4, normalized and log transformed the expression values in each cell\textsuperscript{16}, and then scaled to z scores for each gene across all cells. We then used a sparse svd algorithm\textsuperscript{35} to decompose the scaled expression matrix, and then performed the Barnes-hut implementation of TSNE. For cluster identification we used the FindClusters function in the Seurat package which first identifies the k nearest-neighbors to calculate a shared nearest neighbor graph, and then uses the original Louvain algorithm for modularity based community detection. K parameter for k nearest neighbors was chosen to be k that best optimized the following validity indices (Dunn Index, Davies-Bouldin Index, Silhouette Coefficient). After cluster identification, we disregarded clusters with less than 10 cells, and mapped back gene expression onto cells, as if we had generated a gene by cell UMI matrix using intronic and exonic reads. We then used Wilcoxon-rank-sum test to identify genes that were significantly upregulated in the cells in one cluster vs. all other cells. Clusters were kept if their cells showed at least 2 fold up-regulation of a gene, and merged with another cluster if they did not.

2.2 snATAC-seq alignement, clustering & marker gene identification

Samples were mapped to the hg38 genome, and downstream analysis was performed using the SnapATAC pipeline\textsuperscript{39}.

2.3 Comparing different approaches to snRNA-seq clustering across patients

The snRNA-seq clustering results in Fig 1 are the clustering results of Seurat’s MultiCCA algorithm, designed specifically for integrating single cell data across different
conditions, technologies and species. For Sup Fig 1c: We also used harmony, an algorithm designed to account for variation across samples/conditions in scRNA-seq data. We applied this algorithm to the sparse singular value decomposition of the scaled UMI count matrix (merged across all patients), and to the principal components of the scaled UMI count matrix (merged across all patients). We also applied singular value decomposition to the scaled UMI matrix (merged across all patients) and used those components to perform graph-based clustering.

2.4 Pseudotime Analysis

Pseudotime analysis on cells from each patient was performed using the Monocle package in R.

2.5 Plotting gene expression of normal cell type signatures

Using the set of top differentially identified genes from single cell analysis of the human brain in chapter 2, we defined genes signatures for astrocytes, oligodendrocyte, oligodendrocyte progenitor cells, microglia and neurons. We took the average expression of these genes in each cell, calculated a z-score based on all cells in the sample, and plotted this value onto our snRNA-seq clustering results.

2.6 Gene Set Enrichment Analysis

We used the cited publicly available databases for GO Analysis and Gene Set Enrichment Analysis.

2.7 CNV Identification form snATAC-seq data

In order to detect genomic amplifications from snATAC-Seq, we first got the aggregate counts of all cells within each cluster for 10kb bins across the entire genome.
In order to identify genomic amplifications, we generated a background based on 5000 random 10kb bins that overlap with open chromatin peaks in that cluster. We then calculated a p-value for each 10kb bin based on the number of 10kb windows in the background greater than the observed value and divided by 5000. We first used a cut-off of 0.01 to select the 10kb bins enriched ATAC-Seq signal and merged the bins. We further filtered any regions < 30kb to remove highly accessible promoter regions to determine the regions of genomic amplifications. These were then validated against the amplifications called using bulk WGS data from the tumor sample.
3.6 Figures

Figure 3.1: snRNA-seq and snATAC-seq profiles of GBM Patients
A) Schematic outlining the protocols used to generate each sample B) Statistical metrics about each sample (median number of UMI from snRNA-seq; median number of reads from snATAC-seq) C) UMAP visualization of snATAC-seq data (left) and snRNA-seq data (right) co-clustered with unmatched normal brain tissue from the human frontal cortex and hippocampus; Top figures are colored grey for normal cell types and colored for tumor clusters; Bottom clusters are colored by patient and section of origin D) Bar-plot indicating relative percentage of cells in a given cluster from each patient (Patient1, Patient4, Patient 9, Patient 11, Patient 12, Normal Sample).
Figure 3.2: CNV Identification using snATAC-seq profiles
A) CNV identification workflow B) scATAC-seq profiles in normal and tumor cells in patient 1 at three regions identified to contain different CNVs; Bottom track is a whole genome sequencing track from this patient C) Heatmaps displaying column z-scores of average chromatin open-ness (left) and average gene expression (right) at 26 regions identified to carry CNVs across patients. Rows represent cluster identity from original clustering of all patients merged with normal cells; columns represent the 26 amplified regions
A) CNV Identification Workflow

- Clustering of snATAC-Seq data
- Select cells from each cluster with the read count > 90th quantile
- Get counts for each cell for all 10kb bins in the genome
- Additional filtering:
  - Amplification in 10% of cells
  - Merge 10kb bins by 500kb and remove any regions <20kb

B) CNVs detected by snATAC-seq

C) snATAC-seq

Amplified regions (Normalized by length of region)

C) snRNA-seq

Amplified regions (Expression normalized by number of genes in region)
**Figure 3.3: snRNA-seq and snATAC-seq profiles of GBM Patients**

A) Clustering results of clustering snATAC profiles from GBM4 (patient 4) and GBM1 (patient 1) (left); clustering results shaded in by proximity to tumor core (darker colors are closer to tumor core) (middle); chromatin accessibility profiles of the promoters of amplified genes (right) B) Expression of amplified genes in GBM4; region that is boxed contains cells from patient 4 C) snRNA-seq profiles of tumor cells only from all patients, highlighting GBM1; cells are colored by cluster (left) and by clones described in snATAC-seq data (middle); clustering results on right display one GBM1 clone at a time
Figure 3.4 Inter and intra-tumoral transcriptional heterogeneity

A) Heatmap z-scored by column showing TCGA subtypes of tumor cells based on cumulative gene expression of TCGA molecular subtypes; columns are colored by initial clustering (left); UMAP plot of patient clustering colored by TCGA subtype (middle; UMAP plot of initial clustering results (right));

B) Sub-clustering of tumor cells shows enrichment of pathways involved in promoting Glioblastoma

C) Signatures of normal gliogenesis (snRNA-seq)

D) snRNA-seq pseudotime trajectories of cells in individual patients

Figure 3.4: Inter and intra-tumoral transcriptional heterogeneity
A) Heatmap z-scored by column showing TCGA subtypes of tumor cells based on cumulative gene expression of TCGA molecular subtypes; columns are colored by initial clustering (left); UMAP plot of patient clustering colored by TCGA subtype (middle; UMAP plot of initial clustering results (right)); B) Clustering of tumor cells in all patients annotated with gene set/pathway enrichment C) Expression of markers of typical gliogenesis in the human brain that lead to an astrocytic trajectory (top) and oligodendrocyte trajectory (bottom) D) Pseudotime trajectories of tumor cells, ASC, ASC-like, ODG, OPC, and OPC-like cells in patients GBM1, GBM4, GBM9; T1-T8 represent tumor subtypes
3.7 Supplemental Figures

**Supplemental Figure 3.1 snRNA-seq GBM patient clustering**
A) Heatmap showing cluster specific gene expression and enrichment of select go terms B) Gene expression signatures of normal glial and neuronal cell types in the human brain plotted onto snRNA-seq clustering results C) Clustering results of different techniques to merge snRNA-seq data across all patients D) snRNA-seq clustering results of each individual patient with normal frontal cortex & hippocampus samples; red represents patient cells and grey represents normal brain cells
Supplemental Figure 3.1 snRNA-seq GBM patient clustering continued
Supplemental Figure 3.2 Characterization of snATAC-seq identified CNVs

A) FGF19 expression plotted on snRNA-seq clustering results. Scale is in unique molecular identifiers (UMI)

B) Chart detailing presence of an amplification (red) or deletion (blue) at EGFR and other genes found to be amplified from snATAC-seq data
Supplemental Figure 3.3 snRNA-seq clustering of tumor cells from all patients

A) snRNA-seq Clustering of Tumor Cells

Supplemental Figure 3.3 Inter and intra-tumoral transcriptional heterogeneity

A) snRNA-seq clustering of only tumor cells identified across all patients; clusters are colored in by sample of origin (left); same clustering results as left, but clusters are colored in by cluster ID (right)
Supplemental Figure 3.4 Pseudotime Analysis of ASC-Like cells in GBM11 &12

A) Pseudotime trajectory profiles of ASC, ASC-Like and tumor cells in GBM11 (left) and GBM12 (right).
3.8 Acknowledgements

Chapter 3 is a manuscript in preparation: *Integrative single cell characterization of tumor heterogeneity in flash frozen GBM tumor sections* Ramya Raviram*, Anugraha M Raman*, Sebastian Preissl, Clark Chen, Bing Ren. The dissertation author will be the co first author of this paper.
3.9 References


Appendix: Hydroxymethylation Landscapes in Human Embryonic Cell Lineages and Adult Tissues

This appendix contains an exploratory correlative analysis on hydroxymethylation, which was the focus of my thesis prior to single cell transcriptomic analysis.

A.1 Summary

Hydroxymethylcytosine is an intermediate in the DNA de-methylation pathway; however its role in gene regulation has not been clearly elucidated. In order to better understand this modification, its genomic distribution needs to be characterized, and until several years ago, a method to quantify this mark at base-pair resolution did not exist. By profiling this mark in early embryonic cells and adult tissues, we characterize its genomic patterns at various genomic features, thus revealing characteristics of 5hmC that can be investigated in future studies. In particular we focus on enhancers and intergenic regions: We identify regions genome-wide containing elevated levels of 5hmC, which are enriched at exonic versus promoter regions, and at enhancers, and we show that changes in 5hmC and mC at enhancers are predictive of differentiation-induced changes in enhancer state.

A.2 Introduction

Hydroxymethylation is an epigenetic mark, whose role in gene regulation has not been clearly elucidated. While previous studies have shown that 5hmc plays an important role in embryonic development, the nature of this role remains to be understood. TET proteins catalyze the oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), leading to an unmethylated cytosine through to a series of less stable intermediaries (5fC or 5caC)\(^1\). 5hmC is often thought to play a passive role as a bystander in gene regulation, existing primarily as an oxidation product of 5mC.
and as an intermediate in the 5mC demethylation pathway $^{2,3}$ In contrast, some evidence suggests that 5hmC may play a more active role by directly influencing gene expression $^{4-6}$.

In order to better understand this modification, its baseline genomic distribution needs to be characterized. Until several years ago, a method to quantify 5hmC at base pair resolution did not exist $^7$. Previous methods were either non-specific, of much lower resolution, or jointly reported methylation and hydroxymethylation levels. By using TAB-Seq, Yu et al. created base resolution maps for 5hmC in human embryonic stem cells (hESCs). A useful reference, this map also showed that 5hmC and 5mC tended to be enriched at enhancers, and their levels are inversely correlated near transcription factor binding sites. The dynamic nature of methylation and hydroxymethylation observed at these cis-regulatory elements suggested that 5hmC could be regulating gene expression by acting on cis-regulatory regions. Several other studies have also quantified 5hmC, with the goal of elucidating its functional role. Collectively, these studies have shown that this mark is found in a variety of species as well as tissue types, and that it is enriched at enhancers and genes $^8$. This suggests a functional role for this mark in gene regulation $^{8,9}$.

In this study we characterize 5hmC at base-pair resolution in 5 embryonic cell types and 5 adult tissues to provide a reference for future studies. The 5 cell types are part of a human embryonic stem cell (H1) and H1-derived system thoroughly characterized by Xie et al $^{10}$. Each of these derived lineages represents key stages in early development, and has been profiled for various other epigenetic marks including (H3K27ac, H3K4me1, and 5mC). The 5 adult tissues have also been profiled for
various epigenetic marks\textsuperscript{11}, and enhancers in these tissues have been defined. While dynamic DNA methylation has been shown to occur primarily during early development, this property has only been shown to occur in adult tissues in some signaling pathways\textsuperscript{12}. Therefore by providing comprehensive maps of 5hmC in adult tissues, as well as cells in different stages of early development, we can gain deeper insight into the potential roles of this epigenetic mark. Furthermore, 5hmC has been shown to serve as a biomarker for disease\textsuperscript{13,14}. In order to understand how to use 5hmC to distinguish between aberrant and disease phenotypes, the base line profiles of 5hmC in normal tissues and cells need to be determined.

A.3 Results

A.3.1 Genomewide Basepair Resolution Characterization of 5hmC

By performing TAB-seq, we quantified 5-hydroxymethyl cytosine at base pair resolution in 5 human cell types (human embryonic cells (H1); mesendoderm (MES); trophoblast-like (TBL); neural progenitor cells (NPC); mesenchymal stem cells (MSC)), and 5 Adult tissues (lung, liver, pancreas, spleen, hippocampus) (Fig 1a). Consistent with previous studies that have quantified 5hmC, with affinity purification based methods, and with TAB-seq in the brain, we observed high levels of 5hmC in the Hippocampus (Fig 1b)\textsuperscript{15}. We also observed higher levels of 5hmC in adult tissues than in ES cells (ES cells mean hmC: 2.4%; ES cells standard deviation hmC: 3.16%; Adult Tissue mean hmC: 7.02%; Adult Tissue standard deviation hmC: 10.74%; Mann-Whitney U test (p=0); KS test (p=0); Fig S1a). Genomewide, we found 5hmC to be variable across tissues. To explore the nature of this variation, we examined the distribution of 5hmC at various genomic regions in our various cell and tissue types (Fig 1c). In agreement with previous studies\textsuperscript{7,8}, we observed an enrichment for 5hmC at
promoters, intergenic regions, and enhancers, as compared to random (See supplementary Table 1a).

**A.3.2 Genomewide Characterization of 5hmC levels across Human Tissue and H1 & Derived Cell Lineages**

To better understand the distribution of 5hmC genome-wide, we characterized cytosines into low and high regions of 5hmC (Appendix Sup Fig 2a). We then identified 958 focal regions of elevated 5hmC (High State Domains of hmC (HD-hmC)), commonly marked in all cells and tissues profiled (Appendix Fig 2a). Since these regions were common to all of our cell and tissue types, we characterized the distribution of hmC at these HD-hmC in adult tissues and cell types (Appendix Fig 2b), as well as their sizes (Appendix Fig 2c). Examining the enrichment of these regions at various genomics features, we found HD-hmC regions to be enriched at enhancers in the H1 and H1 derived cells, as well as in adult tissues, as well as at intergenic regions (Appendix Fig 2d, Appendix Table S2). Furthermore, in comparison to random background regions in the genome of similar size, HD-hmC regions where approximately 3 times as likely to fall at exonic regions versus promoter regions (tss +/- 250)(p = 4.8e-4). HD-hmC regions also showed the greatest enrichment at H1 and H1 derived enhancers (Appendix Table S2).

Since we found HD-hmC regions to be both enriched at intergenic regions, as well as, commonly located on the gene body instead of the promoter, we tried to gain more insight into the types of genes found at these regions by conducting GO analysis using genes located in HD-hmC regions (Appendix Sup Table 1B,C). We found that these genes were enriched for biological processes related to negative regulation of
cellular, metabolic, and biological processes (p = 8.5e-3, p=1.1e-3, p = 3.13e-2), and in particular, negative regulation of transcription of DNA-templated transcription, transcription from the RNA polymerase II promoter, and gene expression (p= 3.25e-2, p=4.10e-3,p=4.19e-2). Additionally, we found these genes to be enriched for the cadherin and wnt signaling pathways (p = 1.08e-7, p=1.49e-5)

A.3.3 Hydroxymethylation at Enhancers

Since we observed earlier that hmC was enriched at enhancer regions, an observation consistent with previous studies conducted in human ES cells, we next sought to characterize 5hmC at these genomic regions. Enhancer states are typically characterized into four categories (inactive, primed, poised, and active), and are defined by several chromatin marks. Poised enhancers are most prevalent in ES cells, and are characterized by repressive marks in regions of open chromatin; whereas, active enhancers are characterized by activating marks, such as H3K27ac and RNA Polymerase II. Since poised enhancers are common in ES cells, we focused further analysis on identifying differences between 5hmC and enhancers in active and poised states.

We previously profiled histone marks across human embryonic and derived lineages. Using this data, we identified active and poised enhancers (enhancers marked with both H3K4me1 and H3K27ac and those marked with only H3K4me1, respectively)\textsuperscript{10}. Additionally, we used a previously identified set of active enhancers in adult tissue for tissue enhancer analysis\textsuperscript{11}. To identify target genes for these enhancer, we used the nearest gene, located upstream, within, or downstream of the enhancer.
A.3.4 Enhancer body 5hmC and Gene Expression

Previous studies have suggested that enhancer methylation might affect gene expression through various mechanisms\(^\text{18}\). We wondered if enhancer 5hmC is predictive target gene expression. Using a set of enhancers found across all of cells and tissues (excluding hippocampus), we correlated enhancer 5hmC and target gene expression for each gene. At a 16% FDR rate, we identified 6% of enhancers as having strong correlations (positive and negative) between enhancer 5hmC and gene expression (Appendix Fig 4a, Appendix Sup Table 1q). While this does not imply that the changes in enhancer 5hmC cause the changes in target gene expression, it suggests that the two are either directly related, or related through some confounding factor. GO analysis of these genes revealed an enrichment in several biological processes, including, RNA splicing related terms, chromatin and chromosomal organization; the top molecular functions that were enriched were those related to RNA and nucleic acid binding (Appendix Sup Table 1r).

A.3.5 Predicting Changes in Enhancer State

Enhancer state is determined by the presence of histone marks and open chromatin. However, the mechanisms that cause changes in enhancer state are an area for exploration. In particular, the role of 5hmC in this context is yet to be understood. One hypothesis with mounting support is that TET activity results in demethylation, accompanied by an increase in hydroxymethylation at the enhancer, which then is associated with changes in enhancer state\(^\text{18}\).

In order to further investigate the role of 5hmC at enhancers, and whether this is a passive or active role, we focused on enhancers that changed in state (active to poised or poised to active) in H1 and derived lineages. First, we characterized 5hmC
levels at enhancers identified as active or poised in each cell type (Appendix Fig 4b), and observed that the distribution of 5hmC at active enhancers is significantly different than that of poised enhancers (Appendix Fig 4c, KS Test, p = 2e-202) in H1 and H1 derived cells. 5hmC levels are higher in active enhancers, in comparison to poised enhancers, which is consistent with previous studies \(^7,19\). Next, we looked across enhancers that changed in state, from H1 to a derived lineage, to understand if changes in enhancer 5hmC, could potentially be related to changes in enhancer activity. We noticed that, for enhancers that switched from an active to a poised state during differentiation, changes in 5hmC were significantly different (KS test, p <1e-25 for all cell types) for enhancers that switched from a poised to an active state (Appendix Fig 4d, Appendix Sup Fig 4b).

Since these distributions were different, we next focused on whether differences in mC (delta mC) and 5hmC (delta 5hmC) between enhancers were predictive of switches in enhancer state. By using changes in 5hmC and 5mC levels at enhancer regions, classified as switching from an active or poised to a poised or active state, between H1 and a given derived cell line, we were able to measure how well each of these epigenetic marks could independently predict changes in enhancer state (Appendix Sup Fig 4 f-j; Appendix Fig 4e). In order to understand if changes in 5hmC and 5mC levels, at a given enhancer, which changed in state between H1 and a derived cell line, were correlated in each cell type, we looked at the spearman correlations of these two changes in each of the H1 derived lineages. If this correlation is negative, this would indicate that the difference in mC levels and hmC levels, at a given state-changing enhancer, were inversely correlated; if this correlation is positive, this would
likely indicate that the rate at which cytosines were being methylated, and the rate at which mCs were being converted to 5hmC was much greater than the rate at which 5hmC was being converted to further oxidized forms (5fc,caC). Notably, we found that in MSC, this correlation was negative and strong (\(-.79\)) (Appendix Sup Fig 4f). Next, we independently checked the correlations of changes in 5hmC, and changes in 5mC with change in enhancer state (Appendix Sup Fig 4f). A positive correlation would indicate that an increase in the difference of 5hmC levels or difference in 5mC levels was associated with a switch from a poised state in H1, to an active state in a derived cell type; a negative correlation would indicate the opposite. All correlations were significant, and we found that for delta 5hmC, this correlation was weak and negative for all cell types but TBL, for which it was weak and positive, and for delta mC, this correlation was weak and negative for all cell type but MSC, for which it was weak and positive. Since none of the correlations between delta mC/hmC and change in enhancer state were extremely strong, we next applied a simple classifier to measure the predictive-ness of the change in both of the epigenetic marks (mC, 5hmC) in predicting change in enhancer state (Appendix Fig 4e, Appendix Sup Fig 4g).

To assess the relative importance of changes in enhancer 5hmC/mC enhancer-body in predicting changes in enhancer chromatin state, we examined how well a combination of these marks could predict this change by using a linear SVM (Appendix Fig 4e, Appendix Sup Fig g-j). We found that in all cell types, changes in either 5hmC or 5mC individually, at the enhancer body of enhancers that changed in state between H1 and a derived cell line, were better predictors of changes in enhancer state, than using a combination of the two marks (Appendix Fig 4d).
To understand the unexpected positive correlation between delta 5hmC with delta mC in NPCs, we looked at the relationship of 5hmC and mC at enhancers genomewide. Heatmaps of H1 derived cell types indicate a seemingly parabolic relationship between enhancer mC and 5hmC (Appendix Fig 4e, Appendix Sup Fig 4e). For both active and poised enhancers with low methylation, there is a positive correlation between enhancer 5hmC and mC. However, for high levels of mC, there is a negative correlation between an enhancer 5hmC and mC. This suggests that a certain level of enhancer mC must be present, before the rate at which 5hmC oxidation is faster than the rate at which mC is being oxidized into 5hmC. We calculated this “methylation turning point” (Appendix Sup Fig 4d) for each cell type, and found that this level was lower for active enhancers, than for poised enhancers. Since we were curious about the NPC enhancers that changed in state, following differentiation from H1, we also calculated this turning point for all enhancers that changed in state, in each cell type. We found this turning point to be higher in enhancers that changed form a poised state in H1 to an active state in NPC, as opposed to enhancers that changed from an active state in H1 to a poised state in NPC, explaining the positive correlation we observed earlier between changes in 5hmC and mC between enhancers that change in state from H1 to the NPC lineage. In all other cell types this turning point was higher for enhancers that changed from an active state in H1 to a poised state in the derived lineage, as opposed to enhancers that changed from a poised state in H1 to an active state in the derived lineage.

A.4 Discussion

We observed an enrichment of HD-hmC regions at enhancers, as well as at exonic regions as opposed to promoter regions. The enrichment for HD-hmC at active
enhancers; however, is consistent with previous studies showing high levels of hmC at enhancers\(^7,8,16\). The enrichment of genes found at HD-hmC in the Cadherin and WNT signaling pathways is interesting; however, since we did not observe a strong correlation between gene-body hmC and gene expression for most genes, there is not enough evidence to suggest that gene-body hmC levels could be directly impacting gene expression.

In H1 and H1 derived cells genes with a significant correlation between gene body hmC and gene expression that are enriched for biological processes related to gene regulation, including regulation of gene expression, splicing, chromosome organization, and RNA polymerase activity, also showed a strong correlation between gene body mC and gene expression, indicating, that if methylation or hydroxymethylation are playing a role in regulating these genes, the role of mC is probably stronger than the role of hmC. In Adult Tissues, where dynamic DNA methylation is known to be less prevalent, genes with a strong correlation between gene-body hmC and gene expression, but not gene body mC and gene expression are enriched for biological functions related to splicing, chromatin organization and SAM dependent methyl-transferase activity. This suggests that perhaps, hmC has a distinct role from mC, if any, in potentially regulating these genes.

The role of individual TET proteins at enhancers is a current area for exploration\(^20\). However, it has been shown that \(TET2\) is influential in maintaining 5hmC at enhancers\(^21,22\). Consistent with this observation, we observed that TBL enhancers (active and poised), had the highest levels of 5hmC, compared to other lineage specific enhancers, and the expression of \(TET2\) was highest in the TBL cell type. Our
observations of mC and 5hmC profiles at active and poised enhancers in ES cells suggests that a certain level of enhancer mC must be present, before the rate at which 5hmC is being converted back into C is faster than the rate at which mC is being demethylated into 5hmC. Additionally, it appears that the level of mC that must be present for this to happen is higher in MSC cells than other H1 derived cell types.

If changes in mc and 5hmC are predictive of changes in enhancer state, this suggests that the loss or gain of these epigenetic marks is either causal, or the change in enhancer state is causing a change in these marks. Our results show that the separate predictive abilities of enhancer-body changes in mC and 5hmC, between H1 and derived lineages, in predicting changes in enhancer activity, are relatively similar for most cell types, except for NPC & MSC. This suggests that if 5hmC is playing an active role in the processes that either drive or result from changes in enhancer state, and is not purely a byproduct of DNA demethylation at enhancers, the role of 5hmC is relatively minor. However, in MSCs, since there is a strong inverse correlation between changes in enhancer-body 5hmC and changes in enhancer-body mC, but changes in enhancer-body 5hmC are a much better predictor of change in enhancer activity, 5hmC is likely playing an active role, in this cell type, in the biological process that either drive, or result from change in enhancer activity. This resonates with some hypotheses where enhancer-body 5hmC prevents repressive histone marks from binding onto enhancers, in order to regulate enhancer state and gene activity 18.

These baseline profiles of 5hmC in 5 adult tissues and 5 cell lines serve as a reference for others to study 5hmC, and have allowed us to attempt to identify potential roles of this epigenetic mark in gene regulation, which can be investigated in future
studies. In the broader context of biomedical research, 5hmC has been shown to act as a biomarker for disease stage and outcome for certain diseases\textsuperscript{13}. Therefore, baseline profiles of this mark could have a long-term translational use, by assisting researchers in designing assays to distinguish between normal and aberrant phenotypes.
A.5 Appendix Figures

Appendix Fig 1 Genomewide Basepair Resolution Characterization of 5hmC
A) UCSC Genome Browser visual representation of dynamic, base-pair resolution of hmC B) Violinplots of genome-wide distribution of hmC in adult tissues and H1 and H1 derived lineages C) hmC distributions at cell-type specific enhancers (left); promoters (tss +/- 1kb) (center); intergenic regions (right)
1a
hmC in H1 and Derived Lineages and Human Adult Tissues

1b
hmC levels in H1 and Derived Lineages & Human Adult Tissue

1c
Enhancer hmC Levels for Cell-Specific Enhancers in ES Cells and Adult Tissues

Promoter hmC Levels in ES Cells and Adult Tissues

Gene-body hmC Levels in ES Cells and Adult Tissues
Appendix Fig 2 Genomewide Characterization of common regions of elevated hmC across Human Tissue and H1 & Derived Cell Lineages
A) Visual representation, of a typical HDhmC region (focal region of elevated 5hmC)
B) Distribution of hmC at all HDhmC regions in H1 and H1 derived cell lines and adult tissues
C) Size distribution of HDhmC regions (subset out the zoomed in region)
D) Enrichment of HD-hmC regions at various genomic features in comparison to random size-matched regions (fischer’s exact test used for determining significance)
Appendix Fig 3 Gene-body hmC Variation
A) Gene-body hmC profiles of genes found to have highly variable gene-body hmC in H1 and H1 derived cells and adult tissues, respectively (variability determined by PCA on gene-body hmC). Values in heatmaps are first normalized by gene (row) and then clustered by gene.
B) Bi-clustering of gene by gene spearman correlations across all H1 and H1 derived cell types, for genes with highly variable gene-body hmC in H1 and H1 derived cells. Values in heatmap are spearman correlations.
C) Distribution of spearman correlations of gene-body hmC and expression in H1 and H1 derived Cells (green) and Adult Tissues (red), with corresponding random distributions
D) Gene-body hmC and expression profiles for genes found to be correlated (p <=.015; spearman correlation = -1 or 1) in H1 and H1 derived cells (top) and adult tissues (bottom). Genes with a significant correlation between gene-body hmC and expression in H1 and H1 derived cells tends to show a positive correlation; whereas, those in adult tissues tend to show a negative correlation.
Appendix Fig 4 Enhancer hmC
A) Distribution of spearman correlations between cell & tissue-type specific enhancer hmC and expression of the nearest genes to the enhancer.
B) Distribution of hmC distribution at cell-type specific active and poised enhancers for each H1 and H1 derived cell type.
C) Distribution of hmC at all cell-type specific active and poised enhancers in H1 and H1 derived cells.
D) Changes in hmC and mC between enhancers that either switch from a poised state to H1, to an active state in MSC, or an active state in H1, to a poised state in MSC (left).
E) ROC curve on left, showing AUC for the “simple classifier” (higher AUC indicates better performance); ROC curve on right, with averaged AUCs for each cell type, indicating performance of an SVM in utilizing changes in enhancer body hmC and mC to predict changes in enhancer state. (See S4f-S4j for more details)
F) Left shows mC & hmC profiles at all active/poised enhancers for NPC; compared to enhancers that change in state between H1 and NPC (right) (See S4d,e for profiles of other cell types).
Appendix Sup Fig S1
A) Genomewide distribution of hmC in H1 and H1 derived cells (left); in Adult Tissues (right). Adult Tissues have significantly higher levels of hmC.  B) hmC levels at various genomic features in Adult Tissues (left) and H1 and H1 derived cells (right).

Appendix Sup Fig S2
A) Distribution of hmC genomewide; at regions classified by our hmm implementation as “high” and “low” in hmC.  B) Plot showing likelihood of observed hmC values given different values for “K” (number of states) used in our hmm implementation.
Appendix Sup Fig S3

A) PCA Plots for PCA on Gene-body hmC in H1 and H1 derived Cells (Left); in Human Tissue (right); in all cell and tissue types (bottom)  
B) PCA plots for PCA on Gene-body mC in H1 and H1 derived cells (left); in Human Tissue (right);  
C) Distribution of Spearman correlations for genes found to have a significant correlation between gene-body hmC and expression across all 10 cell and tissue types, with corresponding random distribution (left). Heatmap of gene expression and hmC for the significantly correlated genes (FDR = 16%) (right).
S3a

H1 and H1 Derived Cells Gene-body hmC

Human Tissue Gene-body hmC

All Cells and Tissues Gene-body hmC

S3b

H1 and H1 derived cells Gene-body mc

Adult Tissue Gene-body mc

S3c

Genebody hmC & Expression Correlations

hmC and Gene Expression in Significantly Correlated Genes
Appendix Sup Fig 4
A) Enhancer size distributions for active and poised enhancers in H1 and derived
B) Distributions of changes in hmC for enhancers that changes in state between H1 and each derived lineage. All differences in distributions of delta enhancer hmC are significantly different and p-values are listed in the figure (KS test).
C) Distributions of changes in mC for enhancers that changes in state between H1 and each derived lineage. All differences in distributions of delta enhancer hmC are significantly different and p-values are listed in the figure (KS test).
D) Methylation Turning Point of all cell-type specific enhancers in H1 and H1 derived cells (active and poised), compared to cell-type specific enhancers that change in state between H1 and a derived lineage.
Appendix Sup Fig 4 Continued

E) mC & hmC profiles at cell-type specific enhancers, for those that change in state between H1 and a derived lineage (top), and for all enhancers (active and poised), within H1 and H1 derived cell types (bottom). C) Spearman correlations of delta hmC between state changing enhancers; delta mC between state changing enhancers; and changes in enhancer state for all derived lineages (Top). Below is a table summarizing spearman correlations shown on top.
Appendix Sup Fig 4 Continued

G) Table summarizing the average performance, after 10 fold cross validation, of the simple classifier on hmC and mC independently, as well as the performance of both the Linear SVM and Random Forest (Extra Trees Classifier) when using both epigenetic marks (mC & hmC).

H) Precision and recall of Linear SVM on predicting changes in enhancer state, using 10 fold cross validation (left); Precision and recall of Random Forest (Extra trees classifier) on predicting changes in enhancer state, using 10 fold cross validation (right).

I) ROC curves for the Linear SVM on predicting changes in enhancer state, for each cell type, using 10 fold cross validation.

J) ROC curves for the Extra Trees classifier on predicting changes in enhancer state, for each cell type, using 10 fold cross validation.
A.6 Methods
A.6.1 Data Processing

RNA-seq processing

Publicly available RNA-seq data from\textsuperscript{10} was aligned to hg18, using bowtie, and quantified using cufflinks. Human Tissue Expression data was obtained from GEO (GSE16256). Human Tissue data was mapped to hg18.

hmC and mC processing

We used novoalign to map the H1 and derived cells, and bismark to map the Adult Tissues to the hg18 genome. Basecalls with a depth of less than then 10 and PHRED score below 20 were excluded from all analysis.

A.6.2 Data Analysis

Calculation of hmC at genomic features

Cytosine basecalls were summed and divided by average depth across gene-body (genes defined in HG18); across enhancer body (active and poised enhancers in H1 derived were defined by presence of H3K27ac and H3k4me1, or H3k4me1, respectively\textsuperscript{10}; enhancers in adult tissues were defined by \textsuperscript{11}. Background at each genomic feature was calculated by shuffling cytosine basecalls.

Identification of HD-hmC regions

In order to identify genomic regions containing continuous cytosines that have an increased level of hmC, relative to the regions around them, we used a HMM. Since we wanted to identify regions that had this property across all of our tissues/cells, we applied the hmm to a merged data set, where the cumulative number of cytosines and depth were merged. In order to figure out how many states to use in the HMM, we calculated a score for each state ranging from 3 to 9. We chose 9 states to get a clear
delineation between observations (hmC levels at each cytosine) in the lowest and highest states (Appendix Sup Fig 2).

**Genomewide classification of hmC levels & HD-hmC detection:**

We applied a hidden markov model, assuming that the underlying distribution of the merged base-calls was a mixture of 9 gaussian distributions (using the python library scikit-learn implementation of GMM-HMM), to merged cytosine base calls, across the ten tissues. Cytosine base-calls corresponding with the state associated with the highest levels of hmC were classified as “high hmC bases.” See Appendix Sup Fig2 for the distribution of hmC levels at cytosines classified as low and high. Consecutive base pairs in the same (high or low state) were merged to identify larger domains of hmC.

Domains in the state coinciding with the highest levels of hmC were termed “HD-hmC.” Since we wanted to filter for regions with high levels of hmC, we used a cutoff for the minimum amount of hmC that must be contained by every cytosine in a genomic region termed as HD-hmC. This cutoff was determined based on the figure above, and was set as 10%. Additionally, we observed that median size of these regions was much smaller than the mean size of these regions, so in order to use a more representative set of regions, we chose to the top 50% of HD-hmC regions, in terms of their sizes.

Enrichment analysis of HD-hmC at various genomic features was calculated using Fischer’s Exact test. 5 set containing 958 random regions, sized match to HD-hmC regions were pooled together and overlapped with each of the genomic features below to calculate the odds-ratios/p-values observed below:
Appendix Table S2:

<table>
<thead>
<tr>
<th>Genomic Feature</th>
<th>oddsratio</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>1.8513</td>
<td>3.88831e-15</td>
</tr>
<tr>
<td>Exons</td>
<td>1.79337</td>
<td>0.000237254</td>
</tr>
<tr>
<td>TSS+/- 250</td>
<td>2.22506</td>
<td>0.0185345</td>
</tr>
<tr>
<td>H1 Active Enh</td>
<td>22.2703</td>
<td>1.27349e-107</td>
</tr>
<tr>
<td>H1 Poised Enh</td>
<td>12.1525</td>
<td>3.53969e-45</td>
</tr>
<tr>
<td>MES Active Enh</td>
<td>19.37</td>
<td>1.74753e-76</td>
</tr>
<tr>
<td>MES Poised Enh</td>
<td>1.43037</td>
<td>0.524423</td>
</tr>
<tr>
<td>TBL Active Enh</td>
<td>25.0219</td>
<td>1.01751e-85</td>
</tr>
<tr>
<td>TBL Poised Enh</td>
<td>25.2132</td>
<td>1.55866e-28</td>
</tr>
<tr>
<td>NPC Active Enh</td>
<td>23.9244</td>
<td>4.32331e-93</td>
</tr>
<tr>
<td>NPC Poised Enh</td>
<td>23.7515</td>
<td>3.25566e-24</td>
</tr>
<tr>
<td>MSC Active Enh</td>
<td>30.2501</td>
<td>2.23384e-62</td>
</tr>
<tr>
<td>MSC Poised Enh</td>
<td>19.5799</td>
<td>7.60089e-29</td>
</tr>
<tr>
<td>Tissue Enh</td>
<td>8.5815</td>
<td>1.7244e-112</td>
</tr>
<tr>
<td>Liver Enh</td>
<td>7.79747</td>
<td>5.62824e-44</td>
</tr>
<tr>
<td>Lung Enh</td>
<td>10.1654</td>
<td>1.51309e-78</td>
</tr>
<tr>
<td>Pancreas Enh</td>
<td>8.02689</td>
<td>6.48222e-51</td>
</tr>
<tr>
<td>Spleen Enh</td>
<td>10.3231</td>
<td>5.01439e-69</td>
</tr>
<tr>
<td>Alt Promoter</td>
<td>6.48005</td>
<td>0.0003583</td>
</tr>
</tbody>
</table>

**Gene Enrichment of Genes Falling in HDhmC**

GO analysis was performed using Panther \(^{23}\) (Appendix Sup Table 1c). Genes overlapped by any of the 958 HDhmC regions were compared against a background set of genes (genes overlapped by any of the 5 sets of 958 random, size matched HDhmC regions).

**Gene-body hmC**

We first filtered our gene set to genes with hmC values above a minimum threshold in ES cells and Adult Tissues, respectively. This threshold was determined
using the fully methylated TAB-seq treated pUC19 control. Since this control should theoretically contain no hydroxymethylation, we first calculated TAB-seq values for these cytosines, and then used 2*(variance in these hmC values for each cell type) to serve as a minimum hmC threshold in each cell/tissue type. Genes containing above this “minimum threshold” in at least one cell/tissue type, were kept for further analysis.

PCA—We used the python library scikit learn’s implementation of PCA, and normalized FPKM and hmC matrices, prior to doing PCA (“whiten” = True). The full set of genes mapped to hg19 were used for FPKM values, and the full set of genes from hg18 were used for gene body hmC calculations. Top hits from PCA were defined as genes in the 95th percentile for magnitude of their loadings on PC1. The list of these genes in H1 and H1 derived cells, and adult tissues can be found in appendix supplementary tables (1b,c). See Appendix Sup Fig 3 for PCA plots of gene-body hmC and mC in H1 and H1 derived Cells and Adult Tissues.

We also looked at how well PC1 and PC2 explain the variance in gene expression (RNA-seq) in the same set of genes. Table below contains the percent of variation among the cells and/or tissue types that each PC explains:

<table>
<thead>
<tr>
<th>PCA</th>
<th>ES RNAseq</th>
<th>ES hmC</th>
<th>Adult RNA Seq</th>
<th>Adult hmC</th>
<th>Merged RNAseq</th>
<th>Merged hmC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>.99</td>
<td>.58</td>
<td>.66</td>
<td>.91</td>
<td>.56</td>
<td>.92</td>
</tr>
<tr>
<td>PC2</td>
<td>0.06</td>
<td>.23</td>
<td>.21</td>
<td>.06</td>
<td>.18</td>
<td>.04</td>
</tr>
</tbody>
</table>

When identifying genes that had a strong spearman correlation of gene-body hmC, across the H1 and H1 derived cell lines, we kept genes whose average spearman correlation across the cell types, between itself and all other genes, was greater than or equal to .6.
**Correlations of hmC and Gene Expression**

We calculated the spearman correlations for each gene across all tissues. Our random set used to calculate FDR, was created by shuffling FPKM values, across and within cell/tissue types. Genes identified as significantly correlated across all 10 cell/tissue types were identified using a 16%FDR. Genes identified as significantly correlated across H1 and H1 derived cells and Adult Tissues had (p<=.05) where p was determined by the randomized distribution for each set of cells and tissues.

**Enhancer body hmC**

**Genebody hmC and Target Gene Expression Correlations**

A set of enhancers identified by Leung et al. 2015, common to H1,MES,TBL,NPC,MSC,Lung,Liver,Pancreas,Spleen was used to define enhancers (2 kb regions). Form this list the nearest gene within, upstream, or downstream to the enhancer was identified from all genes in hg18.

**Enhancer hmC and mC**

Methylation Turning Point was calculated using python’s numpy library’s curve fit function. The inflection point of this function was calculated, and used as the “methylation” turning point. (See Appendix Sup Fig 4)

**Enhancer State Switching Predictions**

**Feature Correlations**

In order to determine how well each feature (change in hmC, change in mC) was correlated with change in enhancer state, we looked at the spearman correlations in each tissue type(See S4f for correlations).
**Simple Classifier**

In order to understand the predictiveness of each mark independently, we applied a simple classifier to delta hmC and delta mC respectively. Delta hmC and delta mC values were binarized based on whether or not they were positive or negative. These binary values (predictions), were compared to changes in enhancer state (true classification), which had also been binarized based on whether or not they were switching from an active to poised, or a poised to active state, between H1 and a derived lineage. ROC curves were calculated by comparing the predictions to the true classification. (ROC curve is in Fig 4e and average performance is in Appendix Sup Fig 4g).

**Linear-SVM and Random Forest for Classification of Changes in Enhancer State**

In order to use both delta hmC and delta mC to predict changes in enhancer state, we assessed the performance of two classifiers (Linear SVM and Random Forest) on each cell type. Using 10 fold cross validation, we compared the precision and recall of each classifier, and we compared the average rocs. From this it appeared that the SVM performed better. The python library scikit-learn’s implementations of linearSVM and ExtraTreesClassifier were used to perform the analysis). (See Appendix Sup Fig 4 for average performance, ROC curves and other performance metrics).

**A.7 Acknowledgements**

Gary Hon and Miao Yu created TAB-seq libraries to generate hmC maps in these cell lines and tissues. Siddarth Selvaraj, Gary Hon, and Anugraha Raman mapped the TAB-seq sequencing data. Dr. Joseph Costello provided the hippocampal tissue. Yupeng He provided the bisulfite sequencing data for the adult tissues.
A.8 References


