Winter 1-2023

Characterization of cell type-specific molecular heterogeneity in cancer using multi-omic approaches

Min Kyung Lee
Dartmouth College, min.kyung.lee.gr@dartmouth.edu

CHARACTERIZATION OF CELL TYPE-SPECIFIC MOLECULAR HETEROGENEITY IN CANCER USING MULTI-OMIC APPROACHES

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Cancer Biology

by Min Kyung (Sarah) Lee

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

January 2023

Examining Committee:

____
Brock C. Christensen, Ph.D., Chair

____
Bonnie W. Lau, M.D. Ph.D.

____
H. Robert Frost, Ph.D.

____
Joseph F. Costello, Ph.D.

____
F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

Tumors are composed of heterogeneous cell types each with its own unique molecular profiles. Recent advances in single cell genomics technologies have begun to increase our understanding of the molecular heterogeneity that exists in tumors with particular focus on gene expression and chromatin accessibility profiles. However, due to limitations in methods for certain sample types and high cost for single cell genomics, bulk tumor molecular profiling has been and remains widely used. In addition, other facets of single cell epigenomic profiling, particularly methylation and hydroxymethylation, remains underexplored. Thus, investigations to understand the cell type specific epigenetic heterogeneity and the cooperation among various molecular layers to regulate tumorigenesis are needed. In this thesis, I utilize a multi-omic approach integrating DNA methylation, hydroxymethylation, chromatin accessibility, and gene expression profiles to investigate unique single cell type-specific features in 1) epithelial-to-mesenchymal transition and in 2) pediatric central nervous system tumors. First, I demonstrate the shared and distinct epigenetic profiles that are associated with single cells undergoing epithelial-to-mesenchymal transition. With a multi-omic approach, I identify increased hydroxymethylation in binding motifs of transcription factors critical in regulating epithelial-to-mesenchymal transition. Then, I shift my focus to characterize the cellular heterogeneity in pediatric central nervous system tumors and transcriptomic alterations associated with these tumors, while accounting for cell type composition, with single nuclei gene expression data. I detect novel pediatric central nervous system tumor associated genes that are differentially expressed. Finally, I illustrate the cytosine modification alterations that occur predominantly in the progenitor-like cell types of pediatric central nervous system tumors with a multi-omic approach. I determine associations between cell type-specific hydroxymethylation alterations with cell type-specific gene expression changes. Together, these findings emphasize the need for consideration of cellular identity to determine molecular heterogeneity that exist in various cancer contexts. Moreover, these works collectively suggest the utility of multi-omic approaches to uncover novel insights in underlying tumor biology.

Preface

The works presented here were primarily supported by the Alex’s Lemonade Stand Foundation and NIH grants (R01CA216265, R01CA253976) awarded to Dr. Brock Christensen. Other funding support information for contributors of the work are specified in each chapter. Parts of this thesis were presented at scientific conferences with the kind support of Department of Epidemiology at Geisel School of Medicine, Dartmouth Cancer Center, and Burroughs Wellcome Fund. I am thankful to the Burrough Wellcome Fund Big Data in Life Sciences grant for the opportunity to receive cross-disciplinary training in laboratory and quantitative sciences.

First and foremost, I would like to thank my thesis advisor Dr. Brock Christensen for his support and affirmations throughout my graduate training. I am truly thankful for your guidance and for the freedom you’ve provided to be creative in my research. I would like to thank my Burroughs Wellcome Fund BDLS co-mentor, Dr. Robert Frost, and my thesis committee member, Dr. Bonnie Lau, for your advice and support over the last few years in getting this work completed. I sincerely appreciate your expert feedback that has really improved the works included in this thesis. I would also like to thank Dr. Joseph Costello, for your service and time to participate as my external thesis committee member.

I am entirely grateful for the steadfast support from my parents to pursue my scientific aspirations. Thank you to my cousin, Hannah Song, for your patience and encouragement to help me achieve this goal. Without you all, I would not have made it this far personally and academically.

I would like to thank the past and present members of the Christensen and Salas labs. Thank you all for your help and feedback over the years. I have been so fortunate to work closely with you all. Finally, I would like to express my gratitude to the support I have received from the Program in Experimental and Molecular Medicine, now part of Molecular and Cellular Biology. Thank you for connecting me with amazing friends to do grad school life with and wonderful faculty members who have inspired my scientific career. My graduate training would not have been the positive experience it was, if not for PEMM.

Table of Contents

List of Tables

List of Supplementary Tables

List of Figures

List of Supplementary Figures

Chapter 1

1. Introduction

1.1. Molecular characterization of cellular identity

1.1.1. Cellular identity in lineage commitment and development

The human body is composed of more than 30 trillion cells which can be categorized into around 500 cell types[1–7]. Under various physiological conditions and intrinsic and extrinsic stimuli, a single cell proliferates and differentiates into various cell types to have specialized functions to make up a living organism[8,9]. A cell type can be defined by its morphology, phenotype, function, and lineage in the context of the organ it is in[9,10]. Cellular identity dictates unique molecular components like the epigenome and transcriptome of each cell type[11]. Cells of different functions coordinate with each other to maintain homeostasis under normal physiological conditions. As each cell type has specific roles in the body, it is essential to consider dysfunction that occurs in diseases at the individual cell type level.

1.1.2. Molecular profiling using ‘omics’ technologies

With the rapid advances in next generation sequencing and genome wide measurement technologies, researchers have been able to utilize an ‘omics’ approach to understand essential molecular components in normal and disease biology. To take an ‘omics’ approach means to take an approach to perform a “comprehensive, global assessment of molecules”[12]. Commonly used examples that fall in under this definition include genomics (DNA sequences), transcriptomics (gene expression), epigenomics (epigenetic marks like methylation or chromatin structure), proteomics (proteins), and metabolomics (metabolites). For example, instead of focusing on single candidate genes or markers as it would be done in ‘genetics’, ‘genomics’ would focus on the entire genome[12]. An ‘omics’ approach has reduced biases that exist in candidate gene methods and has improved our understanding of biological systems and networks as a whole[12].

While less commonly performed than single omics measurements, multi-omic approaches hold greater potential for understanding the intricate complexities that exist in biological systems. Biological processes are not one-dimensional but involves coordination among multiple different molecular facets. Thus, multi-omic approaches provide more opportunities to address more causative questions of regulation of normal physiological homeostasis or of disease mechanisms than single omics strategies[12].

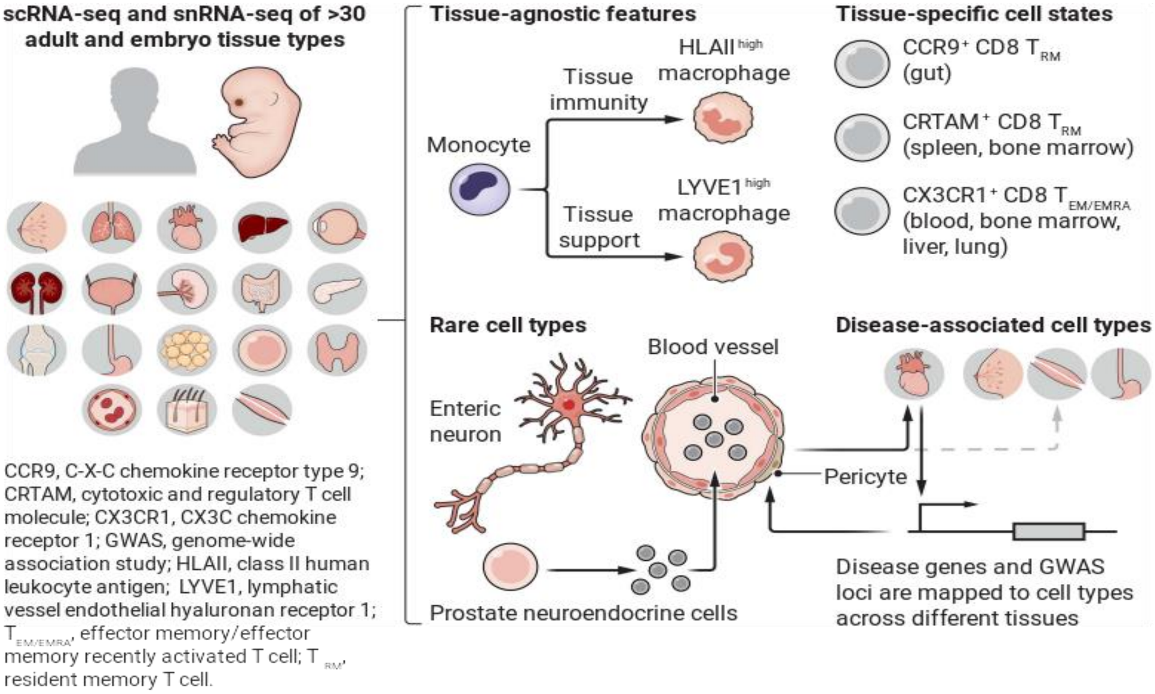
1.1.3. Technologies to identify cell types

Traditionally, distinct cell types have been sorted using approaches like fluorescence-activated cell sorting (FACS) with antibodies for known markers for cell types then profiled by sequencing or arrays. However, antibody-based sorting will only capture pure single cell populations if the antibody and the cell marker itself is highly specific and widely available[11]. If the marker is present on unknown cell types, there may still be a mixture of cell types after sorting to confound analyses. Moreover, distinct markers or antibodies for rare cell types may not have been identified yet, limiting investigations into cell types of interest.

In the past decade, single cell genomic profiling tools have revolutionized our understanding of molecular characteristics and gene functions of specific cell types in particular tissue and disease environment[9,13–16]. Various methods to profile transcriptomics, genomics, epigenomics, and proteomics among others at the single cell level have been developed to allow molecular characterization of single cell types, identification of rare/unknown cell types or cell type composition of tissues, and discovery of genes governing and regulating cellular identity programs[9,11,14,17,18]. Single cell molecular characterization approaches not only provide information on cell types, but also information on cell states that exist in a spectrum in the tissue of interest [12,13].

While single cell technologies have exponentially improved our ability to investigate biological processes and molecular profiles at the single cell level, there are challenges that remain to be resolved as we move forward. Certain cell types are more vulnerable to destruction during the initial tissue dissociation that is required for most single cell technologies. Moreover, if there are cell types that are dependent on the extracellular matrix structure, its transcriptome may be disturbed during the dissociation process and limit accurate molecular profiling for those cell types. Logistically, cost of sequencing for single cell experiments is much higher compared to bulk tissue experiments as it requires greater depth of sequencing to get enough reads per cell.

In addition, there are computational challenges that need to be solved for data generated by single cell technologies. The extremely large datasets from these methods require much higher computational power and tools to reduce technical noise from lower input material and handle zero inflation properties[14,19,20]. Moreover, assigning single cells into specific cell types have been a little more difficult as there are no standard methods for classification[9]. Cell types are currently classified based on expression of markers from previously published studies, expression on sets of genes, or by inference using other annotated single cell studies. Increasing efforts to address the variability in cell type classification have been made experimentally and computationally. One of the most extensive efforts in this realm was from the Tabula Sapiens Consortium, in which they developed a single cell transcriptomic atlas of more than 500 cell types in multiple organs in the human body that may be used as a universal reference as well as to identify tissue-specific, tissue-agnostic, or disease-associated features of cell types (**Figure 1-1**)[4–7].


scRNA-seq and snRNA-seq of >30 adult and embryo tissue types
Tissue-agnostic features
Monocyte
Tissue immunity
HLAII high macrophage
Tissue support
LYVE1 high macrophage
Tissue-specific cell states
CCR9+ CD8 TRM (gut)
CRTAM+ CD8 TRM (spleen, bone marrow)
CX3CR1+ CD8 TEM/EMRA (blood, bone marrow, liver, lung)
Rare cell types
Enteric neuron
Blood vessel
Pericyte
Prostate neuroendocrine cells
Disease-associated cell types
Disease genes and GWAS loci are mapped to cell types across different tissues

CCR9, C-X-C chemokine receptor type 9; CRTAM, cytotoxic and regulatory T cell molecule; CX3CR1, CX3C chemokine receptor 1; GWAS, genome-wide association study; HLAII, class II human leukocyte antigen; LYVE1, lymphatic vessel endothelial hyaluronan receptor 1; $T_{EM/EMRA}$, effector memory/effector memory recently activated T cell; T_{RM}, resident memory T cell.

Figure 1-1. **Single cell atlas of around 500 cell types from human tissues developed by the Tabula Sapiens Consortium**[3–6]**.**

Figure from Liu and Zhang[3].

High level of applicability, extensive use, and relative ease of technological design of single cell profiling technologies for RNA-seq and ATAC-seq have led to commercially available products for single cell RNA-seq (scRNA-seq), single cell ATAC-seq (scATAC-seq) and even MULTI-ome (10X Genomics, Parse Biosciences, etc) which combines scRNA-seq and scATAC-seq data collection. However, for some other molecular characteristics like DNA methylation, development of commercially available single-cell measurement products for widespread use has been lagging. Limitations like the harsh effects of bisulfite treatment on DNA and the need for longer reads for alignment that can be addressed in bulk sequencing but is more difficult to rectify at the single cell level. The methods for single cell epigenome profiling[17,21–23] , apart from sequencing or arraying after experimental cell type sorting, have mostly remained in academic settings.

Bulk tissue measures of genome-scale DNA methylation remain widely used due to limitations in single cell technologies. To address effects from heterogenous cell types that compose tissue of interest, computational approaches have been developed to deconvolute, or separate, signals from various cell types. Numerous cell type deconvolution methods have been developed to identify proportions for cell types that exist in bulk tissue for gene expression (RNA-seq), chromatin accessibility (ATAC-seq), chromatin contacts (Hi-C) and DNA methylation profiles[24–35]. Additionally, methods like CellDMC and Tensor Composition Analysis exist to identify specific associations of differentially methylated loci with a phenotype of interest *and* the cell type driving those alterations[36,37]. CellDMC and Tensor Composition Analysis methods incorporate the cell type fractions as interaction terms when conducting epigenome wide association studies[36,37].

1.2. DNA cytosine modifications

Although almost every cell in each individual has the same DNA, cell types have various phenotype and functions. These different functions are controlled by the epigenome. Epigenetics is “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail change in DNA sequence”[38,39]. Cellular identity is established by complex, intricate coordination of different epigenetic regulatory marks, like DNA cytosine modifications, histone modifications, and chromatin organization[40]. Here, I focus on DNA cytosine modifications as it is one of the major areas of focus of the studies in the following chapters.

1.2.1. Methylation

One of the most well-studied epigenomic marks is DNA methylation. It is a stable, heritable mark that allows for the transfer of gene regulatory programs for cell types from parent to daughter cells[41–44]. Dynamics of DNA methylation, along with its oxidized derivatives, are critical in regulating cell fate and identity[45,46]. DNA methylation is essential to normal development due to its roles in regulating gene expression, ensuring genome stability, maintaining chromatin structure, and regulating splicing[47,48]. Moreover, altered DNA methylation plays key roles in contributing to disease progression.

The overwhelming majority of DNA methylation in humans occurs at the in the context of cytosine linked to a guanine through a phosphate group (CpG) dinucleotides (**Figure 1-2**)[49,50]. DNA methyltransferases add methyl groups from S-adenosylmethionine to the fifth carbon on DNA cytosines to become 5-methylcytosine (5-mC, **Figure 1-3**)[51–53]. DNMT1 enzymes are responsible for maintaining DNA methylation marks during DNA replication that allow for the heritability[54–56]. DNMT3A and DNMT3B are generally responsible for *de novo* CpG methylation and non-CpG methylation, especially during development, and also function to maintain DNA methylation in[56–66].

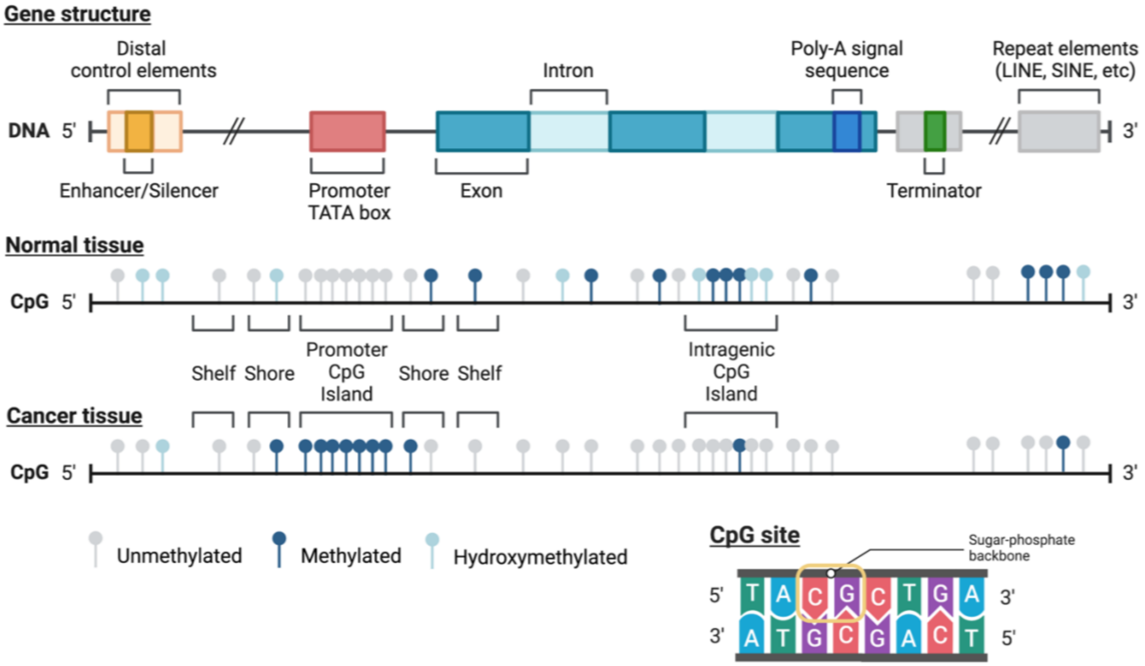


Figure 1-2. **Methylation and hydroxymethylation status at CpG islands over varying regions of the gene in normal and cancer tissue.**

Created with Biorender.com

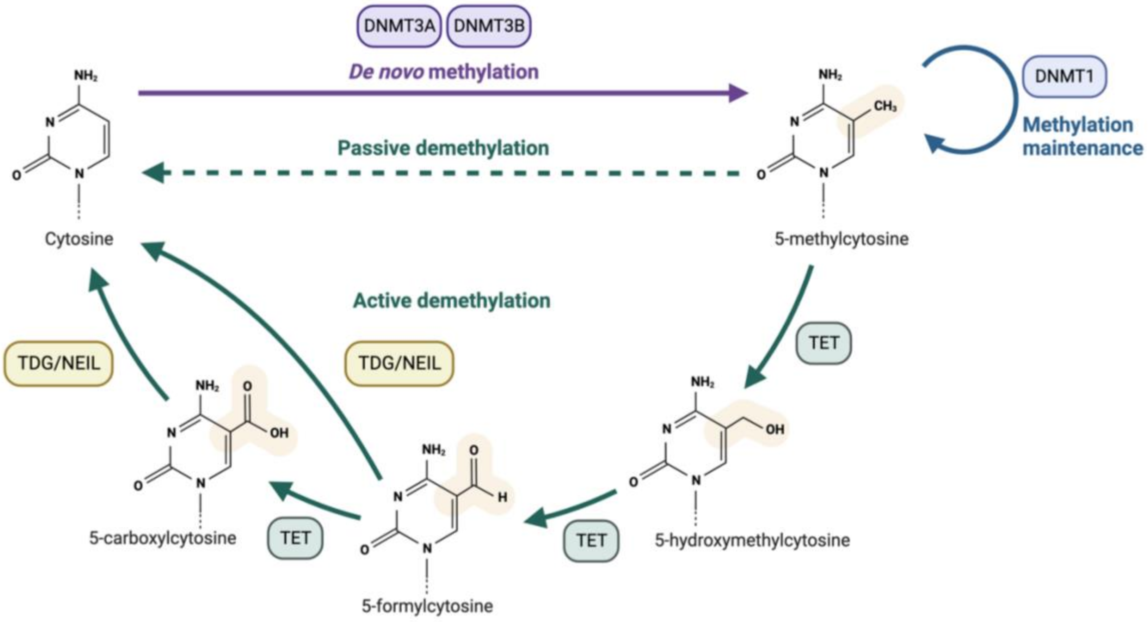


Figure 1-3. DNA methylation and demethylation pathways.
Created with Biorender.com

DNA methylation marks coordinate with histone modifications at key development and lineage specifying genes to repress pluripotency and de-differentiation during the differentiation process[67–74]. As it is essential to encoding cellular identity for the daughter cells, each cell type has distinct DNA methylation profiles[74–81]. Tissue specificity also contributes to the distinct methylation profiles for the various cell types[81–89].

Approximately 1% of the human genome is CpG sites (~30 million sites), and about half of CpG sites are in transposon derived sequences such as short and long interspersed nuclear elements (SINE, LINE respectively) and long terminal repeat (LTR) retrotransposons[59,90,91]. CpGs in repeat sequences are generally highly methylated[59,92,93]. A subset of transposable elements, retrotransposons (or Class 1 transposable elements), are mobile DNA elements of the genome that make up almost half of the human genome[94–97]. Transposable elements can contribute to genome instability and their roles in insertions and chromosomal rearrangements have been associated with certain diseases such as hemophilia and different types of cancers[98–100]. DNA methylation at the transposable elements is critical in repressing their activity to stabilize the genome[93,101–104].

A small proportion of CpGs are in clustered in regions called CpG islands. CpG islands are defined as 200 base pair segments of DNA with more than 50% GC content and observed CpG to expected CpG ratio greater than 0.6[105]. 75% of promoters are in CpG islands and remain largely unmethylated in histopathologically normal cells (**Figure 1-2**)[60,106,107]. To remain unmethylated and protect from DNA methyltransferases, CpG islands recruit complexes that include histone methyltransferases, particularly MLL which methylates H3K4 and is associated with active transcription[108–112]. CpG islands remain unmethylated and repressed by Polycomb complexes which mark the region with H3K27me3 and nucleosomes, which are associated with inactive genes particularly in embryonic stem cells[69,113–115]. Some promoter CpG islands that are methylated are associated with gene silencing in genes that would need to be stabilized in the repressed long term such as genes located on the inactive X chromosome[114]. Contrary to promoter CpG islands, many intragenic CpG islands have high levels of methylation and have been reported to associate with gene expression rather than gene silencing[78,79,116,117]. In addition, intragenic CpG island methylation has been shown to play a role in regulating splicing and polyadenylation and shown to be a product of nearby gene transcription[118–126].

Methyl-binding proteins (MBPs) are responsible for reading DNA methylation marks. Generally, these proteins have methyl-CpG-binding domains (MBDs) that can bind to single symmetrically methylated CpG sites and transcriptional repression domains (TRD) that mediate interactions with other proteins[127–133]. Methyl-CpG-binding protein 2 (MeCP2) was one of the first MBPs discovered[134]. MeCP2 interacts with different partners like histone deacetylases (HDAC1/3) to repress transcription and with BRM (part of the SWI/SNF complex) to remodel the nucleosome[135–139]. Several other

MBPs also coordinate with DNA methylation, histone modifications and chromatin organization to regulate transcription and chromatin structure[128,139].

1.2.2. Hydroxymethylation

While the existence of 5-hydroxymethylcytosine (5-hmC) was discovered in viruses in the 1950s and in vertebrates in 1975[140–142], its functional roles were largely unknown and very underexplored. It was not until in 2009, when Tahiliani et al discovered ten eleven translocation (TET1/2/3) enzymes to be responsible for oxidizing 5-mC and Kriaucionis & Heintz discovered 5-hmC in different neuron types, that research on this modification was revitalized[143,144].

5-mC can undergo active demethylation by TET enzymes which oxidizes the methyl group to produce 5-hmC (**Figure 1-3**)[143]. TET enzymes then oxidize 5-hmC to produce 5-formylcytosine (5-fC) and then oxidize 5-fC to produce 5-carboxylcytosine (5-caC). 5-fC and 5-caC, but not 5-mC and 5-hmC, are excised by thymine-DNA glycosylase (TDG) in the base excision repair pathway or alternatively by NEIL1-2 DNA glycosylases to become an unmethylated cytosine[145–148].

While 5-hmC can be an intermediate in the active DNA demethylation pathway, numerous studies have suggested that it can act as a stable mark on DNA, especially in the context of specific cell types or tissue types, that it has been called the ‘sixth’ base of the genome with 5-mC being the ‘fifth’ base[149–154]. Compared to 5-fC and 5-caC, 5-hmC is 10 to 100 times more prevalent in cell types like embryonic stem cells, neural progenitor cells and some neurons[143,144,152,155,156]. While in most tissues 5-hmC prevalence is relatively very low compared to 5-mC, it can be prevalent up to 40% of 5-mC in some cell types such as Purkinje cells[144,157]. Like 5-mC, 5-hmC is tissue type-specific. Highest levels of 5-hmC are found in the brain relative to other tissue types[150,156,158–160]. Around 50% of the 5-hmC marks all marks and tissue specific differentially hydroxymethylated regions are enriched in the gene body regions (**Figure 1-2**)[159].

5-hmC plays a key role in development and cell differentiation. Some of the same MBPs as for 5-mC can recognize 5-hmC to regulate transcription and chromatin structure[156,161,162]. In the brain and in embryonic stem cells, where there are high levels of 5-hmC, 5-hmC is enriched in gene bodies, promoters marked with bivalent chromatin signature and enhancer regions[149,160,163–171]. Moreover, 5-hmC is enriched in protein-DNA interacting sites, including interacting sites of key developmental genes like *OCT4* and *NANOG*[166,171,172]. But, 5-hmC marks can be context-specific. For example, while 5-hmC is mutually exclusive from trimethylated H3K27 regions in the brain, it is enriched in promoters marked with H3K27me3 in embryonic stem cells[164,165,167,169]. In addition, 5-hmC accumulates particularly in the gene bodies of activated genes associated with neuronal function during neurogenesis[149,167]. Contrary to neuron differentiation, 5-hmC levels decrease during embryonic stem cell differentiation[143,151,173,174]. The different TET proteins regulate 5-hmC in cell type-specific and genomic-context dependent manners. For example, TET1 is preferential to embryonic stem cells while TET3 is critical for regulating the epigenome in oocytes and zygotes[175,176]. Moreover, TET1 regulates 5-hmC in promoters and enhancers and TET2 regulates 5-hmC in gene bodies in embryonic stem cells[170,177].

5-hmC has been thought to play a role in regulating transcription in a genomic context dependent manner as well. For example, 5-hmC in gene bodies has been associated with highly expressed genes[149,178]. Furthermore, enrichment of 5-hmC in promoters have been associated with lowly expressed genes[165,169]. 5-hmC interactions with various transcription factors and other protein complexes like Polycomb repressive complex have been described as one of the possible mechanisms behind transcriptional regulation[166,170,172].

1.2.3. Methods to measure genome-wide cytosine modifications at the single base resolution

The past couple of decades with ever developing microarray and sequencing technologies have exponentially improved methods to measure DNA methylation and hydroxymethylation genome wide. There are three main methods to detect cytosine modifications: antibody, enzymatic, and chemical treatments. The antibody-based methods, called MeDIP-seq and hMeDIP-seq, immunoprecipitate regions of DNA with cytosine modifications using 5-mC and 5-hmC specific antibodies. One enzymatic method to detect 5-mC is called TAPSβ[179]. In this method, β-GT glycosylates 5-hmC to protect 5-hmC during the TET oxidation. Following the oxidation, the DNA is treated with pyridine borane which converts the final oxidized product, 5-caC, to uracil which then is read as thymine. Methylated cytosine will be sequenced as thymine, while hydroxymethylated cytosine is sequenced as cytosine.

While there have been many methods developed to detect DNA cytosine modifications, one of the most used methods to measure DNA methylation is by chemical bisulfite (BS) treatment. For this method, cytosine bases in single stranded DNA are deaminated to uracil upon the treatment with sodium bisulfite. During the PCR amplification step, the treated DNA will replace the uracil (converted cytosines) as thymine[180]. Methylated cytosines are resistant to deamination and will remain to be read as cytosines which allows discrimination of methylated and unmethylated cytosines. However, hydroxymethylated cytosines also are resistant to the deamination from sodium bisulfite[181,182]. Therefore, traditional bisulfite treatment methods for DNA methylation measures do not distinguish 5-hmC from 5-mC[183]. To distinguish between the two, an oxidative bisulfite (oxBS) treatment method, among other methods, were developed[183]. When DNA is first oxidized with a chemical like $KRuO_4$ before the bisulfite treatment, it will convert the 5-hmC to 5-fC. Sodium bisulfite then will convert 5-fC into a thymine. Subtracting bisulfite treated only and oxidative bisulfite treated signals will allow for estimation of 5-hmC[184,185].

Bisulfite treated or oxidative bisulfite treated DNA can be sequenced or measured with a microarray like the Illumina Human Methylation EPIC array. Measuring cytosine modifications with a sequencing or an array approach each have strengths and limitations. Coverage of the genome and measured genomic contexts are similar for both methods of measurement[186]. While sequencing can query more CpGs sites, or other non-CpG sites, coverage, and therefore precision, is limited by the cost from the depth of sequencing needed[186]. Methylation arrays offer better precision, reproducibility, and cost-effectiveness while being limited to only a proportion of CpGs that sequencing offers[186],[187]. These strengths and limitations need to be considered while designing experiments to measure DNA cytosine modifications.

In the studies incorporated in this thesis, we utilize oxidative bisulfite treatment complemented with Illumina Human Methylation EPIC array. The EPIC array is the most recent version of the Illumina methylation arrays, following Human Methylation 27K BeadChip and Human Methylation 450K arrays. It measures around 850,000 CpGs in diverse genomic contexts across the human genome. The methylation arrays result in beta values (β) calculated by dividing the intensity of the methylated signal by the sum of the intensity of unmethylated and methylated signal + 100. A completely unmethylated CpG will have a beta value of 0 and a completely methylated CpG will have a beta value of 1.

1.2.4. Aberrant DNA cytosine modifications in cancers

Non-mutational epigenetic reprogramming has been described as an emerging hallmark and enabling characteristic of cancer[188]. Epigenetic reprogramming contributes to other hallmarks of cancer such as enabling phenotypic plasticity and activating invasive growth programs through processes like epithelial-to-mesenchymal transition[188]. In addition, alterations in the epigenome have been suggested to be associated with predisposition to cancer and be early events in tumorigenesis [189–191].

Genetic alterations have been found in epigenetic modifiers in many tumor types. Among those genetic alterations include mutations and chromosome translocation in epigenetic modifiers of cytosine modifications, *DNMT1/3A/3B, TET1/2,* and *IDH1/2*[191–194]*.* While effects from genetic alterations epigenetic modifiers on DNA cytosine modification profiles still need further exploration, initial studies in tumor types (hematological malignancies and glioma) with frequent mutations in these genes have been associated with changes in cytosine modification profiles. For example, TET mutations in hematological malignancies contribute to hypermethylation particularly at enhancers and sites associated with hematopoietic differentiation and in euchromatin regions and is associated with decrease in hydroxymethylation[195–200]. Moreover, across numerous tumor types, IDH1/2 mutations are associated with DNA hypermethylation at gene bodies and enhancers, hypomethylation at promoters, and greater 5-hmC levels[201–204].

DNA methylation alterations are prevalent across almost all tumor types. Three main mechanisms of DNA methylation alterations exist: 1) hypomethylation of repeat elements, 2) hypermethylation of promoters, and 3) mutation at methylated cytosines[194]. Hypomethylation in the cancer genome has been suggested to contribute to tumorigenesis by increasing mutation rates, promoting genomic instability, and altering chromatin organization[205–211]. Hypermethylation of CpG island promoters is associated with transcriptional alterations, particularly silencing of tumor suppressor genes to drive tumorigenesis (**Figure 1-2**)[212–223]. Methylated cytosines contribute to increased mutations as they are hotspots for deamination that may be repaired incorrectly to a thymine instead of a cytosine and are favored to form DNA adducts[194,224–227].

While mechanisms underlying alterations DNA hydroxymethylation that contribute to tumorigenesis are less clear than that for DNA methylation, loss of 5-hmC is a common characteristic of numerous tumor types (**Figure 1-2**)[178,228–236]. Loss of 5-hmC has also been associated with poor prognosis compared with those who have relatively higher levels of 5-hmC in many tumor types as well[237–243]. The loss of 5-hmC may be an effect of inactivating mutations, downregulation of TET enzymes, or mutations in metabolic genes like IDH1/2 that produce TET cofactors[229,240,244–251]. However, as loss 5-hmC is a shared characteristic across so many tumor types, there is likely another explanation. One potential explanation for the loss of 5-hmC is that it is a mark of high proliferation as 5-hmC levels have been shown to be negatively associated with proliferation[154,178,200,236].

1.3. Tumor heterogeneity

Tumors and their microenvironments are comprised of heterogenous cell types (**Figure 1-4**). Intrinsic and extrinsic pressures provide pressure to drive clonal evolution of tumor cells to result in intratumoral heterogeneity[252]. Initial studies to elucidate tumor heterogeneity and delineate clonal evolution began with the genomic heterogeneity that exists by sequencing different regions of a single tumor[253–255]. However, as selection focuses on phenotype heterogeneity rather than genotype heterogeneity, epigenetic heterogeneity also strongly influences to tumor heterogeneity[188,256,257]. Understanding intratumoral heterogeneity is critical as it has been associated with poor prognosis, tumor progression, and therapy resistance in many different types of cancers[258–267].

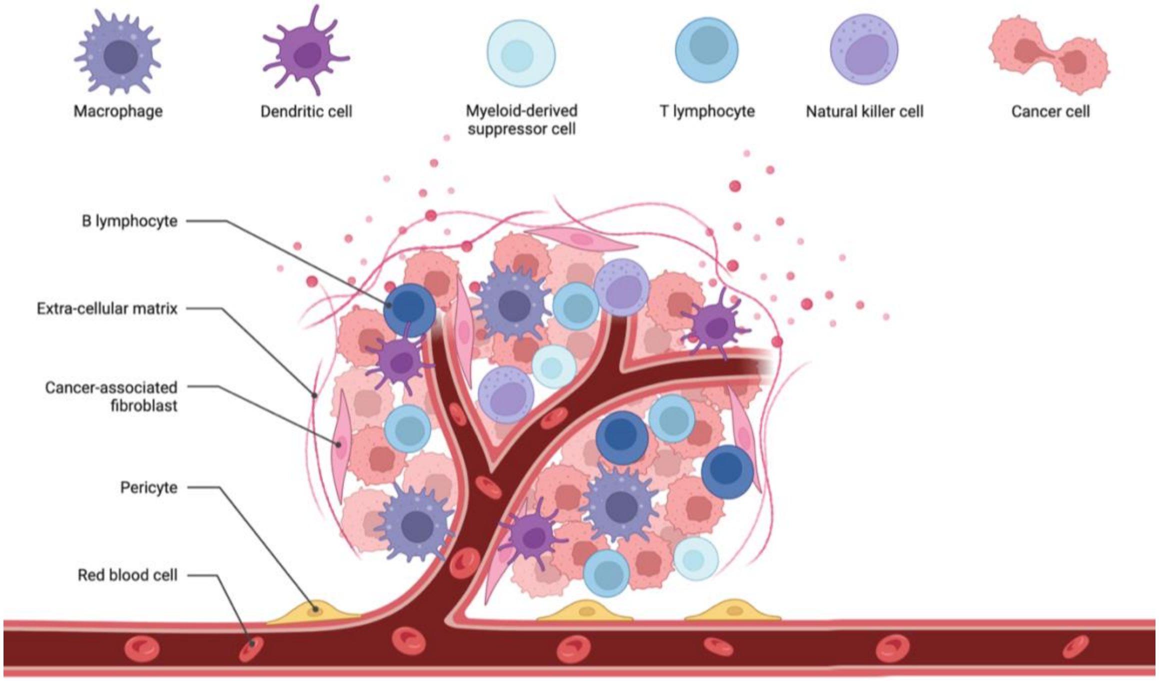


Figure 1-4. Diverse cell types present in the tumor microenvironment.
Created with Biorender.com

The conditions of the tumor microenvironment can induce tumor heterogeneity. As tumor cells are exposed to altered conditions like abnormal levels of growth factors, altered pH levels, structural changes in vasculature, and hypoxia, cells can undergo cell state transitions or adapt to its environment to lead to intratumoral heterogeneity[256,268–272]. Other external factors like cancer treatments can induce selection of clones that are genetically and epigenetically fit to survive[273–278]. These clones that can adapt and survive therapy evolve and progress to build resistance to therapies[261,262].

Cancer has been deemed to be a ‘genetic’ disease from long line of research establishing genomic alterations associations with tumor development and progression[188,279–281]. Within a single tumor, the genome of the tumor cells can vary in mutations, copy number alterations, and structural chromosomal aberrations[256,259,260,263,265,267,275,282–285]. Genetic alterations in key tumor suppressor genes or oncogenes in certain clones provide survival advantage and allows that clone to proliferate further in the tumor[260,263,266,267,282,286–289].

In addition to genetic heterogeneity, it is becoming more evident that epigenetics plays a large role in intratumor heterogeneity[272,282,290,291]. The epigenome is responsive to the factors in the tumor microenvironment even more so than the genome[256,281]. Certain epigenetic changes like hypermethylation of promoters of key cancer progression related genes (ex: *MGMT*, *MLH1*) act similar way to driver mutations to contribute to the intratumoral heterogeneity[265,292–295]. Dysregulated epigenetic modifier enzymes or epigenetic marks can also create transcriptomic heterogeneity[277,294]. Furthermore, epigenome in tumors regulates the highly plastic state which allows tumor cells to transition between cell states or differentiate into different cell types[256,257,290,296–299].

Genetic and epigenetic heterogeneity with the tumor microenvironment all work together to produce intratumor heterogeneity[273,281,291,300,301]. While current single cell technologies and sequencing strategies have allowed us to begin mapping the intratumoral heterogeneity at the single cell level in many tumor types[297,302–313], there is still a lot of work to be done in understanding the mechanisms behind intratumoral heterogeneity especially in understudied cancer types. As experimental and computational approaches to characterize intratumoral heterogeneity is ever evolving, it will be essential to incorporate it when developing new therapeutic strategies and identifying patient populations that would benefit most from these therapies.

1.4. Cancer contexts of focus

1.4.1. Epithelial-to-mesenchymal transition (EMT)

Epithelial-to-mesenchymal transition (EMT) is an essential process that plays critical roles in tumor heterogeneity, metastasis, and therapeutic resistance[314]. EMT is a cellular program in which epithelial cells, with apical-basal polarity and intact cell-cell junction properties, progresses through number of cellular states to gain mesenchymal cell type properties, such as front-back polarity and motility[315]. It is a normal process during embryonic development but is often observed in cancers undergoing invasion and metastasis. While long considered to have been a binary transition from an epithelial cell type to a mesenchymal cell type, it has recently been established that it is a stepwise process in which cells gradually transition into intermediate/hybrid cell states before becoming a mesenchymal cell type[314–316]. Various tumor microenvironment factors like infiltration of inflammatory cells or hypoxia contribute to inducing EMT[317–321].

EMT programs are largely controlled by ZEB1/2, SNAIL, SLUG and TWIST1/2 transcription factors[322]. These transcription factors regulate the other EMT transcription factor expression and are responsible for inducing the transcriptional changes that occurs when cells change states[322]. Epithelial state associated genes like *CDH1* are repressed by SNAIL and ZEB1[323–325]. Mesenchymal state associated genes like CDH2 are induced by ZEB1/2 and SNAIL[323,326,327].

Epithelial cells undergoing EMT does not always end as mesenchymal cells but can terminate its transition in the intermediate EMT state, particularly during tumor progression[315,316,328]. These intermediate EMT cell states are characterized by reduced epithelial features like expression of *CDH1* but have not fully gained mesenchymal cell type characteristics[315,328]. The intermediate EMT states display high levels of stemness and plasticity thereby able to generate phenotypic heterogeneity in tumors[314,329,330]. The plasticity in EMT is governed by various epigenetic factors as evidenced by chromatin signatures of key EMT-associated genes[322,331]. For example, promoters of *CDH1* in CD44+ stem-like cells had bivalent chromatin signature (H3K4me3 and H3K27me) while promoters *CDH1* in the CDH1+CD24+ differentiated cells had the activated chromatin signature (H3K4me3) the human mammary epithelium[332]. Moreover, *ZEB1* promoters in non-cancer stem cells also had a bivalent chromatin signature in basal breast cancer cells allowing these cells to remain plastic to be able to respond to external signals for EMT[333].

Although it is now generally accepted that intermediate states exist as meta-stable states[315], difficulty in isolating the intermediate EMT states have limited our understanding of the molecular features of these cell states. Moreover, while certain epigenomic features like histone modifications have been suggested to be important regulators of EMT, there are still other epigenomic features, like DNA cytosine modifications, that need further exploration. Given the likely responsibilities of the epigenome in EMT and the roles of cytosine modifications in establishing cellular identity, better understanding of cytosine modifications in coordination with other epigenetic factors in EMT will improve our understanding of the biological changes in EMT and may provide targets for preventing EMT progression.

1.4.2. Pediatric central nervous system tumors

Central nervous system (CNS) tumors are one of the most common and deadly cancer types in the pediatric population (0 – 19 years of age)[334,335]. Pediatric CNS tumors are comprised of a variety of tumor types based on histology, immunohistochemistry, and molecular biomarkers[336]. Incidence, survival rate, and treatment strategies varies between the different tumor types (**Table 1-1, Table 1-2**). Pilocytic astrocytoma account for a large number of the pediatric CNS tumor cases with an incidence rate of 0.95 per 100,000 but have a high 10-year relative survival rate of 95.4%. Ependymal tumors are less prevalent and have poorer prognosis with an incidence rate of 0.29 per 100,000 and 10-year relative survival rate of 69.6%. While CNS tumors are considered to be the most common solid tumors in the pediatric population, it is still very rare in the general population, with an incidence rate of 6.29 per 100,000[335]. The rarity in the general population makes it difficult to accrue large enough sample size to study these tumors. When classified further into the various subtypes, it gets even harder to accumulate high statistically powered sample size.

Table 1-1. Epidemiology of pediatric central nervous system tumor subtypes. Bolded tumor types are the major categories of subtypes. Selected for only certain tumor types. Adapted from Ostrom et al.[335]

	5-Year Total	Annual Average	Rate (95% CI)
Diffuse Astrocytic and Oligodendroglial Tumors	**2,248**	**450**	**0.55 (0.53–0.57)**
Diffuse astrocytoma	946	189	0.23 (0.22–0.25)
Anaplastic astrocytoma	365	73	0.09 (0.08–0.10)
Glioblastoma	700	140	0.17 (0.16–0.18)
Oligodendroglioma	164	33	0.04 (0.03–0.05)
Anaplastic oligodendroglioma	22	4	0.01 (0.00–0.01)
Oligoastrocytic tumors	51	10	0.01 (0.01–0.02)
Other Astrocytic Tumors	**4,371**	**874**	**1.07 (1.04–1.10)**
Pilocytic astrocytoma	3,877	775	0.95 (0.92–0.98)
Ependymal Tumors	**1,176**	**235**	**0.29 (0.27–0.30)**
Other Gliomas	**3,133**	**627**	**0.77 (0.74–0.79)**
Glioma malignant, NOS	3,093	619	0.75 (0.73–0.78)
Other neuroepithelial tumors	34	7	0.01 (0.01-0.01)
Neuronal and Mixed Neuronal-Glial Tumors	**2,012**	**402**	**0.49 (0.47–0.51)**
Embryonal Tumors	**2,397**	**479**	**0.59 (0.56–0.61)**
Medulloblastoma	1,652	330	0.41 (0.39–0.43)
Primitive neuroectodermal tumors	208	42	0.05 (0.04–0.06)
Atypical teratoid/rhabdoid tumor	382	76	0.09 (0.08–0.10)
TOTAL	**25,497**	**5,099**	**6.21 (6.14–6.29)**
Malignant	**14,586**	**2,917**	**3.57 (3.51–3.62)**
Non-Malignant	**10,911**	**2,182**	**2.65 (2.60–2.70)**

Table 1-2. Relative survival rates for pediatric central nervous system tumors by subtype and malignancy.
Selected for only certain tumor types. Adapted from Ostrom et al.[335]

	5-Year RS (95% CI)	10-Year RS (95% CI)
Diffuse astrocytoma	81.5 (80.1-82.9)	78.5 (77.0-80.0)
Anaplastic astrocytoma	28.3 (25.3-31.3)	23.3 (20.3-26.4)
Glioblastoma	19.8 (17.8-21.8)	15.9 (14.0-17.9)
Oligodendroglioma	94.6 (92.4-96.2)	89.4 (86.1-91.9)
Anaplastic oligodendroglioma	50.7 (40.7-59.8)	39.8 (29.9-49.5)
Pilocytic astrocytoma	96.8 (96.4-97.1)	95.4 (94.9-95.9)
Ependymal tumors	78.5 (76.8-80.1)	69.6 (67.5-71.5)
Glioma malignant, NOS	70.0 (68.9-71.1)	68.6 (67.5-69.7)
Neuronal and mixed neuronal-glial tumors	79.2 (74.4-83.2)	77.6 (72.6-81.9)
Embryonal tumors	64.8 (63.6-65.9)	59.6 (58.4-60.8)
TOTAL	75.6 (75.2-76.1)	72.1 (71.6-72.6)

Clinical challenges specific to pediatric CNS tumors necessitates deeper investigations into these tumors. First, there are many detrimental late effects from their cancer treatments even after patients have survived their initial pediatric CNS tumors. Childhood CNS tumor survivors face the highest rate of cumulative burden of chronic or disabling conditions later in life than any other tumor type survivors[337]. For instance, cranial radiation, one very commonly used treatment option in these tumors, has been associated with neurocognitive late effects like intellectual and academic decline and physical late effects like stroke[338–345]. Moreover, although mortality rates generally for pediatric cancers have been significantly reduced since the 1970s, reduction in mortality rates for pediatric CNS tumors have not been as drastic[346]. Some cancer types like Hodgkin lymphoma and gonadal tumors have seen more than 80% reduction in mortality since 1975 while pediatric CNS tumors only have had a reduction of around 29%[346]. To address the discrepancies in mortality rate reduction and to improve the quality of life post surviving CNS tumors, treatment strategies for pediatric CNS tumors still need to be further developed.

To begin to understand the underlying mechanisms for disease in pediatric CNS tumors, efforts have been made to profile molecular landscape and develop finer tuned subtypes for some pediatric CNS tumors. One of the earliest efforts to incorporate molecular distinctions was done in medulloblastomas. Utilizing an integrative approach of genomic and transcriptomic medulloblastoma profiles, four major molecular subgroups (WNT, SHH, Group 3, Group 4) of medulloblastomas were established[347–352]. Recurrence and prognosis vary within each molecular subgroup. For example, the WNT subgroup, which is defined by genetic alterations in *CTNNB1, DDX3X, SMARCA4,* and *TP53* and loss of chromosome 6, has a very good prognosis and is less likely to be metastatic[352]. On the contrary, the Group 3 subtype, which is defined by genetic alterations in *SMARCA4, KBTBD4, CTDNEP1,* and *KMT2D* along with chromosomal gain in 1q, 7, 18 and chromosomal loss in 8, 10q,11, and 16q, has a poor prognosis and is likely to be metastatic[352].

Subtypes for ependymoma can be defined by DNA methylation[353]. Methylation profiles, in coordination with localization, histology, and genetic alterations, were able to categorize ependymoma into 9 separate subtypes, 3 each for the anatomical localization (SP: Spine, PF: Posterior fossa, ST: Supratentorial)[353]. Survival rates vary by subtype for ependymoma as well[353,354]. The initial methylation-based classification study indicated ST-EPN-RELA has the lowest 5-year progression free survival rate at 29% and ST-SP has one of the highest 5-year progression free survival rate at 100%[353].

While efforts to incorporate molecular markers for some pediatric CNS tumor types have provided granular understanding and have improved treatment management strategies, additional studies are needed to expand to additional pediatric CNS tumor types and to increase the sample sizes due to the rarity of these tumors. Moreover, as previous studies have established the low mutational burden in pediatric cancers[355–357], it is likely that epigenomic alterations play important roles in pediatric CNS tumor initiation and progression. However, studies on epigenomic contribution to these tumor types remain limited. Thus, additional studies are needed to appreciate the roles that epigenomic aberrations may have in contributing to tumorigenesis of pediatric CNS tumors.

1.5. Summary

Molecular heterogeneity has been demonstrated to be a common feature across numerous tumor types. However, majority of our understanding of heterogeneity that exist have come at single omics layers. Further investigations on the complexities of how each molecular layers work together to regulate the heterogeneity in cancers are needed. In addition, many studies have focused on the molecular alterations in bulk tumor tissue without consideration for cell type composition effects. To capture more granular alterations than at the bulk tissue level, additional studies utilizing single cell genomics technologies or computational deconvolution methods to identify molecular changes at the cell type level are needed. This thesis aims to address some of these underlying molecular complexities that exist in cancers in the context of epithelial-to-mesenchymal transition and of pediatric central nervous system tumors at the cell type-level with integrative, multi-omic approaches.

In Chapter 2, I integrate DNA methylation and hydroxymethylation, chromatin accessibility, and gene expression data to identify roles of epigenomic characteristics in distinct cell states undergoing epithelial-to-mesenchymal transition. In Chapter 3, I characterize the cellular and transcriptomic heterogeneity in pediatric central nervous system tumors compared to non-tumor pediatric brain tissue with gene expression profiles of 84,700 nuclei. In Chapter 4, I build on results of Chapter 3 by integrating single nuclei RNA-seq data with genome wide methylation and hydroxymethylation data to elucidate the epigenetic heterogeneity in pediatric central nervous system tumors and to identify the epigenetic alterations associations with changes in gene expression.

Collectively, these works 1) the demonstrate integrative molecular heterogeneity in understudied cancer cell states and cell types at the cell type-specific level and 2) highlight the importance of incorporating and distinguishing DNA hydroxymethylation from DNA methylation.

Chapter 2

2. Distinct cytosine modification profiles define epithelial-to-mesenchymal cell-state transitions

This chapter was adapted from work published in

Epigenomics, Volume 14, issue 9 on May 2022.

https://doi.org/10.2217/epi-2022-0023

The following authors contributed to the work:

Min Kyung Lee, Meredith S. Brown, Owen M. Wilkins,
Diwakar R. Pattabiraman, Brock C. Christensen

*Supplementary Table 1 – 4 can be found in the online publication.

2.1. Abstract

Epithelial-to-mesenchymal transition (EMT) is an early step in the invasion-metastasis cascade, involving progression through intermediate cell states. Due to challenges with isolating intermediate cell states, genome-wide cytosine modifications that define transition are not completely understood. We measured multiple DNA cytosine modification marks and chromatin accessibility across clonal populations residing in specific EMT states. Clones exhibiting more intermediate EMT phenotypes demonstrated increased 5-hydroxymethylcytosine (5-hmC) and decreased 5-methylcytosine (5-mC). Open chromatin regions containing increased 5-hmC CpG loci were enriched in EMT transcription factor motifs and were associated with Rho GTPases. Our results indicate the importance of both distinct and shared epigenetic profiles associated with EMT processes that may be targeted to prevent EMT progression.

2.2. Introduction

Epithelial-to-mesenchymal transition (EMT) is an early step in the invasion-metastasis cascade, involving progression through a number of cellular states. It is a process by which epithelial cells lose specific properties such as apical-basal polarity, detach from the basement membrane to gain mesenchymal properties such as front-back polarity and motility[315]. Rather than being a binary conversion from an epithelial to a mesenchymal state, the EMT encompasses a step-wise progression to a mesenchymal cell state whereby the cells could display intermediate/hybrid phenotypes of both epithelial and mesenchymal cells[330,358]. As metastasis is responsible for the majority of deaths in cancer patients[359,360], it is critical to understand the molecular underpinnings of EMT.

Cells that reside in an intermediate state display more plasticity than the cells on either ends of the EMT spectrum[330,361–363]. In addition to increased plasticity, intermediate cells have been shown to harbor stem cell characteristics such as self-renewal and increased expression of pluripotent genes[364–366]. Although it is evident that there are intermediate phases when transitioning from epithelial to mesenchymal states[367–369], experimental isolation of these specific states has proven challenging. Consequently, the molecular and functional characteristics and of the intermediate states and their contribution to metastasis are poorly understood.

DNA methylation is a well-studied epigenetic mark, mostly known for its role in regulating gene expression. Methylation of cytosines (5-methylcytosine/5-mC) can occur in the context of Cytosine-phosphate-Guanine (CpG) dinucleotides and the reaction is catalyzed by DNA methyltransferase enzymes (DNMTs). Ten eleven translocation (TET) enzymes can oxidize methylcytosine to form 5-hydroxymethylcytosine (5-hmC), then 5-formylcytosine (5fC), and finally 5-carboxylcytosine (5caC)[143]. Oxidized cytosines can then be deaminated AID then undergo thymine DNA glycosylase-mediated base excision repair to an unmethylated cytosine. While around 80% of mammalian CpG dinucleotides are estimated to be methylated[370,371], hydroxymethylation accounts for a relatively modest proportion of overall cytosine modification and varies greatly with tissue type[372,373]. Although 5-hmC levels are low in relation to 5-mC in human tissues, it is most highly enriched in brain and breast tissues, relative to other tissue types[158]. While a number of studies have shown the importance of DNA methylation in EMT, these studies used traditional bisulfite treatment to measure 5-mC, which does not resolve 5-hmC[374–379]. 5-hmC can be estimated from comparing oxidized-bisulfite treatment to bisulfite treated DNA[183], as traditional bisulfite treatment does not distinguish 5-mC from 5-hmC. In comparison to general repression of transcription from 5-mC, 5-hmC is positively-associated with transcriptional activity and gene expression[380,381]. If the association is a consequence of passive dilution of 5-mC via DNA demethylation, or due to functional actions of 5-hmC is yet unclear and is likely context dependent. However, growing evidence suggests 5-hmC contributes directly to gene regulation in several specific contexts, aside from its role in DNA demethylation. At the chromatin level, 5-hmC has been shown to increase DNA flexibility and mechanical stability, and nucleosome accessibility[382]. Transcription factors and their binding sites have been associated with being colocalized with TET and 5-hmC[168,383–385], which provides possible 5-hmC mechanism of gene expression regulation through transcription factor recruitment[159].

Although decreased global 5-hmC is consistently observed in cancer[178,386–388], few studies have measured cancer-associated 5-hmC changes at nucleotide-resolution. 5-hmC maintenance has been associated with protecting against CpG island hypermethylation, which commonly occurs in cancer[389–393]. Measures of breast tissue nucleotide-specific 5-hmC revealed enrichment within breast-specific enhancers and transcriptionally active chromatin[394]. In ER/PR-negative breast cancer particularly, loss of 5-hmC is associated with poor prognosis[388]. As DNA methylation alterations occur early in breast carcinogenesis and are related with prognosis[395,396], a better understanding of 5-hmC in breast cancer and EMT is needed.

In concert with DNA methylation, chromatin accessibility regulates transcription and cell reprogramming[397]. Interactions with different nuclear macromolecules such as transcription factors and histone modifications shape the topology of chromatin[397]. Specific chromatin accessibility states have been implicated in regulating EMT. Putative enhancers, defined by promoter-distal H3K27ac and H3K4me1 histone modifications have been shown to recruit key EMT transcription factors such as NF-κB and AP-1 in epithelial cells in comparison to TGF-β-treated mesenchymal cells[398–400]. In addition, motifs of key EMT transcription factors (AP-1, ETS) were enriched in accessible chromatin regions of TGF-β transformed mammary epithelial cells[401]. Although transcription factors influencing EMT and metastasis-associated chromatin accessibility have been identified[402–405], gaps in knowledge of chromatin accessibility changes in non-TGF-β-induced EMT cells and cells in EMT intermediate/hybrid states still remain due to challenges in isolating cells in these states. Moreover, better understanding of the relationship between cytosine modifications and chromatin conformation is needed.

Here, we provide a nucleotide-resolution genome-scale map of cytosine modifications and chromatin accessibility for phenotypes spanning the EMT spectrum. We address gaps in understanding of epigenomic changes in the intermediate/hybrid states on the EMT spectrum. Using a novel model derived from estrogen receptor/progesterone receptor negative (ER/PR-negative) breast cancer cells to study terminal and intermediate EMT states, we demonstrate substantial differences in the cytosine modifications profiles of cells in intermediate EMT states; particularly, increases in 5-hmC enriched in key EMT transcription factor motifs. Further, we utilize novel, integrative multicomponent epigenetic analysis to show cytosine modifications coordinate with chromatin accessibility especially at promoters to regulate transcription.

2.3. Methods

Cell culture

Single cell clones, methods of which isolation and characterization are detailed in Brown et al[363] were used. To summarize, six single cell clones were isolated from SUM149PT cells to represent different points of the EMT spectrum. Position on the EMT spectrum was determined by cell morphology, flow cytometry analysis of CD44 and CD104 markers, and mRNA expressions of *ZEB1/2*. Graphic representation of each clones' position on the EMT spectrum can be found in **Figure 2-1.** Transwell assays to measure migration and invasion were conducted and reported for each clone in Brown et al[363].

DNA methylation and hydroxymethylation

DNA conversion and methylation/hydroxymethylation profiling

DNA from each clone of similar passage numbers was extracted using DNeasy Blood and Tissue kit (Catalog ID 69504, Qiagen, Hilden, Germany). DNA was quantified with Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA). ~2µg of DNA underwent oxidative-bisulfite conversion to measure both 5-mC and 5-hmC using the TrueMethyl OxBS Module (Catalog ID 0414-32; Nugen, Redwood City, CA). Epigenome-wide DNA methylation profiling was performed using the Infinium MethylationEPIC Bead Chips (Illumina Inc., San Diego, CA) at the Norris Cotton Cancer Center Genomics Shared Resource Core.

Quality control and processing

Raw intensity files produced from the MethylationEPIC Bead Chips were preprocessed using the *minfi* R/Bioconductor analysis pipeline (v1.34.0) annotation file version *ilm10b4.hg19*[406,407]. 695 technical probes and 33,360 SNP associated probes were excluded. Quality control was performed using *ENmix* R package[408]. 301,580 probes that failed to meet a detection p-value of 0.00005 in > 30% of the samples and 5% of the CpGs were excluded. High number of CpGs that failed to pass the quality control may have been due to 1) oxidation further damaging the DNA on top of the bisulfite treatment and 2) signal distributions being distorted from the oxidation measurement as the quality control measures were developed for bisulfite converted DNA. After these exclusions, 545,515 CpGs remained for analysis. The filtered data was then normalized using *preprocessFunnorm* in *minfi* to remove unwanted technical variation.

Annotations of CpGs such as genomic context or relation to CpG Island were provided in the Illumina EPIC B4 manifest and UCSC hg19 reference genome files. “Promoter”, “Intergenic”, “Intron” and “Exon” genomic contexts were defined by finding overlapping genomic regions of the CpGs and each context using the UCSC hg19 reference genome annotation. “DNase hypersensitive site” context was defined by having a record in the “DNase_Hypersensitive_NAME” in the annotation. “Gene body” transcriptional context was defined by having a “Body” in the *UCSC_RefGene_Group*. Likewise, “3’ UTR” and “5’ UTR” regions were defined by having “UTR3” and “UTR5”, respectively, in the *UCSC_RefGene_Group*. Relation to CpG Island were defined by the “Relation_to_UCSC_CpG_Island” in the Illumina EPIC annotation file. If no record of relation to the CpG island was indicated, the CpG was considered to be in the “Open Sea” region. For analysis testing enrichment of CpGs measured on the Illumina EPIC array to ATAC regions, GRCh38 annotation file from Zhou et al was used[409].

CpGs annotated to open chromatin regions were defined by their overlap with open chromatin regions from ATAC-seq data. CpGs were determined to be in enhancers if they were located in distal intergenic regions (within 10 – 15kbps upstream and downstream of gene) of the ATAC-seq consensus peaks. CpGs were determined to be in open promoters if they were located in promoters of the ATAC-seq consensus peaks.

5-hmC estimation

5-hmC beta values were estimated using the *fitOxBS* function in the *OxyBS* package[184]. Instead of naive subtraction of signals from oxidative-bisulfite treated probes from bisulfite only treated probes, the *OxyBS* package uses maximum likelihood estimation of the signal intensities from the oxidative-bisulfite treated and bisulfite treated DNA from the Illumina EPIC array to determine the parameters for unmethylated, hydroxymethylated and methylated CpGs.

Analysis

Principal component analyses were performed using 5-hmC and 5-mC beta values using *princomp* function in R. Differential methylation and hydroxymethylated analyses were conducted using *limma* (v3.44.3) and *qvalue* (v2.20.0) R packages in R (v4.0.2)[410,411]. Differentially methylated and hydroxymethylated CpGs were identified by fitting into a linear regression model, testing for differences in beta values CpG-by-CpG in groups of clones based position on the EMT spectrum (distal vs intermediate). Linear regression models were fit by using *lmFIt* and *eBayes* functions. E, EM1, M2, and P were considered as distal clones. Intermediate group was comprised of EM2, EM3, M1 clones. The differentially methylated CpGs were deemed to be significant at the q-value threshold of 0.01.

Differentially hydroxymethylated and methylated CpGs were compared to the 545,515 CpGs used in analyses to test for enrichment at specific genomic contexts using Fisher's exact test. Functional significance of these CpGs were assessed using the Genomic Regions Enrichment of Annotations Tool (GREAT)[412].

ATAC-seq

ATAC-seq and preprocessing

ATAC-seq for 2 replicates per clone was performed as described in Buenrostro et al[413]. Similar passage number (+/- 1 passage) of the clones as for the DNA methylation and hydroxymethylation measurements were used. Same processing methods and detailed descriptions can be found in Brown et al[363].

Briefly, ATAC-seq data was then processed using the publicly available ENCODE ATAC-seq pipeline (https://www.encodeproject.org/pipelines/ENCPL792NWO/). Illumina adapter and transposase sequences were trimmed using *Cutadapt*[414] (v1.9.1) with parameters "--minimum-length 5 -e 0.1". Trimmed reads were aligned to hg38 human genome using *Bowtie2*[415] (v2.2.6) in "--local" mode with parameters "-X 2000 -k 2". Duplicate reads were identified and filtered from final alignments using *MarkDuplicates* (*Picard Tools*[416]). To account for insertion of adapter sequences by the transposase, alignments were converted to tagAlign files and shifted +4 bp and -5 bp on the + and – strands, respectively. *MACS2*[417] (v2.1.1) *callpeak* command with parameters "--shift -75 --extsize 150 --nomodel --keep-dup all --call-summits -p 1.0E-10" were used to call peaks. The peaks were filtered against the ENCODE hg38 blacklist. The Irreproducible Discovery

Rate (IDR) method was used to identify a set of reproducible peaks across biological replicates using an IDR threshold of 0.05.

ATAC-seq analysis

Principal component analyses were performed using variance stabilizing transformed ATAC-seq counts using *princomp* function in R. Low level regions were filtered out using *filterByExpr* using *edgeR* (v3.30.3)[418]*.* Open chromatin regions containing dhmCpGs were annotated using *TxDb.Hsapiens.UCSC.hg38.knownGene* R annotation file package and the *annotatePeak* function in *ChIPseeker* (v1.24.0)[419],[420]. Enriched biological pathways associated with the differentially accessible regions were identified using the *ReactomePA* (v1.32.0)[421].

We tested for over-representation of TF binding site motifs of dhmCpGs containing consensus ATAC peaks compared to all ATAC peaks. We scanned these peaks for TF motif occurrences using R-package *motifmatchr*[422]. Position frequency matrices for human TF motifs used as input to motifmatchr were downloaded using R-packages *JASPAR2020*[423] and *TFBSTools*[424]. Over-represented TF motifs in each peak set were identified through hypergeometric testing using the *phyper* R function, with all peaks identified in that clone used as the background set. TF motifs with an FDR-adjusted hypergeometric *P*-value <0.05 were deemed as over-represented.

RNA-seq

RNA extraction and preprocessing

RNA was collected using Qiagen RNeasy plus kit (Catalog ID: 74034, Qiagen, Hilden, Germany) and quantified using a NanoDrop (Thermo Fisher Scientific - ND-2000-US-CAN). Same processing methods and detailed descriptions can be found in Brown et al[363].

To summarize, raw single-end RNA-seq data were trimmed of polyA sequences and low-quality bases using *Cutadapt* (v2.4)[414]. Reads were aligned to human genome hg38 using *STAR* (v 2.7.2b)[425] with parameters “--outSAMattributes NH HI AS NM MD --

outFilterMultimapNmax 10 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1". Quality of alignments was assessed using *CollectRNASeqMetrics* (*Picard Tools*)[416] and duplicate reads were identified (but retained) with *MarkDuplicates* (*Picard Tools).* Gene-level abundance estimates were generated using *RSEM* (v1.3.2)[426] using the rsem-calculate-expression command with the parameters "--strandedness reverse --fragment-length-mean 313 --fragment-length-sd 91".

2.4. Results

We utilized a previously derived model of six single-cell clones from SUM149PT, a heterogeneous ER-/PR- inflammatory breast cancer line, that represent cell states present along the EMT spectrum. The EMT state of each clone was determined by cell morphology, flow cytometry for CD44 and CD104 markers, and immunofluorescence staining for Vimentin/E-cadherin, as well gene expression of canonical EMT markers (*SNAI1*, *ZEB1*, *CDH1*, *VIM*, and others), detailed in previous work[363]. More epithelial-like clones had low CD44 and high CD104 expression, while more mesenchymal-like clones had high CD44 and low CD104 expression. Intermediate clones had high CD44 and high CD104 expression. *VIM* and *ZEB1/2* increased in expression along with progressive position on the epithelial to mesenchymal transition spectrum, while *CDH1* and *OVOL1/2* decreased in expression (**Figure 2-1**, gene expression data in Brown et al)[363]. These clones were ranked as epithelial (E), three distinct intermediates (EM1, EM2, and EM3), two unique mesenchymal-like clones (M1 and M2), and compared here with the parental cell line (P). Phenotypically, the intermediate clones (EM1, EM2, EM3) displayed higher migratory and invasive behavior, and higher tumor initiation and metastasis formation potential compared to the clones on the either edges of the EMT spectrum (E, M1, M2) (**Figure 2-1**, specific data for each clone reported in Brown et al)[363].

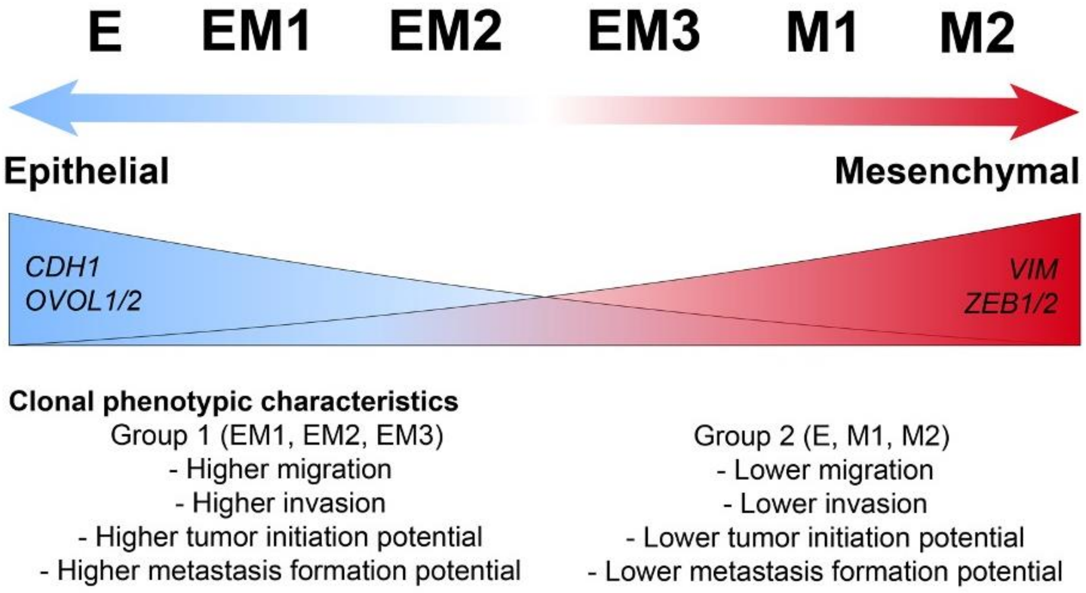

E
EM1
EM2
EM3
M1
M2
Epithelial
Mesenchymal
CDH1
OVOL1/2
VIM
ZEB1/2
Clonal phenotypic characteristics
Group 1 (EM1, EM2, EM3)
- Higher migration
- Higher invasion
- Higher tumor initiation potential
- Higher metastasis formation potential
Group 2 (E, M1, M2)
- Lower migration
- Lower invasion
- Lower tumor initiation potential
- Lower metastasis formation potential

Figure 2-1. Summary of characteristics of isolated single cell clones that reside in specific epithelial-to-mesenchymal transition spectrum.
Specific data for each clone that are summarized in this figure are reported in Brown et al[363]. Gene expression of epithelial markers (*CDH1* and *OVOL1/2*) are highest on the most epithelial-like clone and decreases sequentially as clones display more mesenchymal characteristics. Gene expression of mesenchymal markers (*VIM* and *ZEB1/2*) are lowest in the most epithelial-like clone and increases sequentially as clones display more mesenchymal characteristics.

We first measured genome-scale cytosine-specific DNA methylation (5-mC) and hydroxymethylation (5-hmC) levels, using the Illumina EPIC methylation array. As expected[373], a relatively small subset of measured CpGs were hydroxymethylated, with average 5-hmC beta values much smaller than that of 5-mC across all clones (**Figure 2-2A, 2-2B**). Average 5-hmC beta values and 5-mC beta values were negatively correlated, at marginal significance (*R* = -0.72, p= 0.071) with increased global 5-hmC and decreased global 5-mC abundance in intermediate clones (EM2, EM3, M1; **Figure 2-2B, 2-2C**).

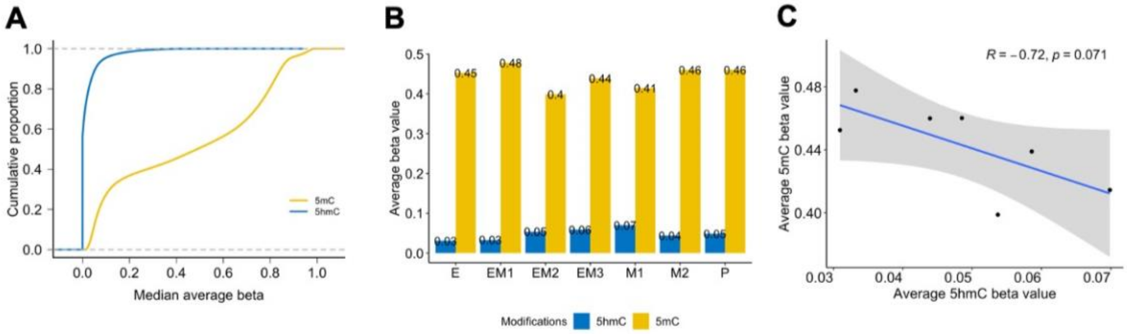


Figure 2-2. 5-hmC and 5-mC levels in the EMT clonal cell line model.
A) Cumulative density of median 5-hmC and 5-mC beta values. **B)** Average 5-hmC and 5-mC beta values per clone. **C)** Pearson correlation of 5-hmC beta values and 5-mC beta values.

To identify which distal clone the clones along the EMT spectrum were similar, we compared each 5-hmC profile of EM1, EM2, EM3, M1 to the 5-hmC and 5-mC profile of clones on the extreme ends on the EMT spectrum (E and M2). 5-hmC profiles of EM1, EM2, EM3 had more similar 5-hmC profiles to the 5-hmC profile of E, in which the number of CpGs with little to no change were higher in E compared to in M2 (**Figure 2-3A**). The 5-hmC profile of M2 had very similar number of CpGs with little to no change in comparison to 5-hmC profiles of E and M2. The 5-hmC profiles of EM1, EM2, EM3, and M1 were all more similar to 5-mC profile of E rather to 5-mC profile of M2 (**Figure 2-3B**). Our results suggest that EM1, EM2, EM3, M1 clones likely were derived from the most epithelial clone, and provide models of states on the epithelial-to-mesenchymal transition.

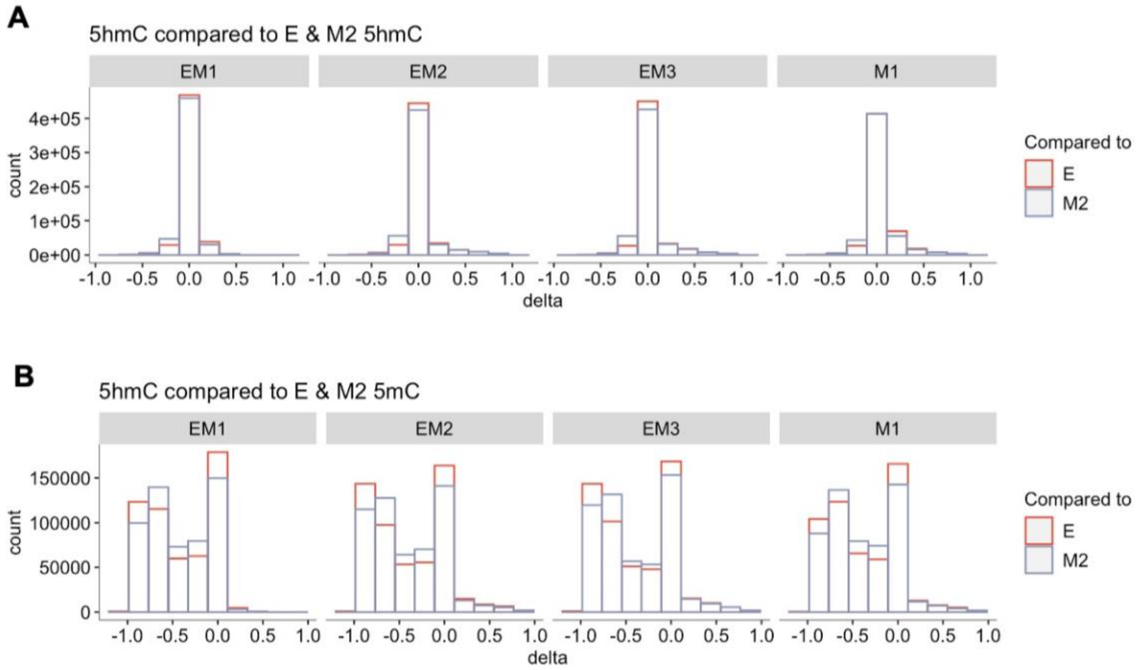


Figure 2-3. Clones in between the most extremes of the EMT spectrum are more similar to the most epithelial clone.
A) Delta change in 5-hmC in EM1, EM2, EM3, M1 compared to 5-hmC in E and M2. Comparison to E indicated with red boxes. Comparison to M2 indicated with blue boxes. **B)** Delta change in 5-hmC in EM1, EM2, EM3, M1 compared to 5-mC in E and M2. Comparison to E indicated with red boxes. Comparison to M2 indicated with blue boxes.

Genome-wide DNA cytosine modification profiles in EMT clones

To determine associations between EMT phenotypes (migratory and invasive behavior) of clones and DNA cytosine modifications, first, we analyzed correlations between global 5-mC and 5-hmC beta values with average migration and invasion levels that had previously been determined in Brown et al[363]. There were no statistically significant correlations between global DNA cytosine modification levels and migration and invasion levels (**Supplementary Figure 2-1A – 2-1D**).

In addition to correlations between global levels of DNA cytosine modifications, we conducted epigenome wide association study to identify specific CpGs that are associated with high migration and invasive properties. Migration and invasion assays shown from Brown et al indicated that clones (EM1, EM2, EM3, P) with greater than the median migration and invasion levels were determined to have high migratory and invasive properties (**Supplementary Figure 2-2A**)[363]. While it was surprising that the EM1, EM2, EM3 clones were more migratory and invasive than the mesenchymal clones, previously established traits of mesenchymal cells did not discern between mesenchymal and intermediate states when determining migratory and invasive behavior. It is possible that because these cell states were not distinguished, the migratory and invasive behavior of the intermediate clones influenced the notion that mesenchymal cells were more likely to be migratory[427]. Only one differentially hydroxymethylated CpG were determined to be associated with high migratory and invasive cellular phenotypes under the FDR < 0.1 significance level (**Supplementary Figure 2-2B**). There were no differentially methylated CpGs associated with high migratory and invasive cellular phenotypes under the FDR < 0.1 significance level (**Supplementary Figure 2-2C**).

To compare genome-scale similarity of DNA methylation profiles among all clones, we compared the 5-hmC and 5-mC beta values using principal component analysis (PCA). PCA results indicated that 5-hmC and 5-mC beta values clustered into two distinct groups: one group of E, EM1, M2 and another group with EM2, EM3, M1 (**Figure 2-4A, 2-4B**). In downstream analyses for this study, EM2, EM3, M1 were defined as intermediate clones and E, EM1, M2, P were defined as distal clones. These two groups were slightly different from groupings identified by the clones' cellular phenotypes summarized in **Figure 2-1** and in the original development of the model. Furthermore, the groups identified by genome-scale 5-mC and 5-hmC beta values were different than PCA clustering from chromatin accessibility profiles from ATAC-seq (**Supplementary Figure 2-3A**) and gene expression profiles from RNA-seq (**Supplementary Figure 2-3B**). Non-negative matrix factorization hierarchical clustering with 5-mC, 5-hmC, and chromatin accessibility profiles revealed similar clustering results from RNA-seq and ATAC-seq **(Supplementary Figure 2-3C**). Following the PCA results, distinct grouping of clones into intermediate and distal was supported by unsupervised hierarchical clustering of the top 5% most variable CpGs (27,276 CpGs) which were chosen based on distribution of variances across CpGs (**Supplementary Figure 2-4A, 2-4B**). Unsupervised clustering identified highly distinct intermediate and distal clone clusters (**Figure 2-4B, 2-4C**) and highlighted the greater relative abundance of 5-hmC in intermediate clones compared to distal clones (**Figure 2-2B**) at the CpG-specific level.

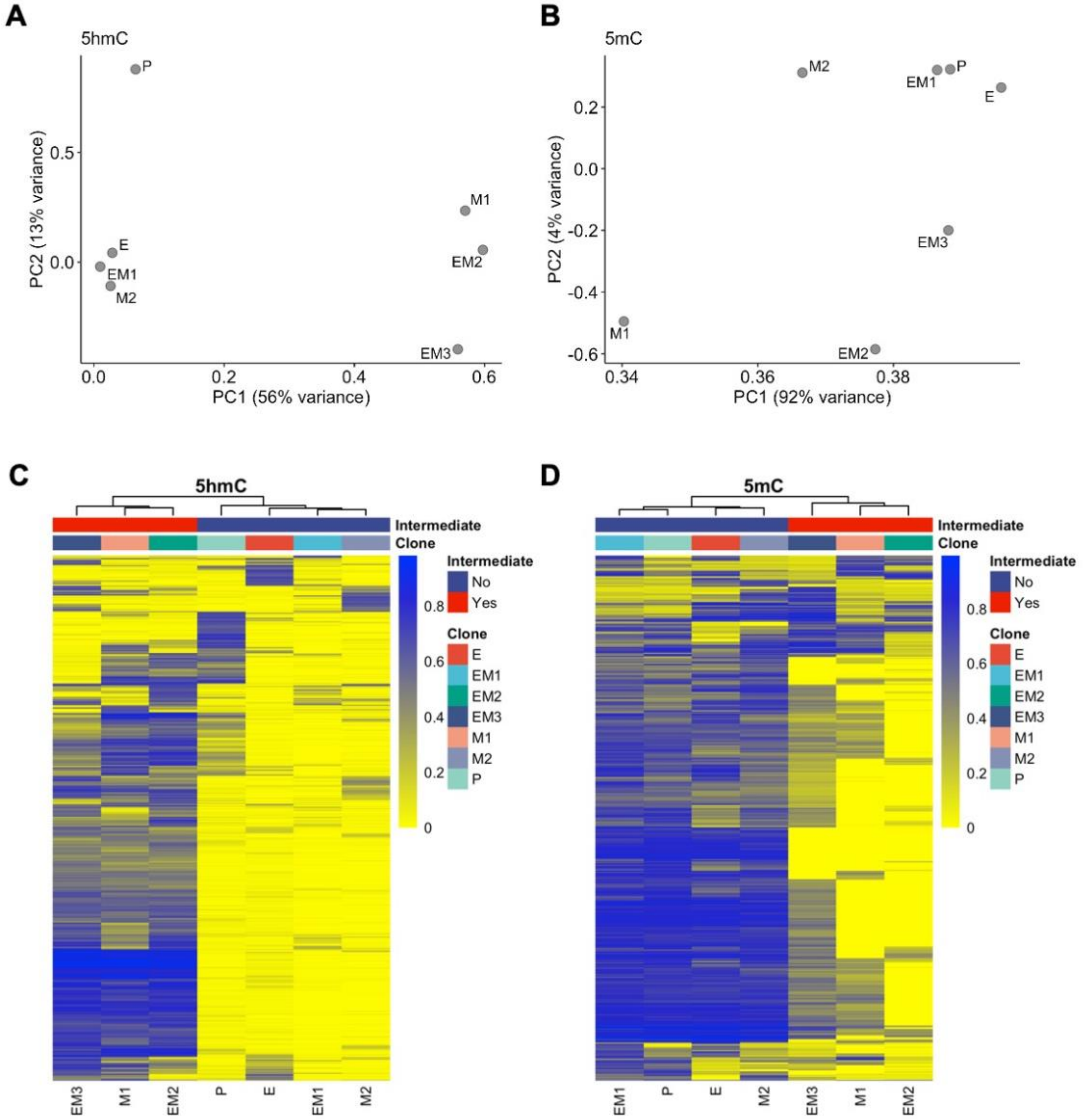


Figure 2-4. Distal and intermediate clones have distinct methylation and hydroxymethylation profiles.
Results from principal component analysis of **A)** 5-hmC and **B)** 5-mC beta values. Heatmap of unsupervised clustering of the top 5% (27,276 CpGs) most variable **C)** 5-hmC and **D)** 5-mC CpGs. Color scale ranges from yellow (low beta value) to blue (high beta value). Horizontal tracking bars indicate clones and position on the EMT spectrum.

We next used a candidate gene approach to investigate if EMT-associated 5-hmC and 5-mC loci distinguished intermediate from distal clones. We performed unsupervised clustering on beta-values of 439 CpGs annotated to epithelial genes (*CDH1, CLDN1, EPCAM, ITGAB4, KRT8* and *OCLN)*, mesenchymal genes (*CDH2, FN1, ITGB1, MMP19, MMP2,* and *VIM*), and EMT-related transcription factors (*SNAI1, SNAI2, TWIST1, ZEB1* and *ZEB2*). Intermediate clones clustered separately from distal clones for both 5-mC- and 5-hmC-associated genes, and a subset of CpGs annotated predominantly to epithelial genes (*OCLN, CDH1, KRT8, EPCAM*) had high 5-hmC among intermediate clones in cluster #4 (**Figure 2-5A, 2-5B**) many of which tracked to promoter regions (**Figure 2-5C**). Together, it suggests potential role of 5-hmC in regulating epithelial genes during the EMT process.

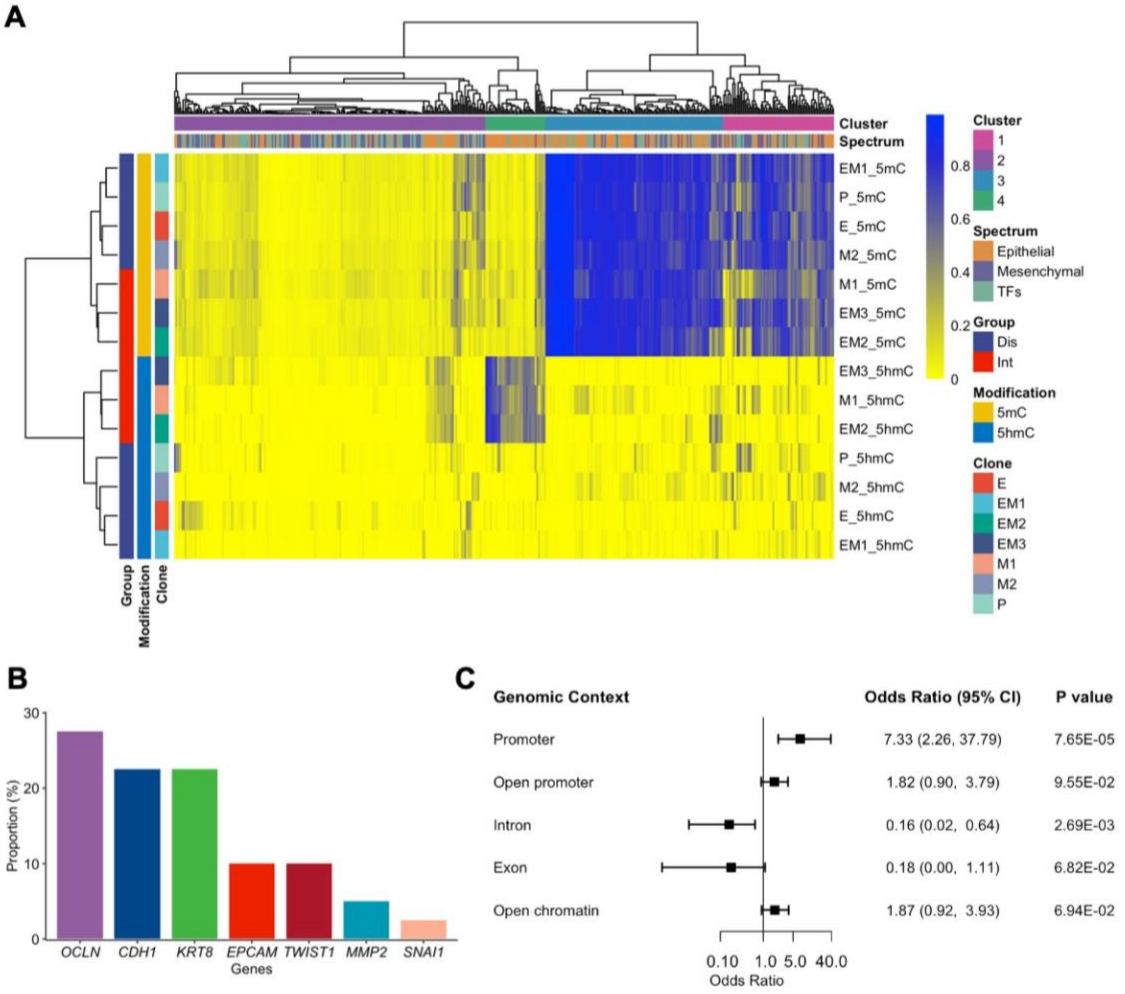


Figure 2-5. Intermediate clones have higher hydroxymethylation among epithelial genes.
A) Heatmap of unsupervised clustering of 5-hmC and 5-mC in a set of 231 CpGs within epithelial genes (*CDH1, CLDN1, EPCAM, ITGB4, KRT8,* and *OLCN*), mesenchymal genes (*CDH2, FN1, ITGB1, MMP19, MMP2,* and *VIM*), and transcription factors (*SNAI1, SNAI2, TWIST1, ZEB1,* and *ZEB2*). Vertical tracking bars indicate DNA modification, clones, and position on the EMT spectrum. Horizontal tracking bars indicate EMT marker group (epithelial genes, mesenchymal genes, transcription factors) and hierarchical clustering group from when height = 1.6. **B)** Proportions of the genes annotated to the 40 CpGs in cluster #4 of the hierarchical clustering from the heatmap. **C)** Enrichment of genomic contexts of the CpGs in hierarchical cluster 4. 40 CpGs in cluster 4 were compared to all 231 CpGs in EMT-related genes using Fisher's test.

To determine if overall 5-hmC and 5-mC abundance was related to expression of cytosine modifying enzymes (DNMTs and TETs), we leveraged RNA-seq to test the correlation of average methylation and gene expression levels. Only *TET1* gene expression was significantly positively correlated with global average 5-hmC beta values (*R* = 0.86, p = 0.024), and none were correlated with 5-mC (**Supplementary Figure 2-5A, 2-5B**). 5-hmC and 5-mC beta values of DNMT and TET CpGs with unsupervised clustering did not identify extensive variation in cytosine states at cytosine modification enzyme genes (**Supplementary Figure 2-5C**). However, a small subset of CpGs (n CpGs = 18 of 241 total), located within TETs (*TET1* = 33%, *TET2* = 28%, *TET3* = 17%), exhibited higher 5-hmC in intermediate clones (**Supplementary Figure 2-5C**).

Together, these findings suggest there are variable patterns of genome-wide 5-hmC and 5-mC based on clonal EMT status, not with clonal phenotypes.

Differential methylation and hydroxymethylation in intermediate clones

Next, we conducted an epigenome wide association study (EWAS) comparing cytosine modifications at the nucleotide level to identify differential cytosine modifications between intermediate and distal clones. Overall, we identified 17,862 significantly differentially hydroxymethylated CpGs (dhmCpG, FDR < 0.01), between distal and intermediate clones, almost all of which had increased in 5-hmC in the intermediate clones (**Figure 2-6A, Supplementary Table 2-1**), including EMT associated genes such as *SNAI1* and *TWIST1*. There were 7,903 significantly differentially methylated CpGs (dmCpG, FDR < 0.01), most of which had decreased in 5-mC in intermediate clones (**Figure 2-6B, Supplementary Table 2-2**), including EMT associated cell type markers *CDH1* and *MMP19*. For further downstream analyses, dhmCpGs were subset for only CpGs increasing in 5-hmC. dmCpGs were subset for only CpGs decreasing in 5-mC. Among CpGs with increased 5-hmC and decreased 5-mC, only 33 CpGs overlapped (**Figure 2-6C**). Expanding to the gene-level, 1,365 genes had both dhmCpGs and dmCpGs among intermediate clones (**Figure 2-6D**). Genomic contexts with enrichment of dhmCpGs were generally depleted among dmCpGs (**Figure 2-6E; Supplementary Table 2-3**). While dhmCpGs were enriched in regulatory regions (open chromatin regions, enhancers, 5’UTR, promoters, TSS1500, TSS200) and in the first exon, dmCpGs were enriched within exons and introns, suggesting different cytosine modifications act on different genomic regions in regulating the EMT process. Our results suggest that while some differential cytosine modification mark may act on the same gene, generally, the two DNA cytosine modification marks act on different regions of the genome to coordinate EMT processes.

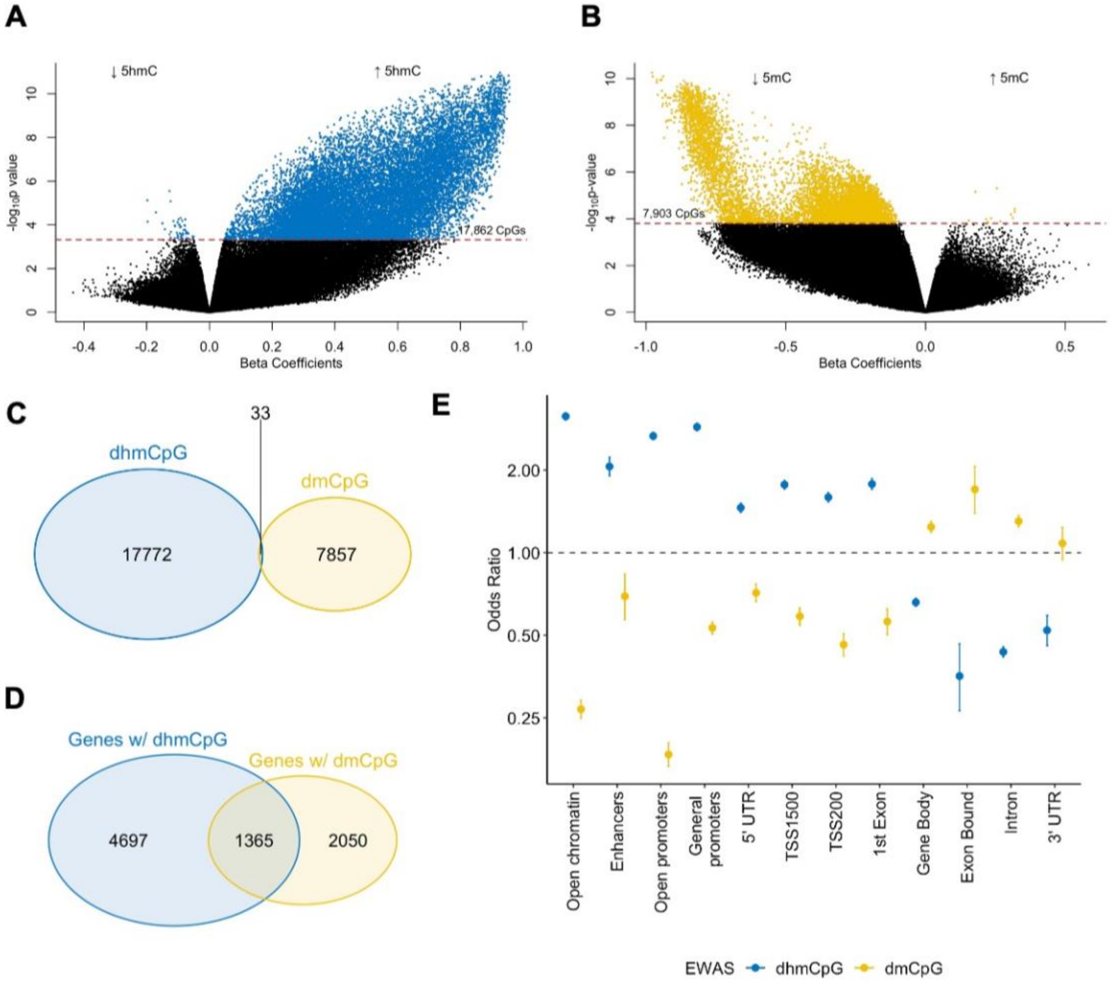


Figure 2-6. Differential 5-hmC CpGs are distinct from the differential 5-mC CpGs.
Volcano plots indicating **A)** 17,862 significantly differentially hydroxymethylated CpGs and **B)** 7,903 significantly differentially methylated CpGs under FDR Q-value of 0.01, in intermediate clones in comparison to distal clones. Red dashed lines indicate the -log10(p-value) at FDR q-value of 0.01. Venn diagrams comparing **C)** dhmCpGs vs dmCpGs and **D)** genes annotated to dhmCpGs vs genes annotated to dmCpGs. dhmCpGs were subset for only CpGs increasing in 5-hmC. dmCpGs were subset for only CpGs decreasing in 5-mC. **E)** Enrichment of dhmCpGs and

dmCpGs at different genomic contexts. Odds ratios calculated by Fisher's exact test. dhmCpGs enrichment indicated in blue. dmCpGs enrichment indicated in yellow.

Genomic Regions Enrichment of Annotations Tool (GREAT) analyses revealed that dhmCpGs were associated with fatty acid-related molecular functions (MF), such as peroxisomal fatty-acyl-CoA transporter activity (FE = 19.00) and long-chain fatty acid transporter activity (FE = 7.46), as well as RNA polymerase II transcription factor-related molecular functions such as RNA polymerase II TF sequence-specific DNA binding (FE = 1.18) and RNA polymerase II regulatory region DNA binding (FE = 1.17, **Supplementary Figure 2-6A**). Similarly, dmCpGs were associated with RNA polymerase II-related molecular functions such as RNA polymerase II transcription coactivator binding (FE = 7.62) and cofactor binding (FE = 6.95, **Supplementary Figure 2-6B**). Additionally, dmCpGs were associated with metal ion transmembrane activity (FE = 1.44). Collectively, these results support the role of differential cytosine modifications in RNA polymerase II related regulation of transcription to influence intermediate EMT phenotype.

Potential roles of 5-hmC in regulating epithelial to mesenchymal transition

As increased hydroxymethylation and decreased methylation is traditionally associated with increased gene expression, we wanted to determine whether the dhmCpGs and dmCpGs were acting in regions of open chromatin as identified by ATAC-seq. Out of 42,510 open chromatin regions containing a CpG that was measured on the Illumina EPIC array, 12.03% of the open chromatin regions contained dhmCpGs in contrast to 1.59% of the open chromatin regions containing dmCpGs (**Figure 2-7A**). Interestingly, the only pathways significantly associated with the open chromatin regions containing dhmCpGs were related to Rho family of GTPase, which have been extensively shown to function as cellular switches in coordinating cell polarity and migration by regulating the cytoskeleton (**Figure 2-7B**)[428]. Expression of majority of genes in the RHO GTPase cycle pathway is high in EM1, EM2, EM3 clones (**Figure 2-7C**).

To identify additional molecular processes dhmCpGs in open chromatin regions may regulate, we conducted transcription factor motif enrichment analysis. Motif enrichment analysis found 571 transcription factors (TF) significantly associated with open chromatin regions with dhmCpGs in intermediate clones compared to only 4 TFs in distal clones under the FDR < 0.05 threshold (**Figure 2-7D, Supplementary Table 2-4**). In the intermediate clones, motifs for key EMT transcription factors (ZEB1 and SNAI2) were enriched among open chromatin regions with dhmCpG, implicating 5-hmC in EMT process-associated gene regulation. In addition, motifs for GRHL2, a suggested EMT pioneer transcription factor that has been shown to be associated with epigenetic remodeling, also were enriched in consensus open chromatin regions with dhmCpGs, but not in consensus open chromatin regions with dmCpGs (**Supplementary Table 2-4**)[429,430]. While not known specifically to play roles in EMT, other TF motifs, particularly motifs of GATA2 and SPI1, were also found to be in open chromatin regions with dhmCpGs. Together, these results suggest increase in 5-hmC may play a regulatory role in the epithelial to mesenchymal transition process by acting in Rho GTPase associated genes and acting on binding sites of EMT associated transcription factors.

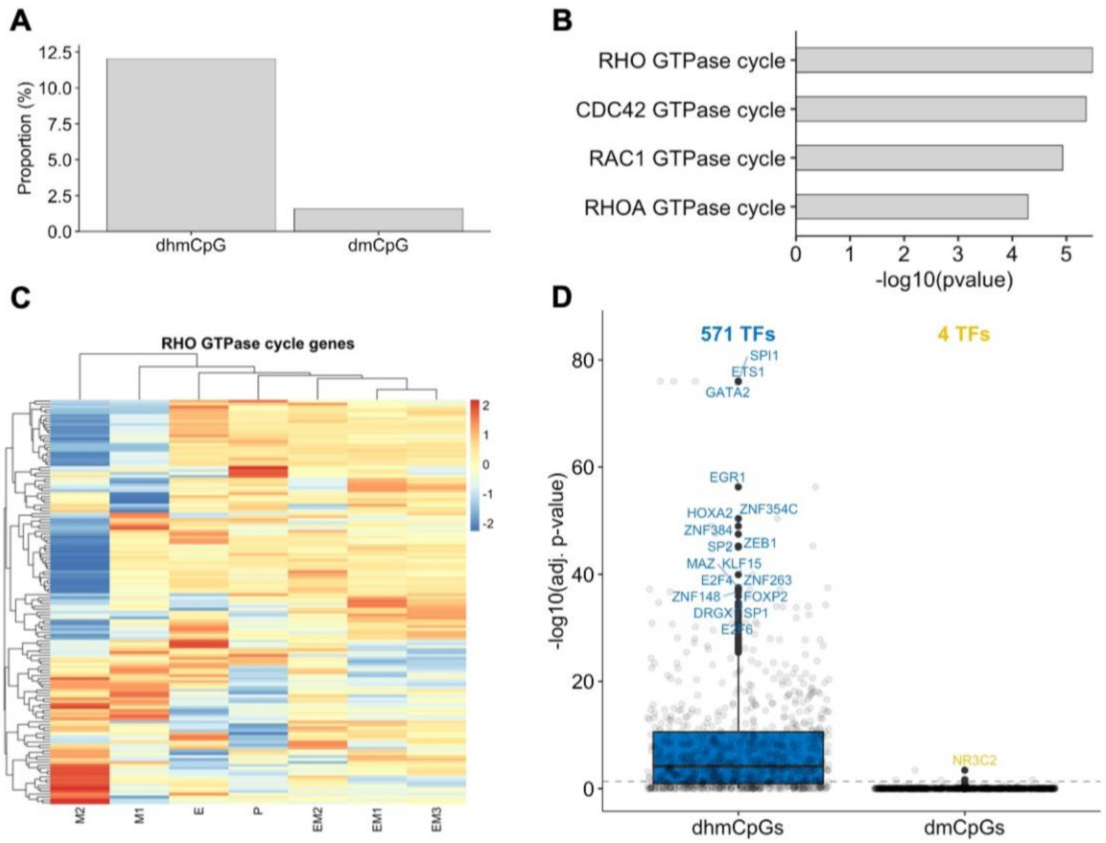


Figure 2-7. dhmCpGs in open chromatin regions are associated with Rho GTPase family and EMT-specific transcription factor motifs.
A) Proportion of open chromatin regions with dhmCpGs and dmCpGs in open chromatin regions containing CpGs analyzed from the Illumina Methylation EPIC array. **B)** Reactome pathways associated with open chromatin regions containing dhmCpGs. **C)** Gene expression z-scores of genes in the RHO GTPase reactome pathway for each clone. Red indicates high expression. Blude indicates low expression. **D)** Transcription factor motifs associated with open chromatin regions containing dhmCpGs and dmCpGs.

2.5. Discussion

Widely used standard bisulfite conversion used to study DNA methylation is unable to distinguish between 5-mC and 5-hmC. Using a tandem oxidative-bisulfite treatment approach, we measured both cytosine modifications to understand their unique distribution across distal and intermediate EMT states. The majority of previous studies measuring 5-hmC have been limited to global 5-hmC levels in tissues of heterogeneous cell types including tumors, where extremely low levels of 5-hmC were observed[158,231,387]. Here, identifying differences in cell state-specific, nucleotide-specific 5-hmC is a strength of our approach. The intermediate clones in our EMT model system suggests that genome-wide patterns of hydroxymethylation are associated with specific EMT phenotypes, suggesting a potential role of 5-hmC in mediating EMT related processes. Moreover, through multi-component approach of epigenome profiling, we show that EMT phenotypes are underscored by substantial epigenetic differences.

Previous work establishing this model system has demonstrated that the intermediate clones represent a population of tumor cells with high migratory and invasive properties. We identify open chromatin regions with dhmCpGs are particularly associated with Rho family of GTPases, family of GTPases that regulates cell polarity and migration by coordinating the cytoskeleton[428]. Rho GTPases have been well documented to play a role in epithelial to mesenchymal transition in tumors[431]. While Rho GTPases have been implicated in tumor progression, mutations in Rho proteins are not common and do not favor initiation or progression of tumors which have called for study of other mechanisms of deregulation Rho proteins[432]. Our study suggests that increasing 5-hmC may be implicated in the epithelial to mesenchymal transition which in turn may contribute to deregulation of Rho proteins. In addition, we show dhmCpGs are associated with motifs of key EMT transcription factors which may indicate recruitment of various transcription factors by 5-hmC may be a potential mechanism regulating the intermediate clones' high migratory and invasive potential. Our results suggest that targeting increases in 5-hmC in intermediate cells may impede the maintenance of this state and/or force lineage commitment, effects that could lead to altered metastatic propensity.

Prior literature has already indicated that DNA methylation states change during TGF-β induced EMT[433]. Similarly, our natural (non-TGF-β induced) EMT model suggests that DNA cytosine modifications exhibit altered genome-wide patterns during the EMT process. Our results indicate that these altered patterns may regulate the existence of cells in various EMT states, thereby enabling tumor heterogeneity. Alterations in cytosine modifications and chromatin accessibility towards a less repressive state suggests that the multi-level epigenome is essential in regulating the dynamics of EMT.

Lastly, our study highlights the importance of multicomponent measures of epigenetic states. Utilizing ATAC-seq in combination with 5-mC and 5-hmC methylation array profiles allowed for identification of the significance of the Rho GTPases that was not evident in only DNA cytosine modification analyses. Moreover, combined datasets allowed for identification of potential role of 5-hmC regulating EMT-related transcription factors. However, the array-based approach may not have revealed CpG loci in relevant accessible chromatin, a limitation that may be overcome with a whole genome bisulfite, oxidative-bisulfite sequencing approach. It highlights the complex epigenetic landscape that is required in the EMT process.

2.6. Conclusion

Our study addresses current gaps that exist in understanding of specific cytosine modifications (5-mC and 5-hmC) roles in EMT and their associations with other epigenetic changes. Clones exhibiting intermediate EMT phenotypes had distinct, more open epigenetic states with increased 5-hmC, decreased 5-mC and more accessible chromatin compared to clones exhibiting more distal EMT phenotypes. Open chromatin regions containing CpG loci with increased 5-hmC enriched in motifs of key EMT transcription factors, ZEB1 and SNAI2, indicate likelihood of multi-component epigenetic regulation during EMT. Epigenetic profiles at the cytosine and chromatin level associated with EMT processes that contribute to gene regulation may be targeted to prevent the progression of EMT.

2.7. Future perspectives

Roles of cell state specific epigenomic changes, specifically in multiple DNA cytosine modification marks, in regulating epitheial-to-mesenchymal transition are only just beginning to be identified. Utilizing multiple genome-wide epigenomic assays will improve understanding of how different parts of the epigenome interact to regulate EMT, which may yield new therapeutic targets to prevent EMT. With novel epigenetic targets, therapeutic strategies to prevent cancer progression into metastasis may be developed for clinical use.

2.8. Author contributions

MKL and MSB carried out the experiments. MSB and DRP conceived the original epithelial-to-mesenchymal transition experimental design. MKL conducted analyses with the help from OMW and BCC. BCC supervised the project. All authors discussed the results and contributed to the final version of the manuscript.

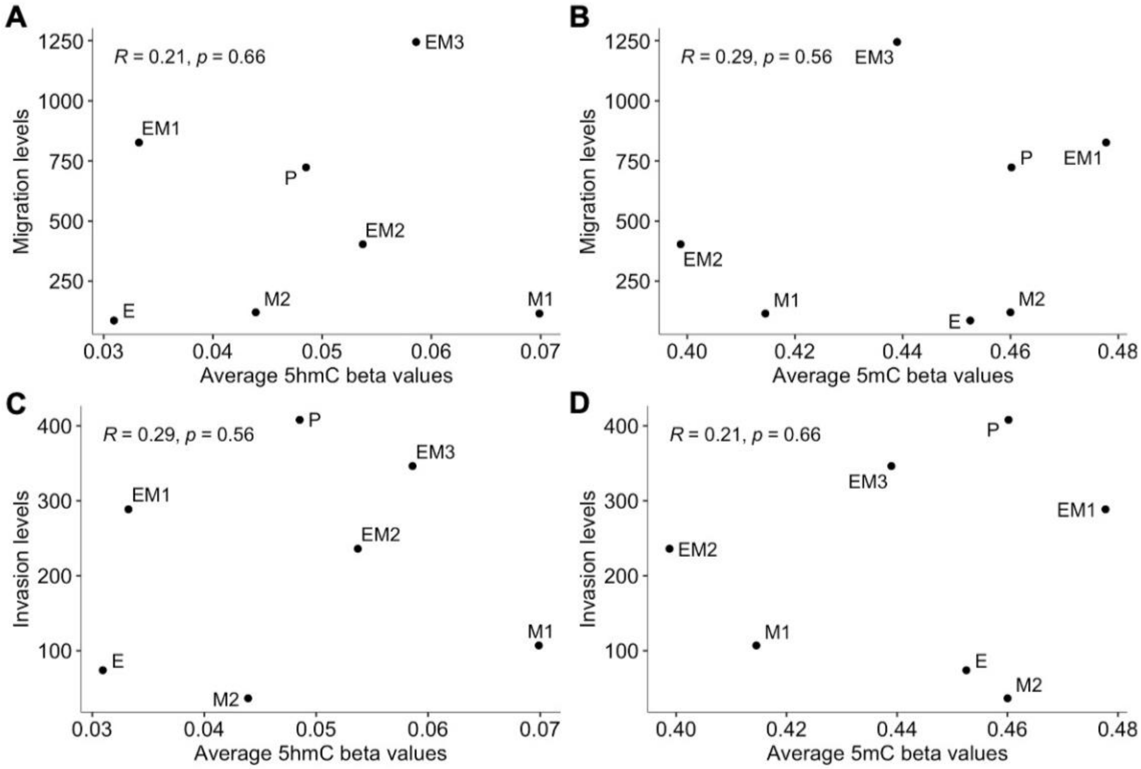
2.9. Acknowledgements

Illumina EPIC arrays, ATAC-seq, and RNA-seq experiments were all carried out at Geisel School of Medicine in the Genomics Shared Resource, which was established by equipment grants from the National Institutes of Health and National Science Foundation and is supported in part by a Cancer Center Core Grant (P30CA023108) from the National Cancer Institute.

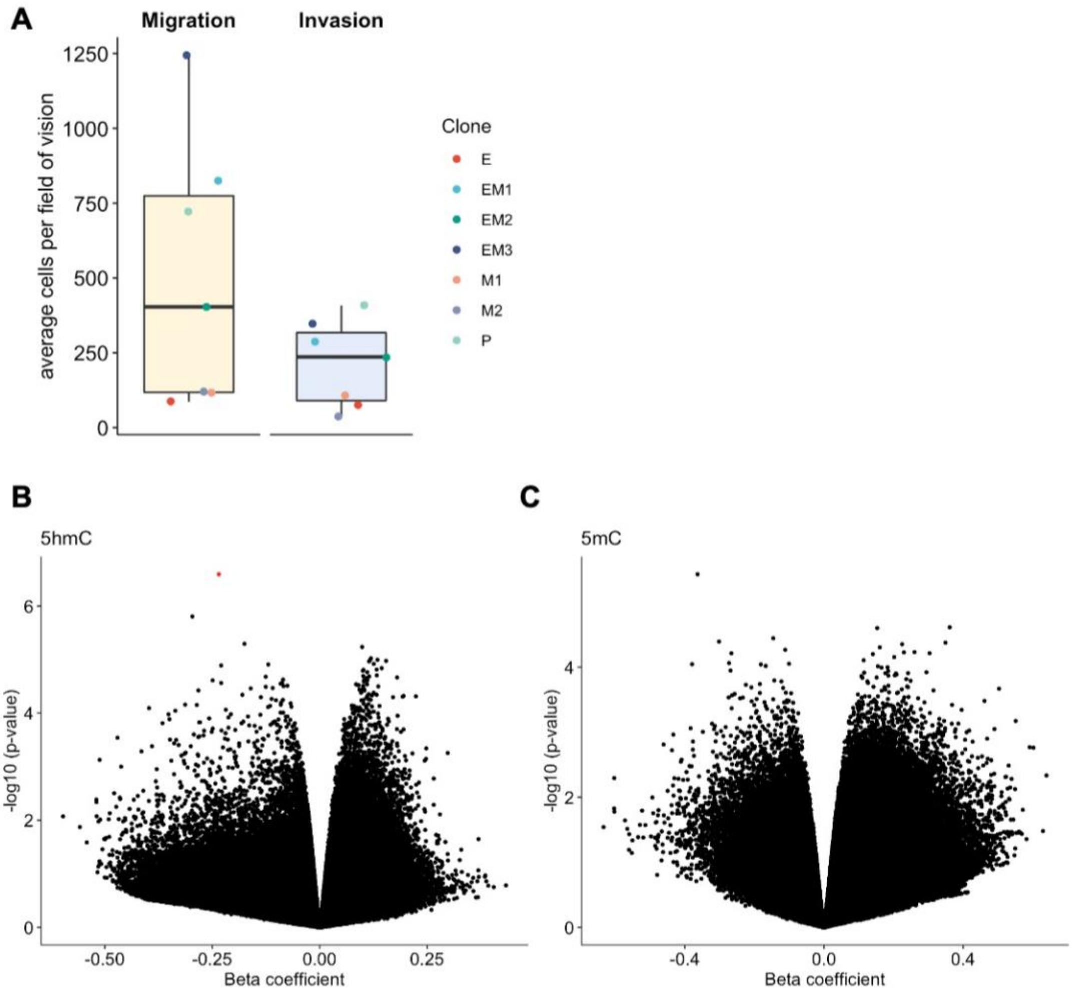
2.10. Funding sources

This work was supported by a Prouty Pilot Grant from Friends of the Norris Cotton Cancer Center, funding from The Elmer R. Pfefferkorn & Allan U. Munck Education and Research Fund at the Geisel School of Medicine at Dartmouth and an NCI Cancer Center Support Grant (5P30CA023108). This work was supported by the National Institues of Health 5R00CA201574 (to D.R.P.), R01CA216265 (to B.C.C), and R01CA253976 (to B.C.C). O.M.W is supported by Cancer Center Core grant (National Cancer Institute 5P30CA023108) and National Institutes of Health 5P20GM130454.

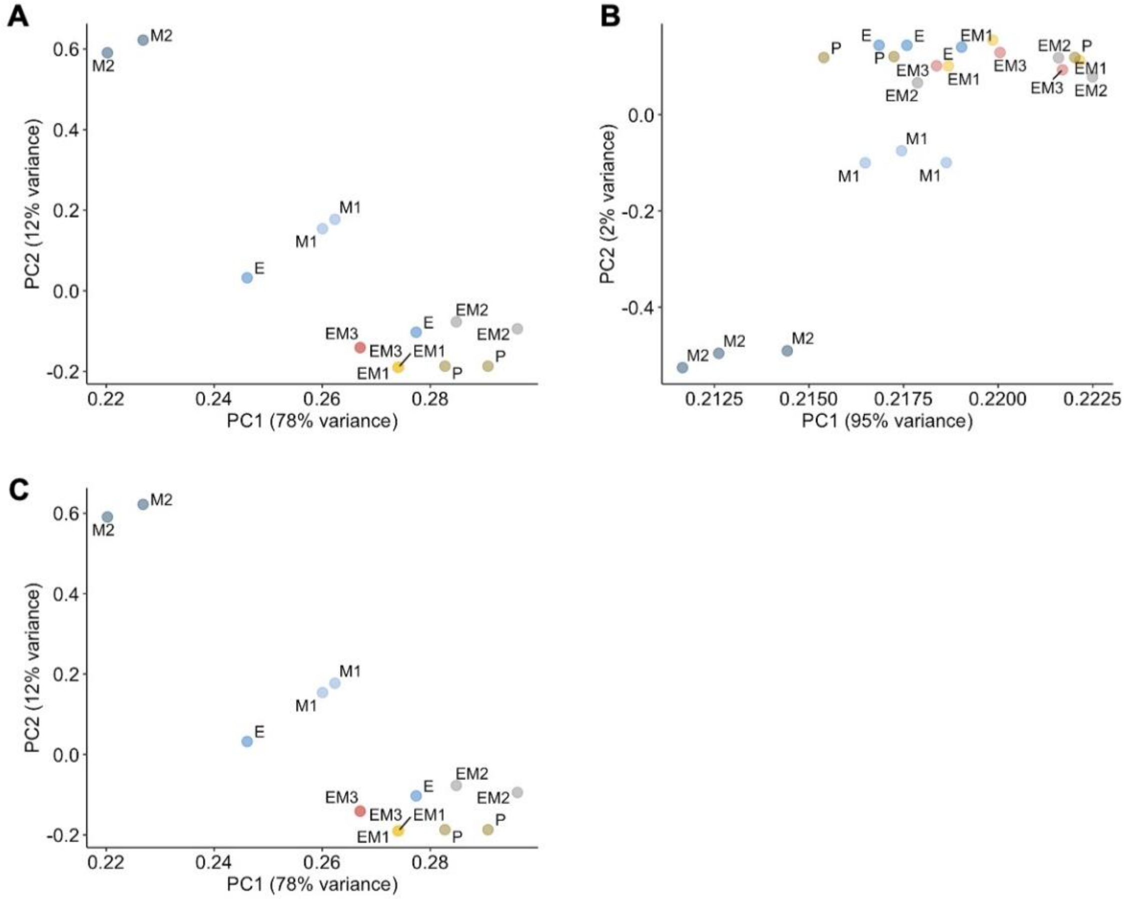
2.11. Supplemental materials



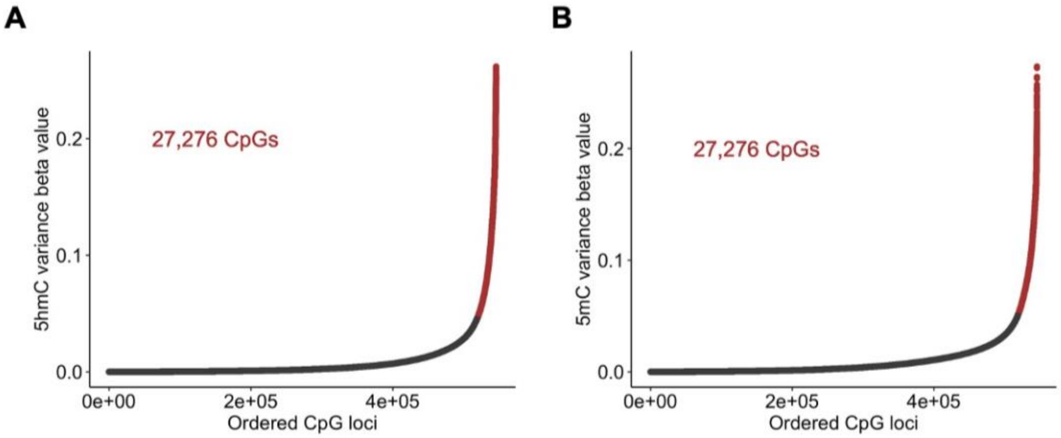
Supplementary Figure 2-1. Correlation between global cytosine modification beta values and clonal phenotypes.
Spearman correlation between migration levels and global **A)** 5-hmC and **B)** 5-mC beta values. Spearman correlation between invasion levels and global **C)** 5-hmC and **D)** bet values. Each point labelled by clone name.



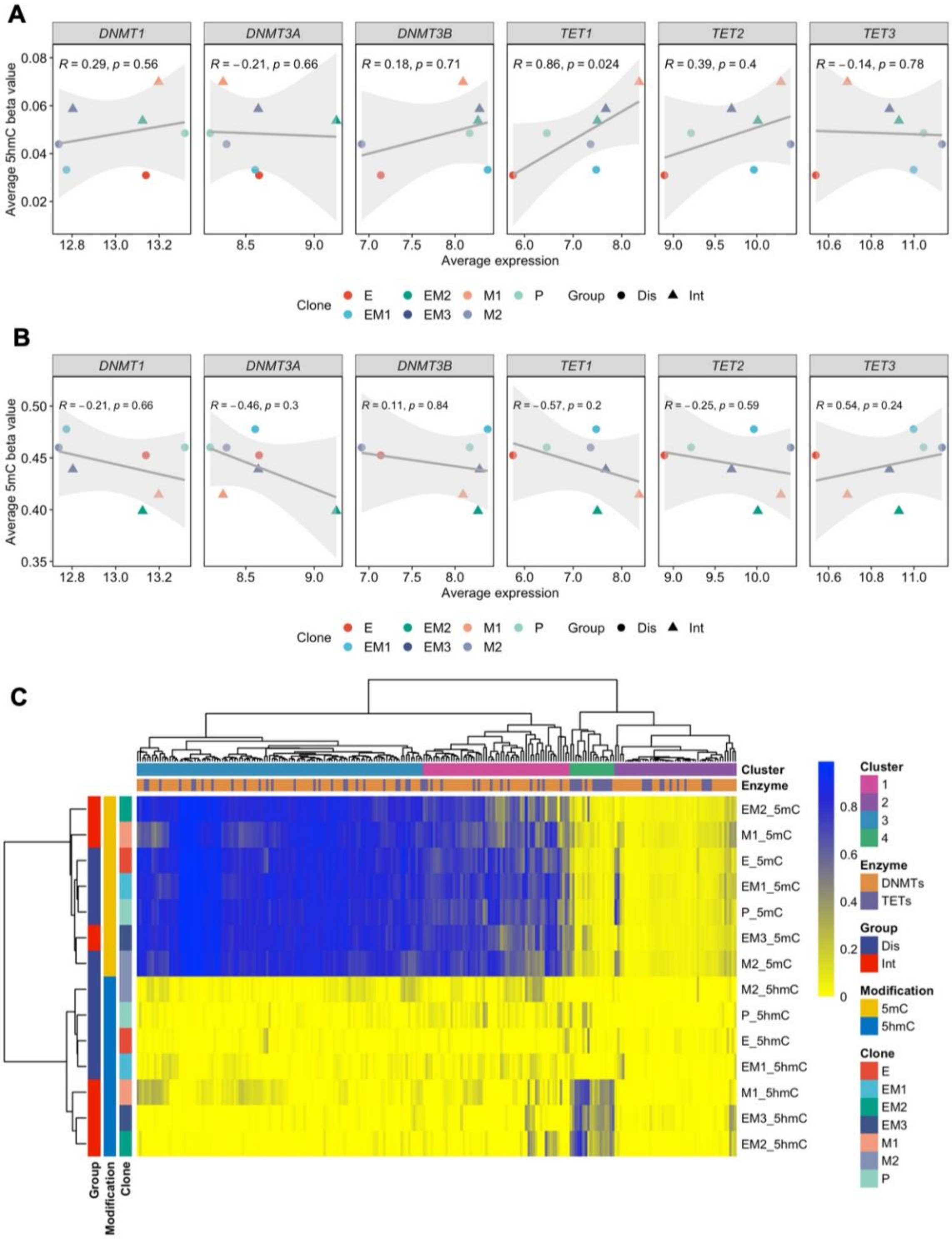
Supplementary Figure 2-2. Results of epigenome wide association studies of DNA cytosine modification marks and high migratory/invasive properties. **A)** Distribution of average cells per field of vision for migratory and invasive properties for each clone from transwell assays. **B)** Volcano plots indicating 1 differentially hydroxymethylated CpG and **C)** no differentially methylated CpG in high migratory/invasive clones compared to low migratory/invasive clones. Differentially hydroxymethylated CpG marked in red.



Supplementary Figure 2-3.
Principal component analysis results of **A)** open chromatin accessibility from ATAC-seq and **B)** gene expression from RNA-seq.

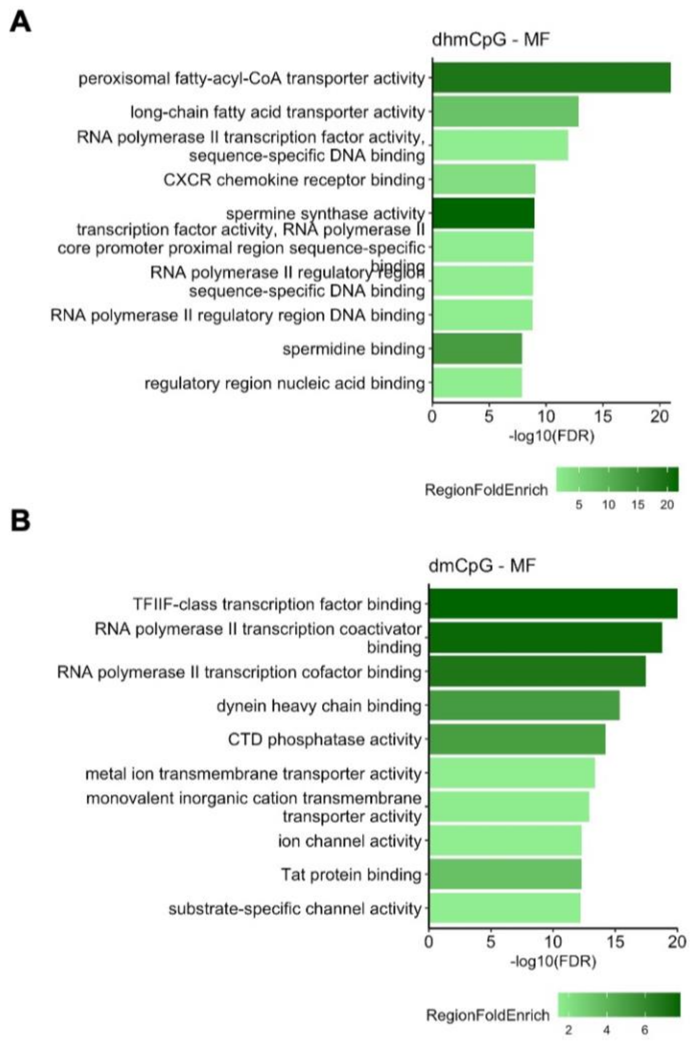


Supplementary Figure 2-4.
Distribution of variance of **A)** 5-hmC and **B)** 5-mC in 545,515 CpGs used in analyses. Red points indicate CpGs with top 5% variance.



Supplementary Figure 2-5. Among DNA methylation modulating enzymes, only TET enzymes indicate subtle differences among clones.
A) Correlation between gene expression levels of DNA methylation and demethylation enzymes and total average 5-hmC beta values. **B)** Correlation between gene expression levels of DNA methylation and demethylation enzymes

and total average 5-mC beta values. Correlation was calculated with Spearman correlation. Shapes of the points in the scatter plot indicate the group, either distal (circle) or intermediate (triangle), of the clone. Colors of the points represent each clone of the EMT spectrum. **C)** Unsupervised clustering of 5-mC and 5-hmC beta values of CpGs located in DNMT and TET genes. Vertical tracking bars indicate DNA modification, clones, and position on the EMT spectrum. Horizontal tracking bars indicate DNA methylation modulating enzyme and hierarchical clustering group from when hierarchical clustering dendrogram height = 1.6.



Supplementary Figure 2-6.
Molecular functions associated with **A)** dhmCpGs and **B)** dmCpGs from Genomic Regions Enrichment Annotations Tool (GREAT) analyses.

Chapter 3

3. Tumor type and cell type-specific gene expression alterations in diverse pediatric central nervous system tumors identified using single nuclei RNA-seq

The following authors contributed to the work:

Min Kyung Lee, Nasim Azizgolshani, Joshua A. Shapiro, Lananh N. Nguyen,
Fred W. Kolling, George J. Zanazzi, Hildreth Robert Frost, Brock C. Christensen

*Only subsets of supplementary tables were included due to file size.

3.1. Abstract

Central nervous system (CNS) tumors are the leading cause of pediatric cancer death, and these patients have an increased risk for developing secondary neoplasms. Due to the low prevalence of pediatric CNS tumors, major advances in targeted therapies have been lagging compared to other adult tumors. We collected single nuclei RNA-seq data from 35 pediatric CNS tumors and three non-tumoral pediatric brain tissues (84,700 nuclei) and characterized tumor heterogeneity and transcriptomic alterations. We distinguished cell subpopulations associated with specific tumor types including radial glial cells in ependymomas and oligodendrocyte precursor cells in astrocytomas. In tumors, we observed pathways important in neural stem cell-like populations, a cell type previously associated with therapy resistance. Lastly, we identified transcriptomic alterations among pediatric CNS tumor types compared to non-tumor tissues, while accounting for cell type effects on gene expression. Cell type-adjusted transcriptomic alterations were associated with nonsense mediated decay and translation associated pathways. Our results suggest potential tumor type and cell type-specific targets for pediatric CNS tumor treatment. In this study, we address current gaps in understanding single nuclei gene expression profiles of previously under investigated tumor types and enhance current knowledge of gene expression profiles of single cells of various pediatric CNS tumors.

3.2. Introduction

Central nervous system (CNS) tumors account for ~25% of pediatric cancer cases and are the leading cause of cancer death in children and adolescents in the United States[334]. Incident pediatric CNS tumors are comprised of many histologically distinct tumor types including pilocytic astrocytomas (15.2%), embryonal tumors (9.4%), and neuronal/mixed neuronal-glial tumors (7.9%)[335]. Survival rates vary widely among tumor types, with a good 10-year survival of 95.4% for pilocytic astrocytomas and a poor 10-year survival of 15.9% for pediatric high-grade gliomas[335]. Pediatric CNS tumor patients are at risk of developing secondary neoplasms, with a 30-year cumulative incidence of malignant secondary neoplasms ranging from 4.7 – 7.8%[434,435]. The standard of care treatments for primary CNS tumors include surgery, radiotherapy, and chemotherapy with relatively limited options for targeted therapy compared to tumors in other anatomic regions.

Recent advances in identifying molecular subtypes in various pediatric CNS tumor types have been made utilizing genomic, transcriptomic and epigenomic data as reflected in the 2021 World Health Organization classification of CNS tumors[336]. For example, medulloblastoma can be classified into four separate molecularly defined subtypes: WNT-activated, SHH-activated and *TP53*-wildtype, SHH-activated and *TP53*-mutant, and non-WNT/non-SHH[351,436–439]. In addition, supratentorial ependymoma can be categorized into *ZFTA* fusion-positive or *YAP1* fusion-positive[353,440]. A better understanding of the molecular variations that exist even among each tumor type has led to novel treatment options. For example, Larotrectinib and entrectinib, targeted therapies for *NTRK* fusion, which has been found in brain tumors, have been approved by the Food and Drug Administration to treat some brain tumors that are metastatic or unresectable with surgery[441,442].

In addition to the molecular characterization of bulk pediatric CNS tumor tissue, emerging work has begun to investigate the transcriptome and cellular states that exist in these tumors at the single cell level. One of the first single cell transcriptomics contributions focused on *H3K27M -altered* pediatric gliomas (n=6, and 3,300 cells) showed that tumors are mainly composed of progenitor cell-like oligodendrocyte populations, rather than differentiated malignant cells[443]. Later, Gojo et al. identified that cellular hierarchies in primary ependymomas (n=28) reflect impaired neurodevelopment and that undifferentiated programs can infer prognosis[305]. Moreover, Gillen et al. revealed that subpopulations in ependymomas (n=26) impact tumor molecular classification of bulk transcriptomes[444]. In medulloblastomas (n=25 and 9,000 cells), Hovestadt et al. identified specific subpopulations associated with molecular subtypes[437]. For example, Group 4 medulloblastoma are composed of differentiated neuronal-like neoplastic cells, while the other three groups are composed of subgroup-specific undifferentiated and differentiated neuronal-like malignant populations[437].

While these single cell and single nucleus transcriptomics studies in 85 total primary CNS tumors to date have improved our understanding of cell states in pediatric CNS tumors, there is still much to be investigated to advance optimal therapeutic options for both primary cancer treatment and reduction of secondary neoplasms. Due to limited sample availability for these rare pediatric CNS tumors, progress in single cell level characterization of these tumors has been relatively slow. Here, we characterized single nuclei gene expression profiles of 35 pediatric CNS tumors and 3 non-tumor pediatric brain tissues. Our study augments previous studies by incorporating single nuclei gene expression profiles of additional pediatric CNS tumor types (dysembryoplastic neuroepithelial tumors, gangliogliomas, etc.) and non-tumor pediatric brain tissue which have not yet been published to our knowledge.

3.3. Methods

Study population

This study of pediatric central nervous system tumors was approved by the Institutional Review Board Study #00030211. Tumor and non-tumor tissues were collected from patients treated at Dartmouth Hitchcock Medical Center from 1993 to 2017. Patients consented to use of tissues for research purposes. Histopathologic tumor type and grade for each sample were re-reviewed according to the 2021 WHO classification of CNS tumors and categorized into the major tumor types[336]. Tumor types included in this study are astrocytoma, embryonal tumors, ependymoma, glioneuronal/neuronal tumors, glioblastoma, and Schwannoma. The average age at diagnosis of subjects from whom the tumor tissues were derived from in this study was 9.3 (range: 0.75 – 18). Male subjects accounted for 62.9% of the tumor samples and female subjects accounted for 37.1% of the tumor samples. Non-tumor brain tissues were obtained from pediatric patients with epilepsy who underwent surgical resection. The average age at diagnosis of subjects from whom the non-tumor samples were derived from was 6.2 (0.58 – 11). Male subjects accounted for 33.3% of the non-tumor samples and female subjects accounted for 66.7% of the non-tumor samples. Specific demographic characteristics of patients for the study are provided in **Table 3-1** and sample information for each subject are provided in **Supplementary Table 3-1.**

Table 3-1. Subject demographics.

	Non-tumor	**Tumor**
Sample size (N)	3	35
# of nuclei	17,451	67,249
Mean (Range)	Pooled	1921.4 (234 – 5795)
Age		
Mean (Range)	6.2 (0.58 – 11)	9.3 (0.75 – 18)
Sex		
F	2 (66.7)	13 (37.1)
M	1 (33.3)	22 (62.9)
Location		
Subtentorial	0 (0.0)	22 (62.9)
Supratentorial	3 (100.0)	13 (37.1)
Tumor type		
Astrocytoma		8 (22.9)
Embryonal		6 (17.1)
Ependymoma		11 (31.4)
Glioneuronal/Neuronal		8 (22.9)
Glioblastoma		1 (2.9)
Schwannoma		1 (2.9)
Grade		
Low (1 + 2)		20 (57.2)
High (3 + 4)		13 (37.1)
NEC/NOS		2 (5.7)

Identification of genetic variation with bulk tissue RNA-seq

RNA was collected using Qiagen RNeasy plus kit (Catalog ID: 74034, Qiagen, Hilden, Germany). RNA-seq libraries were prepared following the Takara Pico v3 low input protocol and sequenced on Illumina NextSeq500.

Raw RNA-seq data were trimmed for polyA sequences and low-quality bases using *cutadapt* (v2.4)[414]. Reads were aligned to human genome hg38 using *STAR* (v 2.7.2b)[425]. Duplicate read identification and other quality control checks for read alignment were performed using CollectRNASeqMetrics and MarkDuplicates in *Picard Tools*.[416] Reads containing N were split using SplitNCigarReads function in the Genome Analysis Toolkit (GATK)[445,446]. Bases quality scores were recalibrated using known variants from the GATK resource bundle and with the BaseRecalibrator and ApplyBQSR functions in GATK[445,446]. Somatic SNV and indels were called with Mutect2 in tumor-only mode[445,446]. Only variants with at least read depth of 10, 5% allele frequency, read depth of 5 for the alternate allele were kept for analysis. The variants were then filtered for variants in sex or mitochondrial chromosomes, RNA editing sites, repeat masker regions, and variants in Panel of Normal (from GATK) references. Variants were then annotated using the Funcotator function in GATK[445,446].

Identification of copy number variation with DNA methylation arrays

DNA were treated with sodium bisulfite following the TrueMethyl® oxBS Module (Tecan Genomics Inc, Redwood City, CA). Converted DNA were hybridized to Infinium HumanMethylationEPIC BeadChips. Raw idat files from the EPIC arrays were processed using preprocessNoob function in *minfi* in R[406]. Copy number variations of tumor samples were estimated in comparison to non-tumor samples using the CNV.fit function in *conumee* package in R[447].

Nuclei isolation, sample multiplexing, and single nuclei RNA-sequencing

Nuclei from fresh frozen tissues were isolated following the Nuclei Pure Prep nuclei isolation kit (Sigma-Aldrich, Catalog ID: NUC201) with some modifications. To summarize, ~10mg of tissue were washed with PBS to remove extraneous OCT the samples were frozen in. The tissue was homogenized with both wide and narrow pestles submerged in 2.5mL of the lysis buffer in a Dounce homogenizer. The lysate mixed with 4.5mL 1.8M sucrose cushion were gently layered on top of the 2.5mL of 1.8M sucrose cushion in Beckman ultracentrifuge tubes. Samples were centrifuged for 45 min at 13,000 RPM at 4°C in an ultracentrifuge. Samples were multiplexed with lipid-tagged oligonucleotides following the MULTI-seq protocol[448]. Nuclei were resuspended in 1% BSA PBS and filtered with 70um and 40um Flowmi filters. Nuclei were quantified with Cellometer K2 (Nexcelom, Lawrence, MA). We aimed for 2,500 – 5,000 nuclei per sample to be sequenced.

Libraries for single nuclei RNA-seq were prepared following the 10x Genomics Single Cell Gene Expression workflows (10x Genomics, Pleasanton, CA) and were sequenced on Illumina NextSeq500 to average 45,000 reads per cell. 10X Cell Ranger software was used to align sequences to GRch38 pre-mRNA reference genome and generate feature-barcode matrices for downstream analyses.

Pre-processing snRNA-seq data

To filter low quality nuclei, only those with greater than 200 and less than 10,000 features and less than 5% of reads that map to the mitochondrial genes were used in downstream analyses. Pooled nuclei were demultiplexed by hashtag oligonucleotides using HTODemux function in Seurat v4[449–452]. Pooled samples were also demultiplexed using Vireo, a genotype based demultiplexing method[453]. We performed genetic demultiplexing analysis using genotype data following the methods described in Weber et al.[454], implemented in a Nextflow workflow[455]. Briefly, bulk RNA-seq reads from each sample were mapped to the reference genome (GRCh38.p13) using STAR[425]. Pooled single-nuclei RNA-seq reads were mapped to the reference genome using STARsolo[456]. Variants among the samples within each pool were identified and genotyped with bcftools mpileup[457] using the mapped bulk reads. Individual cells were then genotyped only at the sites identified using the bulk RNA using cellsnp-lite (mode

1a)[458]. Cell genotypes were used to identify the sample of origin for each cell using Vireo[453]. Code for the genetic demultiplexing workflow can be found at https://github.com/AlexsLemonade/alsf-scpca/tree/main/workflows/genetic-demux.

To integrate the methods, we first used sample identity assigned from the hashtag oligonucleotides. If the nuclei were confidently assigned a sample, it was compared to the genotype-based sample assignment. Those that did not match the same sample were filtered out. If the nuclei were assigned as a doublet or to none of the samples, the nuclei were assigned to a sample based on the genotype-based approach. 84,700 nuclei with confident sample assignment were used in analysis.

As our dataset included a very large number of nuclei to be integrated and was expected to have certain cell types only present in certain samples, we used the reciprocal PCA integration approach on the 2,000 most variable features to combine the nuclei from each sample. We first found the integration anchors with the FindIntegrationAnchors function then used the IntegrateData function in Seurat v4 to integrate all our filtered nuclei[450–452].

Dimension reduction and clustering of snRNA-seq data

The integrated dataset was scaled using the ScaleData function in Seurat. First, PCA dimensionality reduction analyses were done to identify 100 principal components (PCs). To further reduce the dimensionality and cluster our nuclei by their gene expression profile, we conducted UMAP analyses on the 50 PCs with highest standard deviation with RunUMAP function in Seurat[449,459]. Then, we clustered our cells using FindNeighbors (n_neighbors = 30) and FindClusters (resolution = 1.0) function in Seurat[449].

Gene set enrichment testing

Gene set enrichment tests at the single cell level were conducted using the Variance-Adjusted Mahalanobis (VAM) method[460]. The vamForSeurat function from the VAM R package was used to calculate enrichment scores for each nucleus. Brain cell type specific gene sets from the Molecular Signatures Database (MSigDB) v7.5.1 were used to validate our single cell identities[461–466]. For identifying cell types, p-values were calculated from the cumulative distribution function values generated by VAM. Nuclei were considered to be associated with a specific brain cell type-pathway if the VAM-generated p-value was ≤ 0.05. Nucleus-level pathway scoring was also conducted using VAM for pathways in the MsigDB Pathways Interaction Database (PID) collection[467]. PID Pathways were considered to be enriched in each nucleus at the FDR adjusted p-value threshold of 0.1 for the VAM-generated p-values.

Stemness scores for each nucleus were calculated using the stemness-associated gene list from Tirosh et al[468] and the AddModuleScores function in Seurat.

Differential gene expression and pathways

Differential expression analysis between tumor nuclei and non-tumor nuclei were conducted using monocle3[469–472]. Differential expression analyses were conducted only on the top 4,000 most variable features identified from the FindVariableFeatures Seurat function. The unadjusted differential expression testing was done using the fit_models function in monocle3 (v1.0.0) R package with the quasi-poisson distribution with the non-tumor nuclei being the referent gene expression profile[469,471,472]. The adjusted differential expression testing was done with the same quasi-poisson distribution with non-tumor nuclei being the referent but including the major cell type identity in the model. Gene types for each gene used in the differential expression testing were annotated using the org.Hs.eg.db[473], Human genome annotation package, and mapIds function in the AnnotationDbi R package[474]. Pathways associated with the differentially expressed genes were identified using the Reactome pathways and ReactomePA R package[421].

Pathways important for each cell cluster were identified using FindAllMarkers function in Seurat v4 with the Wilcoxon rank sum test in Seurat on the binary classification of PID pathways enrichment for each nuclei[475]. Log fold change and minimum percentage of cells enriched in each pathway were both set to 0. To identify the pathways with greater number of nuclei with enriched pathway per cluster, we selected pathways that were only positive in direction in the FindAllMarkers options.

Statistical testing

Observed proportion of genes that were either increased or decreased in the same direction in the shared differentially expressed genes among all the tumor types (60.9%) were compared to expected proportion of genes that would be increased or decreased in the same direction across all the tumor types (3.13%) using a one-sample proportion test. The expected proportions were determined based on the permutations of direction of change compared to non-tumor for the six tumor types.

3.4. Results

Samples from pediatric central nervous system tumors and non-tumor pediatric brain tissue were obtained from patients being treated at Dartmouth-Hitchcock Medical Center and Dartmouth Cancer Center from 1993 to 2017. Non-tumor pediatric brain tissues from the supratentorial regions were collected from patients undergoing surgical resection for epilepsy. Patient characteristics are described in **Table 3-1**. Pathological re-review for histopathologic tumor type and grade were done according to the 2021 World Health Organization CNS tumor classification system and categorized into the broader tumor types to balance sample size per tumor type[336]. Specific diagnoses for each sample can be found in **Supplementary Table 3-1**.

Genetic variants were identified using bulk tissue RNA-seq data for all tumors except for two tumors due to low bulk RNA-seq data quality. Copy number variations (CNV) were determined using bisulfite treated DNA methylation array data. Genetic and cytogenic variations varied among tumors and tumor types (**Figure 3-1**). Interestingly, across all but one tumor sample, tumors had genetic variants in *MALAT1*. Many of the genetic variants detected within the pediatric CNS tumors were associated with epigenetic processes. For example, almost half of the tumors, across tumor types, had genetic variants in *HIST1H1E* (14/33). CNV patterns in some tumor types were as expected from previous literature. For instance, 5 out of the 9 ependymoma had chromosome 1q gain, which has been considered to be an early tumorigenic event in ependymoma[476,477].

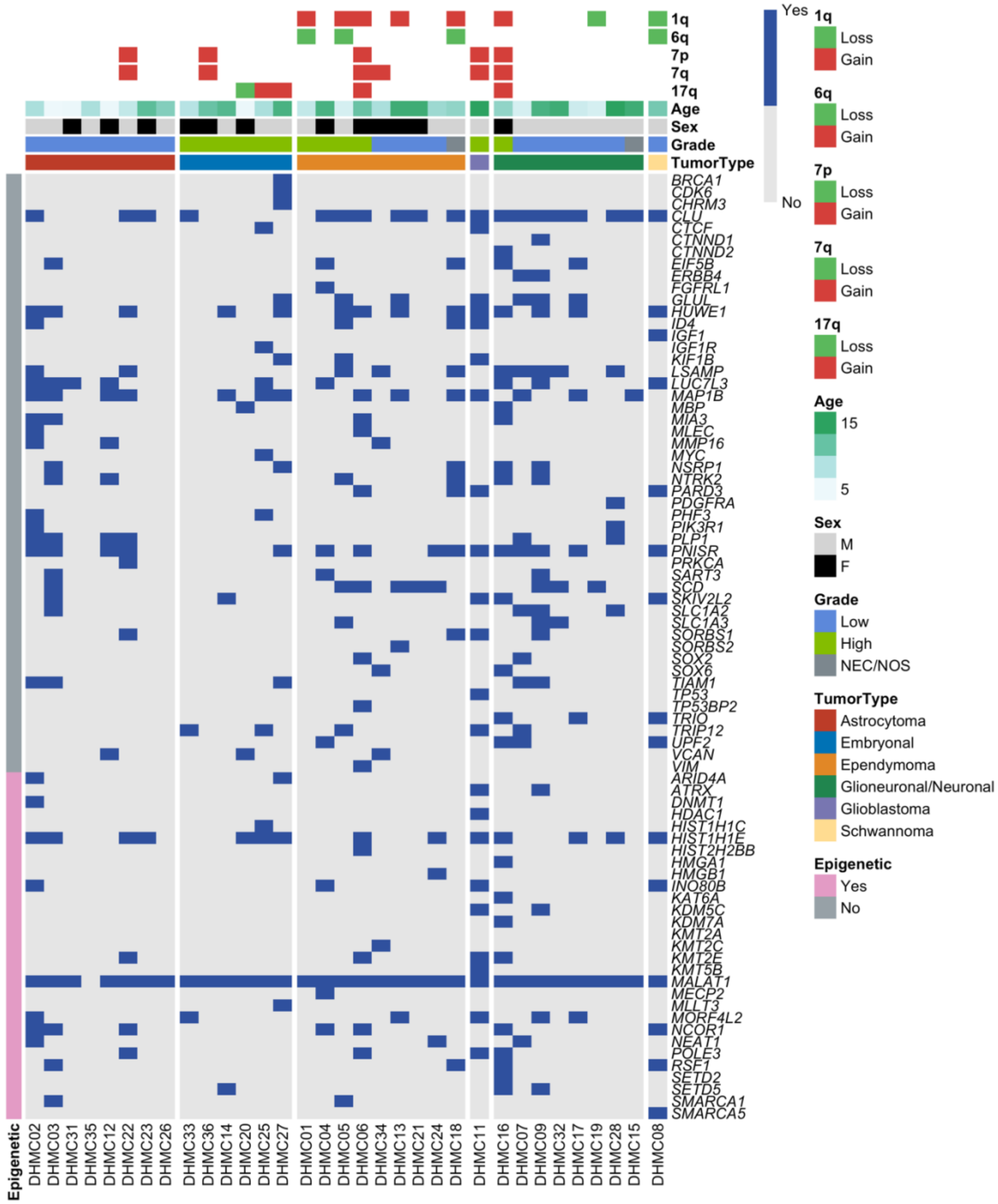


Figure 3-1. Genetic and cytogenic characteristics of pediatric CNS tumors. Heatmap of presence of genetic variant in select genes. Blue squares indicate presence of genetic variant. Gray squares indicate the genetic variant was undetected. Vertical tracking bars indicate whether the gene is associated with epigenetic processes. Horizontal tracking bars correspond to each patient’s age, gender, grade, tumor type, and copy number variations in select chromosomes.

Integrated de-multiplexing method to increase single nuclei RNA-seq data yield

Using lipid-tagged hashtag oligonucleotides (HTO), 34 samples (out of 38 total samples) were multiplexed in 17 pools to collect 10X genomics snRNA-seq data[448]. The distribution of samples across sequencing runs and pools is provided in **Supplementary Table 3-1**. As many nuclei were not tagged with sufficient HTO to be efficiently demultiplexed in downstream analyses, we aimed to augment demultiplexing by analyzing sequencing-derived genotype data from each nucleus together with HTO information and assign additional nuclei to specific samples (**Figure 3-2A**). To summarize our demultiplexing process, we first used HTOs to assign the nuclei to their respective samples. For samples that were assigned confidently with the HTO, we filtered to keep only the nuclei that were assigned to the same sample concordantly using genotype information. For nuclei that were either unassigned to a sample or assigned as a doublet with HTO, we assigned nuclei to samples using genotype information (detailed in the methods section). The final set of nuclei per sample were comprised of the filtered nuclei from HTO and genotype identified nuclei. An example of the single nucleotide variants identified per pool along with their assigned sample can be found in **Figure 3-2B**. An example of how many nuclei were obtained for one pool, during each step, is shown in **Figure 3-2A** on the right. The integrated demultiplex method classified an average of 1,921 nuclei per sample (range = 234 – 5795, **Table 3-1**). The number used in downstream analysis per sample is included in **Supplementary Table 3-1**. The total number of demultiplexed nuclei was increased 47.4% (additional 27,248 nuclei) using the integrated approach over the HTO-only method, and 15.6% (additional 11,445 nuclei) over the genotype-based method alone (**Figure 3-2C**). Gene expression profiles for a total of 84,700 nuclei were used for downstream analyses.

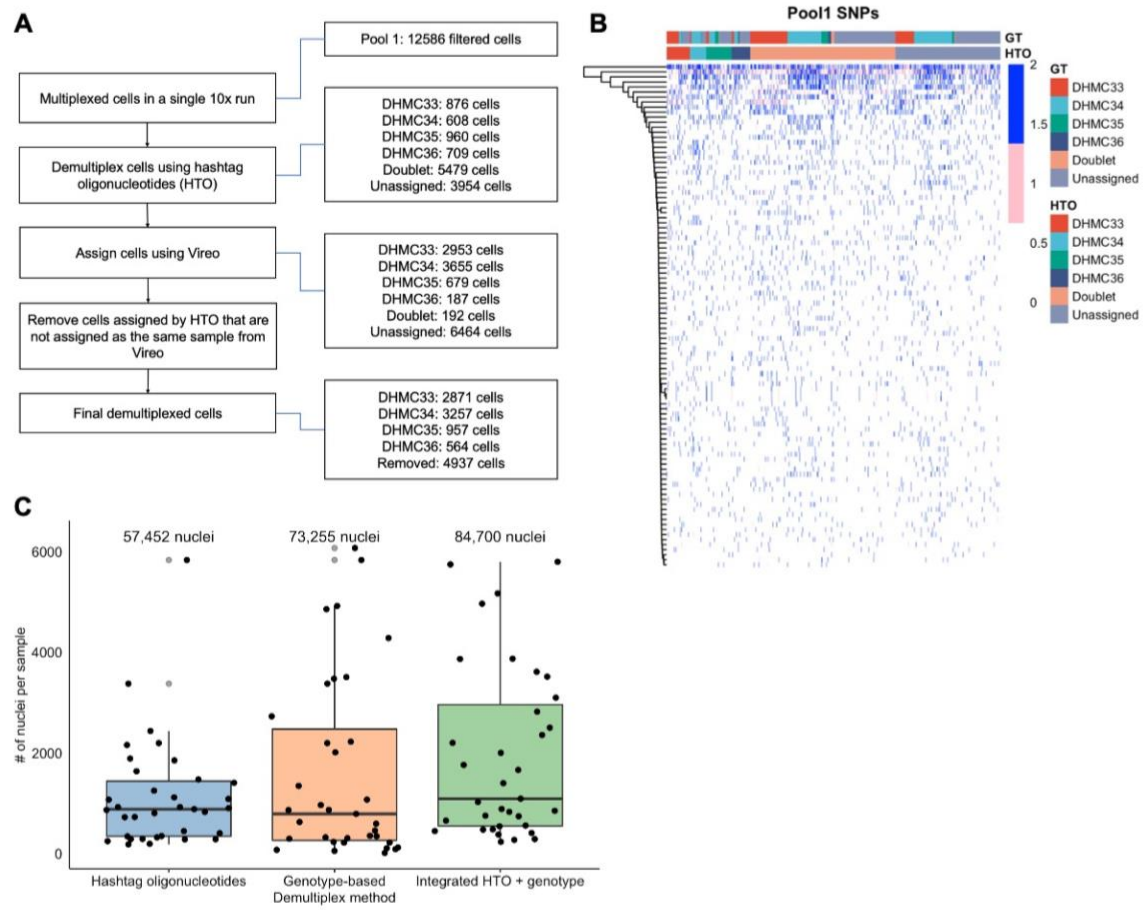


Figure 3-2. Integrative method to demultiplex pooled samples increases nuclei per sample from single-nuclear RNA-seq data.
A) Diagram of integrated method for demultiplexing pooled samples. Multiplexed samples were first demultiplexed using hashtag oligonucleotide (HTO) counts. Cells assigned using HTO were filtered for those that did not match the sample assignment from genotype-based method. Cells unable to be assigned to a sample from HTO were assigned based on genotype information. On the right are the number of cells retained at each step of the integrated demultiplex method for Pool #1. **B)** Example of genotype information (Pool #1) used to demultiplex samples. Blue indicates 100% alternate allele presence. Pink indicates heterogeneous alternate allele presence. White indicates no alternate allele depth presence. Tracking bars indicate the samples assigned based on HTO or genotype (GT). **C)**. Number of nuclei assigned per sample based on hashtag oligonucleotides, genotype-based method, or integrated method. The total number of nuclei obtained for each method is labeled on the top of the boxplot.

Cell type heterogeneity in pediatric central nervous system tumors and non-tumor pediatric brains

Out of 84,700 nuclei, 67,249 nuclei (79%) were from pediatric CNS tumors and 17,451 nuclei (21%) were from non-tumor tissue (**Figure 3-3A**). Across all samples, snRNA-seq data revealed 58 clusters that were grouped into 16 major cell types: astrocytes (AST), embryonal tumor cells (EMB), endothelial cells (EN), macrophage/microglia (MAC/MG), neurons (NEU), excitatory neurons (NEU_EX), granular neurons (NEU_GN), inhibitory neurons (NEU_INH), interneurons (NEU_INT), neural stem cells (NSC), oligodendrocytes (OLIG), oligodendrocyte precursor cells (OPC), radial glial cells (RGC), stromal cells (ST), T cells (TC), and unipolar brush cells (UBC) (**Supplementary Figure 3-1A, Figure 3-3B**). The clusters were classified into cell types using classical markers for cell types found in the brain (**Table 3-2**). The following gene markers for cell types were used: *GFAP* and *AQP4* for astrocytes; *FN1* and *COL4A1* for endothelial cells; *CSF1R* and *PTPRC* for macrophage/microglia; *RBFOX3* and *RELN* for neurons and unipolar brush cells; *GAD2* for inhibitory neurons and interneurons; *SOX2* and *CD44* for neural stem-like cells; *MOG* and *PLP1* for oligodendrocytes; *PDGFRA* for oligodendrocyte precursor cells; *VIM, NES,* and *PAX6* for radial glial cells; *FAP* for stromal cells; *CD3E* for T cells. Not all gene markers corresponded to expected expression levels for the major cell types. For example, the neural stem cells (NSCs) did not express classical neural stem cell like genes (*SOX2* and *CD44*) but were identified by enrichment testing of neural stem cell/neural progenitor-like cell gene sets. Because the embryonal tumor cells (EMB) clusters were unlike any other classical cell type found in the brain, the cells in these clusters were classified as embryonal tumor cells. These marker-based cell type classifications were subsequently validated by enrichment of cell type-specific pathways using the Variance-adjusted Mahalanobis method, a single cell-level pathway enrichment method (**Figure 3-3C, Supplementary Figure 3-1B**)[460–466,478]. The cell type-specific pathways used for enrichment testing were derived from single cell RNA-seq experiments of developing human and mouse brains.

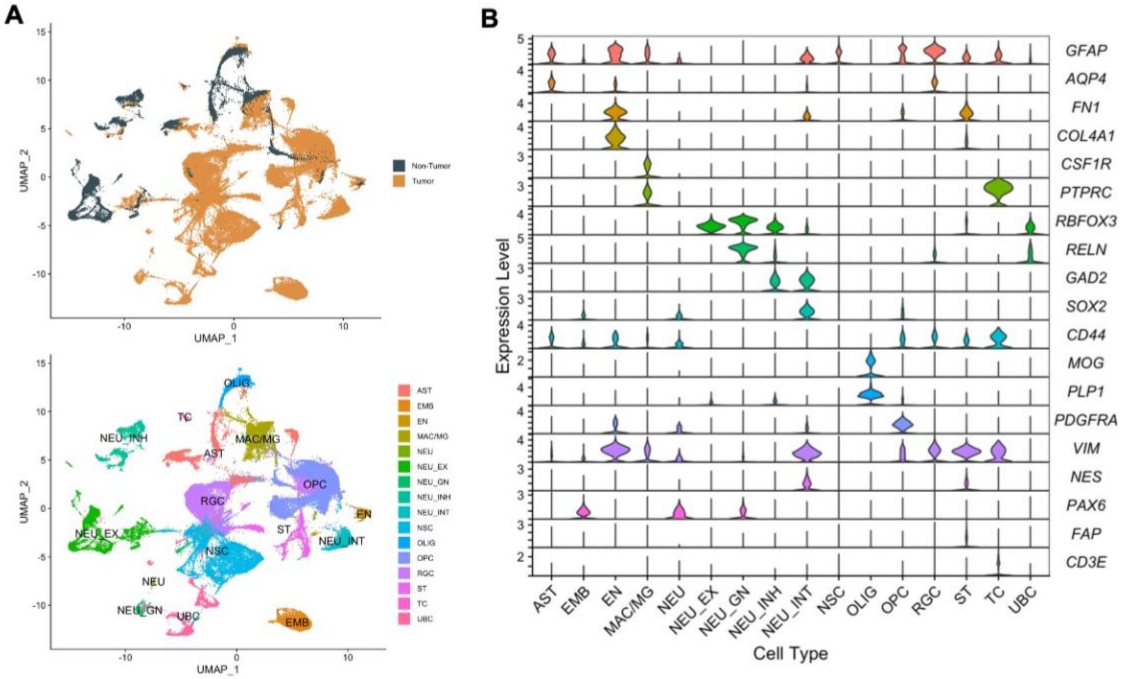


Figure 3-3. Heterogeneity of cell types in pediatric CNS tumor tissue and non-tumor pediatric brain tissue.
A) UMAP of the 84,700 nuclei colored from tumor and non-tumor tissue. Dark green indicates nuclei from non-tumor tissue. Orange indicates nuclei from tumor tissue. **B)** UMAP of the 84,700 nuclei colored by major cell type. **C)** Gene expression levels of classical gene markers for cell types present in the brain by major cell type cluster. Astrocytes (AST): *GFAP* and *AQP4*; Endothelial cells (EN): *FN1* and *COL4A1*; Macrophage/microglia (MAC/MG): *CSF1R* and; Neurons and unipolar brush cells (NEU, NEU_EX, NEU_GN, UBC): *RBFOX3* and *RELN*; Inhibitory neurons and interneurons (NEU_INH, NEU_INT): *GAD2*; Neural stem cells (NSC): *SOX2* and *CD44*; Oligodendrocytes (OLIG): *MOG* and *PLP1*; Oligodendrocyte precursor cells (OPC): *PDGFRA*; Radial glial cells (RGC): *VIM, NES,* and *PAX6*; Stromal cells (ST): *FAP*; T cells (TC): *CD3E.*

Table 3-2. Classic markers for cell types in the brain

Cell Type	Markers
Astrocytes	*GFAP; AQP4*
Endothelial cells	*FN1; COL4A1*
Macrophage/Microglia	*CSF1R; PTPRC*
Neurons/Unipolar brush cells	*RBFOX3; RELN*
Inhibitory neurons/Interneurons	*GAD2*
Neural stem-like cells	*SOX2; CD44*
Oligodendrocytes	*MOG; PLP1*
Oligodendrocyte precursor cells	*PDGFRA*
Radial glial cells	*VIM; NES; PAX6*
Stromal cells	*FAP*
T cells	*CD3E*

To identify stem-like phenotypes in our tumor nuclei population, we investigated the expression levels of classically used markers of cancer stem cells (*ITGA6, CD44, PROM1, NES, MSI1, MYC, NANOG, SOX1, SOX2, POU5F1, VIM, SDC1, SDC2, GPC1, GPC2*), as well as an enrichment score for stemness from Tirosh et al[468,479]. Levels of expression for genes classically used for to isolate stem-like cells in literature varied among the different cell types (**Supplementary Figure 3-2**). Interestingly, cell types expected to be more differentiated, like astrocytes, had relatively high levels of *CD44* and *VIM,* and these genes were expressed in many of the cell types. In addition, although the NSC-like cluster had a high stemness score, the expression of cancer stem cell markers was minimal. Unexpectedly, the UBC-like clusters also had elevated stemness scores. While gene expression levels may not always correlate with protein expression, our results indicate cell types identified using classical stem cell markers may not capture all tumor cells with stemness features.

Next, we tested for potential associations of clinical variables with tumor stemness scores. We first assessed the distribution of stemness scores among nuclei in each sample and determined the median stemness score (**Supplementary Figure 3-3**).

We found that the stemness scores were higher in embryonal compared to other tumor types or non-tumor tissue (**Supplementary Figure 3-4A**). Specifically, embryonal tumors had significantly higher stemness scores compared to astrocytomas (*P*-value = 0.03), ependymoma (*P*-value = 0.02), and glioneuronal/neuronal tumors (*P*-value = 0.029). Compared with low grade tumors, high grade tumors had higher stemness scores (*P*-value = 0.008, **Supplementary Figure 3-4B**), and somewhat unexpectedly, stemness score was positively correlated with age (R = 0.47, *P*-value = 0.004, **Supplementary Figure 3-4C**). No difference in stemness score was observed between tumors in the subtentorial and supratentorial regions of the brain (*P*-value = 0.600, **Supplementary Figure 3-4D**). Our results indicate that stemness level of single cells is associated with tumor type and grade, which may be important when considering potential for therapy resistance and metastasis and when developing targeted therapies.

To reveal any specific cell populations that are only present in a restricted set of tumor types, we evaluated the association between cell type proportions and tumor type (**Figure 3-4, Supplementary Figure 3-5**). Non-tumor tissue contained nuclei from all major expected cell types found in normal brain, including astrocytes, oligodendrocytes, and excitatory and inhibitory neurons which demonstrated the high-quality data derived from the non-tumor tissues. Some tumor samples had small proportions of cell types normally present only in non-tumor tissue, such as excitatory neuron cluster #5 (NEU_EX5) and inhibitory neuron cluster #2 (NEU_INH2). These cases are likely the result from the inclusion of cells from the tumor margin. Non-tumor tissues had limited numbers of nuclei from progenitor-like cell types, like NSCs, RGCs, or UBCs. While OPCs are a progenitor cell type, they are also found in normal brain tissue. The non-tumor OPCs were limited to the OPC4, a population transcriptionally distinct from tumor OPCs residing in OPC1-3.

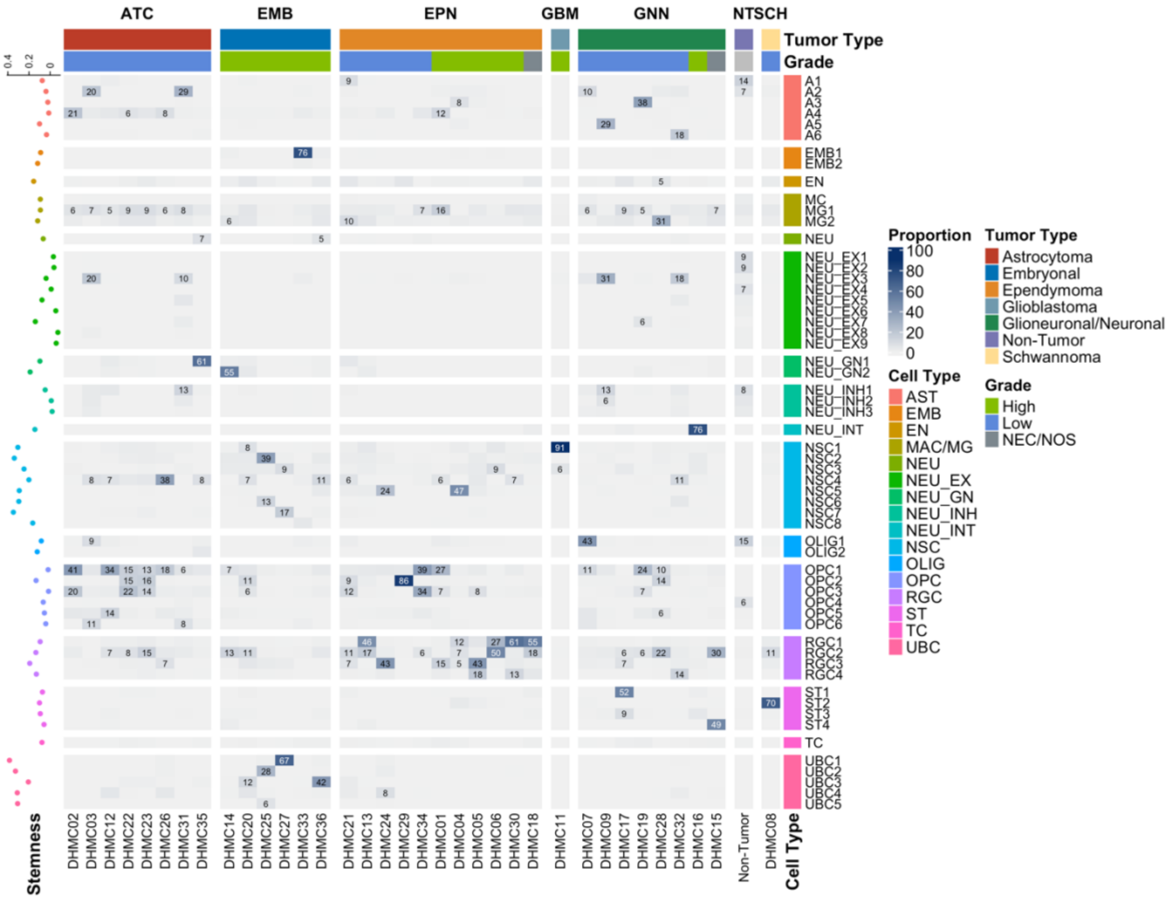


Figure 3-4. Tumor type-specific presence of cell types.
Heatmap of the proportions (%) of each cell type present in each sample. Scatter plot on the left of the heatmap indicates the median stemness level of each cell type. Horizontal tracking bars indicate the tumor type and grade of each sample. Vertical tracking bars indicate the major cell types of the nuclei. The cell types with greater than 5% are labeled within each cell. ATC: Astrocytoma; EMB: Embryonal tumors; EPN: Ependymoma; GBM: Glioblastoma; GNN: Glioneuronal/neuronal tumors; NT: Non-tumor; SCH: Schwannoma.

Some cell types were exclusive to a specific tumor. For example, the glioblastoma sample was comprised of 91% NSC1, and an ependymoma sample consisted predominantly (86%) of OPC2. MG2 was present at higher proportions (mean = 3.3%, range = 0.3 – 31.2%) in tumors compared to non-tumor tissue (0.9%). All astrocytomas had at least small proportions of A4, OPC1, and OPC5. The embryonal tumors had cell types that were more neuronal (apart from EMB cell types) like NSCs and UBCs. Large proportions of ependymoma samples were made of RGC clusters. The glioneuronal/neuronal tumor type samples were more varied in terms of which cell types were more present in each tumor. The expanded cell types were consistent with some known cell types of origin for these tumors, such as the RGCs in the ependymomas.

Cell type-specific pathway enrichment in pediatric CNS tumors

First, to determine cell type-specific pathway enrichment in each nuclei of the tumor samples, we conducted a pathways analysis at the single cell level using the Variance-adjusted Mahalanobis (VAM) method, which computes cell-level pathway scores that account for the technical noise and inflated zero counts of single cell RNA-seq data[460]. We used 196 pathways from the MSigDB Pathway Interaction Database (PID) collection for our enrichment testing[461,462,467]. The cell-level enrichment p-values generated by VAM were corrected for false discovery rate using the Benjamini-Hochberg method and classified to be significantly enriched in each nucleus if the FDR adjusted p-value was less than 0.1 as binary classifications (enriched or not enriched).

Next, we determined any pathways that were more specific for each cell type to determine pathways important in each cell type. The PID pathways were considered to be important/specific to the cell type under adjusted p-value < 0.05 threshold in the differential enrichment test. For cell types with a limited presence in tumor tissues, like many of the excitatory neurons and A1, we observed no pathways that were specific to the clusters (**Supplementary Figure 3-6A, Supplementary Table 3-2**). The immune-related cells (MG1, MG2, and TC), which were present in tumor tissue at slightly higher levels than in non-tumor tissue, had more than 44% of the PID pathways specific to these cell types. The high percentage of PID pathways that were important in the immune-related cell types is likely due to the relatively greater number of cytokine and other immune-associated pathways are included in the PID database.

All NSC clusters, except for NSC6 (3.6% pathways), had more than 10% of PID pathways that were important to the NSCs (range = 11.73 – 42.35%, **Supplementary Figure 3-6A, Supplementary Table 3-2**). While there were no shared pathways that were considered to be important in all 8 NSC clusters, there were numerous pathways shared among majority of the NSCs (**Figure 3-5A**, **Supplementary Figure 3-6B**). The retinoic acid pathway and telomerase pathway were considered to be important in 7 of 8 NSC clusters (**Figure 3-5B**). Aurora-B, PLK1, FOXM1, E2, ATR, FOXO, Retinoic Acid pathways were considered to be important in just 6 of the 8 NSC clusters. Our results provided potential cell type-specific targets within these PID pathways important for each cluster for future therapeutic strategies.

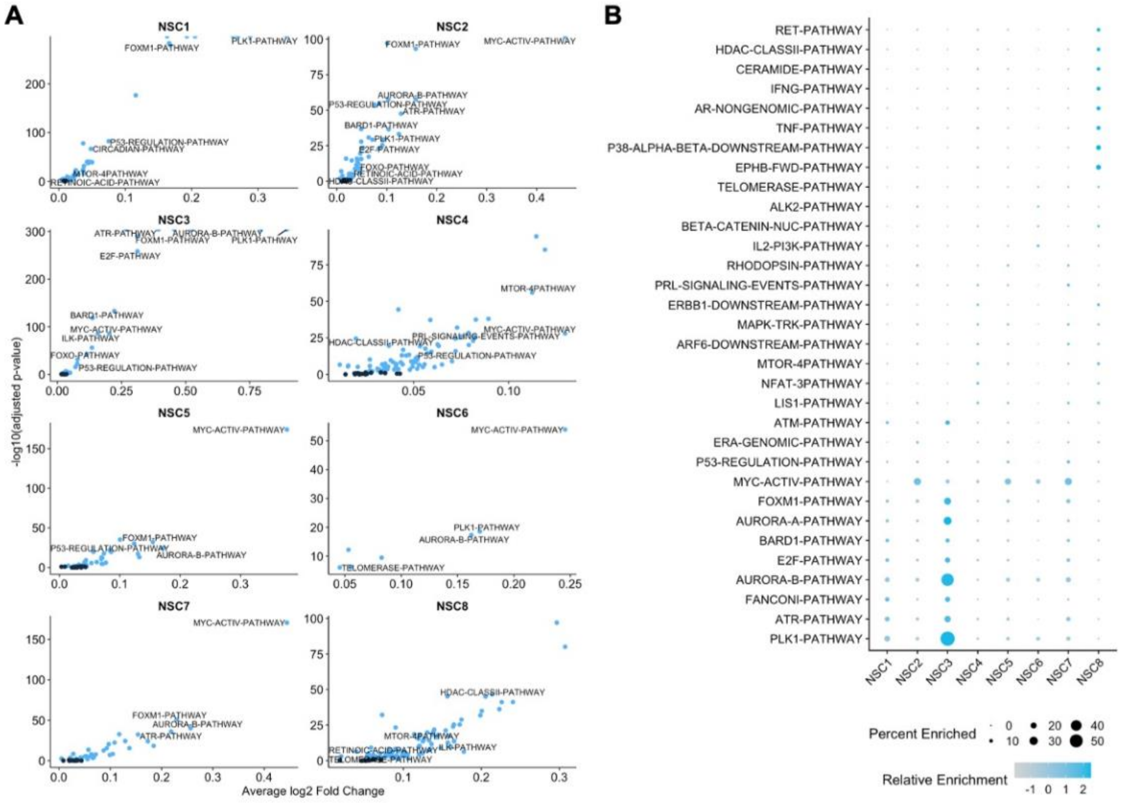


Figure 3-5. Enriched pathways in neural stem cell-like cells in pediatric CNS tumors.

A) Differentially enriched pathways from Pathways Interaction Database (PID) in the NSC subpopulations compared to all other cell clusters in pediatric CNS tumors. Blue points indicate statistically significantly enriched pathways at adjusted p-value threshold of 0.05. Labeled pathways indicate more commonly enriched pathways in the NSC subpopulations. The few points that appear to be cut-off have -log10(adjusted p-value) of infinity as the adjusted p-values were essentially zero.
B) Relative enrichment and percentage expressed in cluster of the top enriched pathways per NSC clusters. Color indicates relative enrichment. Size indicates percentage expressed in each NSC cluster.

Transcriptomic alterations in tumors compared to non-tumor at the single cell level

We next aimed to determine transcriptomic alterations in pediatric CNS tumors compared to non-tumor pediatric brain tissue. In bulk differential gene expression analyses, it is typically not possible to account for the impact of cell composition differences on gene expression levels[4,480–482]. Here, using single nuclei level data, we compared expression of the 4,000 most variable genes in nuclei from each tumor type to the gene expression of nuclei in non-tumor tissue, controlling for cell-type composition differences (**Figure 3-6A**). Genes were considered differentially expressed if they met the FDR < 0.05 threshold.

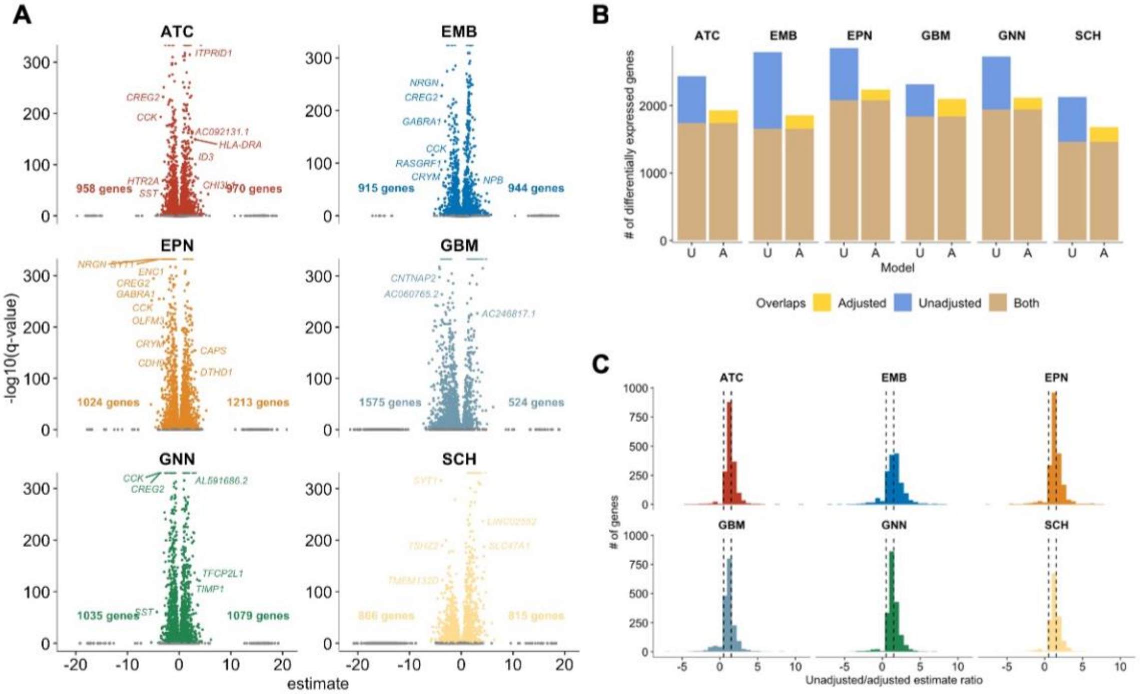


Figure 3-6. Transcriptomic alterations in pediatric CNS tumor cells compared to non-tumor pediatric brain cells.
A) Volcano plot of differentially expressed genes for each tumor type compared to non-tumor tissue, adjusted for major cell type. Number of genes on the left of the volcano plot indicate genes that are downregulated compared to non-tumor tissue. Number of genes on the right of the plot indicate genes that are upregulated compared to non-tumor tissue. **B)** Comparison of the number of differentially expressed genes in the adjusted model and the unadjusted model per each tumor type. **C)** Distribution of differential expression estimates in the unadjusted model to

estimates in adjusted model per tumor type. Dashed lines at 0.5 and 1.5 to indicate genes with similar estimates in the two models.

As expected, adjusting for cell type proportions reduced the number of significantly differentially expressed genes compared with cell-type-unadjusted analyses. However, importantly, cell-type-adjusted analyses identified on average 200 genes per tumor type that not observed in unadjusted models. (**Figure 3-6B, Supplementary Figure 3-7A, 7B, Supplementary Table 3-3**). Genes uniquely identified in cell-type-adjusted models represent underlying tumor biology that was obscured by variation in cell type proportions composing the tumor microenvironment across subjects (**Figure 3-6C**). For example, *WNT3A,* a gene shown to mediate glioblastoma progression[483] was shown to be upregulated in glioneuronal/neuronal tumors and Schwannoma only using the adjusted analysis (**Supplementary Table 3-3**). Furthermore, the unadjusted model often gave estimates that were contrary to the direction of change from the adjusted model. For example, *FAT2* was significantly decreased (estimate = -1.80) in embryonal tumors relative to non-tumor tissue in the adjusted model but significantly increased (estimate = 0.42) in embryonal tumors in the unadjusted model. Also, *FGFR2* had significantly increased in expression (estimate = 0.76) in the Schwannoma nuclei relative to non-tumor tissue in the adjusted model but was significantly decreased in expression (estimate = -0.52) in the unadjusted model.

Using cell type-adjusted models, we detected tumor type-specific alterations in gene expression compared to non-tumor tissue. In astrocytomas, we identified 958 significantly downregulated and 970 significantly upregulated genes compared to non-tumor tissue (FDR < 0.05). Genes upregulated in astrocytomas include *ID4, CD74* and *FOS.* The differentially expressed (DE) genes in astrocytomas were associated with translation-related and nonsense-mediated decay-related processes (**Supplementary Figure 3-8A, Supplementary Table 3-4**). Embryonal tumors had 915 downregulated and 944 upregulated genes relative to non-tumor tissue that were associated with rRNA processing and translation-associated processes (**Supplementary Figure 3-8B,**

Supplementary Table 3-5). In embryonal tumors, the topmost DE genes included many ribosome-associated genes like *RPS2, RPLP1,* and *RPL13A* as well as histone H3.3 related genes like *H3F3A* and *H3F3B*. Ependymomas had 1024 downregulated and 1213 upregulated genes compared to non-tumor tissue. The topmost DE genes were *IGFBP5, CFAP54* and *COLEC12.* Similar to astrocytomas, DE genes in ependymomas were associated with translation and nonsense-mediated decay related processes (**Supplementary Figure 3-8C, Supplementary Table 3-6**). Glioneuronal/neuronal tumors had 1,035 downregulated and 1,079 upregulated genes relative to non-tumor tissue; these genes that were associated with extracellular matrix and integrin-related processes and MET signaling (**Supplementary Figure 3-8D, Supplementary Table 3-7**). *TAFA1, ALK,* and *VAV3* were some of the topmost DE genes in glioneuronal/neuronal tumors. In the glioblastoma, there were 1,575 downregulated genes and 524 upregulated genes that were associated with RNA processing and translation-related processes (**Supplementary Figure 3-8E, Supplementary Table 3-8**). Some genes that were topmost DE in glioblastoma nuclei include *RMST, ID4* and *PBX3.* Lastly, in the Schwannoma, there were 864 downregulated genes and 813 upregulated genes relative to non-tumor tissue that were associated with elastic fibers and RHO/RAC1 GTPases cycles (**Supplementary Figure 3-8F, Supplementary Table 3-9**). *CEMIP, THSD4,* and *GPC6* were among the topmost DE genes in the Schwannoma nuclei. Only the top 10 most associated pathways are reported in **Supplementary Figure 3-7.** The list of differentially expressed genes and their associated pathways per tumor type are listed in **Supplementary Table 3-3 – 3-9**, respectively.

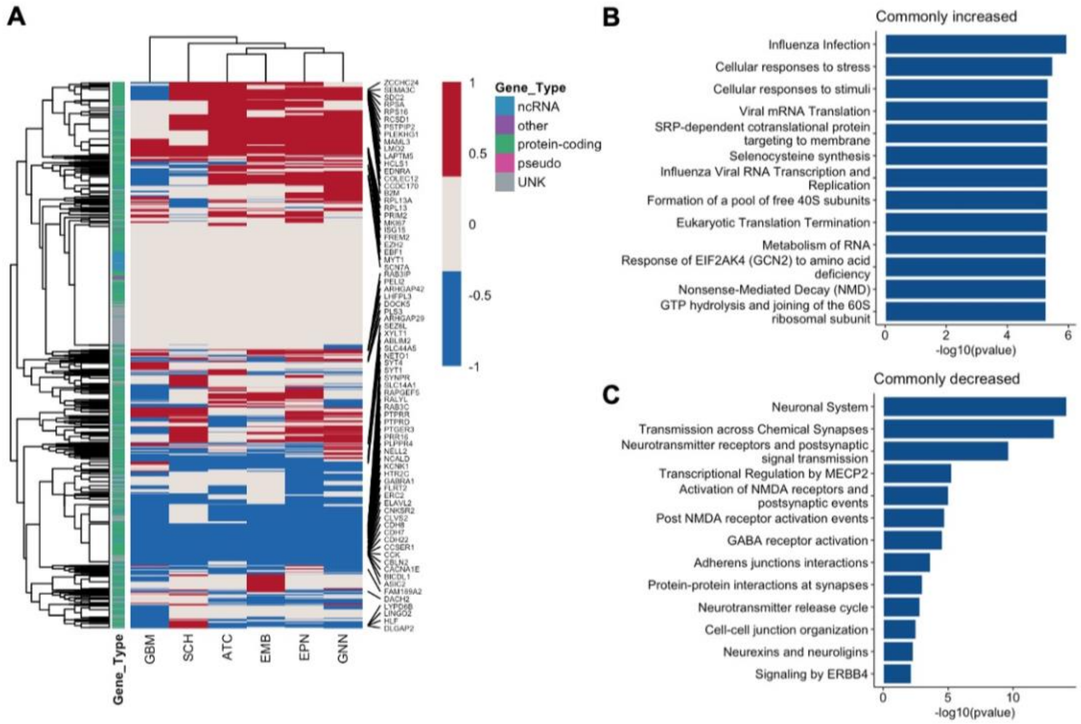


Figure 3-7. Adjusting for cell type identity identifies novel genes associated with pediatric CNS tumor types.
A) Heatmap of differential expression direction and significance in all 4000 genes tested in differential expression analyses. Red indicates significantly upregulated in the tumor type compared to non-tumor tissue. Blue indicates significantly downregulated in the tumor type compared to non-tumor tissue. Gray indicates the gene is not significantly differentially expressed. Tracking bar indicates the gene type. **B)** Top Reactome pathways associated with genes commonly upregulated across all tumor types. **C)** Top Reactome pathways associated with genes commonly downregulated across all tumor types.

Of the 4,000 most variable genes that were used in differential gene expression analysis, there were 558 genes that were differentially expressed in all six of the tumor types, 717 in five of the tumor types, and 596 in four of the tumor types compared with non-tumor tissue (**Figure 3-7A, Table 3-3**). There were differentially expressed genes specific to a single tumor type: 43 genes for astrocytomas, 61 for embryonal tumors, 52 for ependymomas, 68 for glioneuronal/neuronal tumors, 98 for glioblastoma, and 57 for Schwannoma. While 60.9% (340/558) of the differentially expressed genes shared among all the tumor types were either increased or decreased the same direction, the remainder of genes varied in the direction of change based on tumor type compared to non-tumor tissue. The proportion of genes that either increased or decreased in the same direction for the shared significantly differentially expressed among all tumor types were significantly higher than expected (*P-value* < 2.2×10^{-16}). Protein-coding genes with increased expression across all tumor types included *E2F7, ETS1, EZH2, ID3/4, MKI67, PIK3R3,* and *TOP2A.* We conducted a pathways analysis of the genes with increased expression across all tumor types, and genes with decreased expression across all tumor types with Reactome pathways[421]. Interestingly, translation or nonsense mediated decay related processes having increased expression across all tumor types compared to non-tumor tissues (**Figure 3-7B**). Shared decreased protein-coding genes across all tumor types included *FOXP2, GABRA1/2/4/5, NRGN, SST,* and *SYNPR*. Even when differential gene expression analyses were adjusted for cell type, across all tumor types, there was decreased expression in genes associated with neuronal system such as transmission across chemical synapses and activation of NMDA or GABA receptors (**Figure 3-7C**). Hierarchical clustering of the differentially expressed genes revealed that transcriptomic alterations were similar in ependymomas and glioneuronal/neuronal tumors and likewise in astrocytomas and embryonal tumors (**Figure 3-7A**).

Table 3-3. Number of significantly differentially expressed genes shared among all or subsets of tumor types

Number of tumor types	Number of genes shared among tumor types
0	868
1	379
2	424
3	458
4	596
5	717
6	558

3.5. Discussion

In this study, we characterized gene expression profiles of 84,700 nuclei from snRNA-seq of 35 pediatric CNS tumors and 3 pediatric non-tumor brain tissues. We utilized an integrated hashtag oligonucleotide and genotype-based methods to maximize the number of sample-assigned nuclei from our multiplexed snRNA-seq experiment. Although the original MULTI-seq[448] work showed that multiplexing nuclei was feasible, some difficultly encountered with the approach in our study may have been attributable to use of fresh frozen samples that had been stored in the freezer for decades. In our study, we detail a novel approach to increase the number of cells assigned to a specific sample from pooled sequencing runs by integrating a genotype-based approach to demultiplex snRNA-seq data. Future studies are expected to benefit from our integrated demultiplexing method to maximize data usage while decreasing the cost of snRNA-seq experiments.

Our study incorporates pediatric CNS tumor types that have not yet been characterized with single cell or single nuclei RNA-seq such as gangliogliomas. Moreover, we incorporated non-tumor pediatric tissues in our experiment, which to our knowledge have not been included in previous pediatric CNS tumor single cell RNA-seq studies. We describe changes in cell type proportions specific to each tumor type and use this information to identify the gene expression profiles and pathways enriched across tumor and normal samples through a cell type-adjusted analysis.

We characterized major cell subpopulations in specific tumor types, some of which have not been previously established. This includes the expansion of oligodendrocyte precursor cell (OPC) subpopulations in astrocytomas, and unipolar brush-like cells (UBC) with high stemness levels enriched in embryonal tumors. In the ependymomas, there was a significant presence of radial glial-like cells (RGC). Some glioneuronal/neuronal tumors featured stromal cells (ST) that were less present in other tumor types, demonstrating significant variability even within subtypes of tumors. The glioblastoma sample was predominantly comprised of a neural stem cell-like cell population. The Schwannoma sample was comprised of a specific stromal cell type. Despite some overlap in the major cell types between tumor and non-tumor nuclei, their gene expression profiles were distinct. For example, the OPC4 cluster is unique to non-tumor nuclei, while tumor OPCs reside in OPC1-3. Some neuron-like clusters (i.e. NEU_EX3) that were present in tumors had very limited presence in the non-tumor samples. Our results suggest distinct tumor-associated gene expression alterations even if the tumor cell may resemble a normal brain cell type.

Our study supported some key findings from previous scRNA-seq experiments in ependymomas. Gojo et al along with other studies identified radial glial like cells as potential cells of origin in ependymomas[305,484,485]. Our results corroborate this finding with an abundance of radial glial cells in our ependymoma samples. Moreover, Gojo et al indicate that stem-like cell populations are associated with more aggressive ependymomas[305]. Our results indicate a similar pattern in our expanded pediatric CNS tumor types, in which higher grade tumors are associated with cells with more stemlike features. Our study also supported results from Reitman et al, who demonstrated that pilocytic astrocytoma tumors are overall comprised of OPCs and mature glial-like cells[486]. Our results indicated a similar pattern in which much of our pilocytic astrocytoma samples were comprised of varying OPC clusters and couple of astrocyte-like clusters. The similarity of our results with previously published studies supports our results and previous findings in separate patient populations.

We identified the pathways enriched in varying cell types, with a focus on neural stem like cells. Since NSCs have been shown to be associated with therapy resistance, metastasis, and tumor malignancy, it is important to specifically consider NSCs when treating pediatric CNS tumors and reducing risk for secondary neoplasms[487–493]. We determined potential targetable NSC-specific pathways. While some commonly enriched pathways like MYC and FOXM1 in NSCs may be considered very difficult to target as MYC and transcription factors are considered to be less druggable, there were more easily targetable pathways enriched in NSCs like Aurora-B kinase and retinoic acid pathway.

With our cell type-adjusted approach, we addressed a critical confounder in differential gene expression analyses to identify transcriptomic alterations that exist in tumors compared to non-tumor tissue. Although the number of significantly differentially expressed genes decreased in the cell type-adjusted model compared to the cell type-unadjusted model, the adjusted model identified novel genes associated with tumors that would not have been uncovered in the unadjusted model. Moreover, the significantly differentially expressed genes exclusive to the unadjusted model likely stem from variations in cell type proportions, rather than from the underlying tumor biology that would be necessary for discovering effective therapeutic targets.

The pathways associated with the differentially expressed genes across the multiple tumor types in the cell type-adjusted model (translation associated processes like peptide chain elongation and translation initiation/termination along with nonsense mediated decay (NMD) processes) suggest the importance of these pathways commonly being dysregulated in pediatric central nervous system tumors. Previous studies have suggested the importance of downregulation of NMD responses in the differentiation of neural stem cells[494–496]. Moreover, high levels of NMD factors were sufficient to keep the stemness of neural stem cells[494]. Interestingly, our results indicate upregulation of NMD associated genes across all pediatric CNS tumor types in comparison to non-tumor pediatric brain which suggest the potential mechanism of upregulation of NMD maintaining more stem-like cells in these tumors. As more stem-like cells contributes to therapy resistance and recurrence, further studies investigating the NMD pathways and how they can be exploited to be potential therapeutic targets in pediatric CNS tumors are necessary.

Our study characterizes the heterogeneity that exists across pediatric CNS tumor types in comparison to non-tumoral pediatric brain tissue at the single cell level. We also identify potential tumor type and cell type-specific molecular characteristics that may be used therapeutic targets for the various pediatric CNS tumors from primary tissue samples. Although there were very limited samples for Schwannomas and glioblastoma, our study included thousands of nuclei from these tumor types to gain a better understanding of cells that exist in these tumor types that previous studies have not investigated yet. From our results, complementary preclinical *in vitro* and *in vivo* experiments are needed to validate these targets to advance these potential targets as therapeutic options in the clinic.

3.6. Author contributions

MKL and NA carried out the experiments. NA, LNN, GJZ obtained samples and clinical data. MKL performed data analyses with the help of HRF and BCC. JAS processed single cell level genotypes demultiplexing. BCC supervised the projects. All authors read and approved the final manuscript.

3.7. Acknowledgements

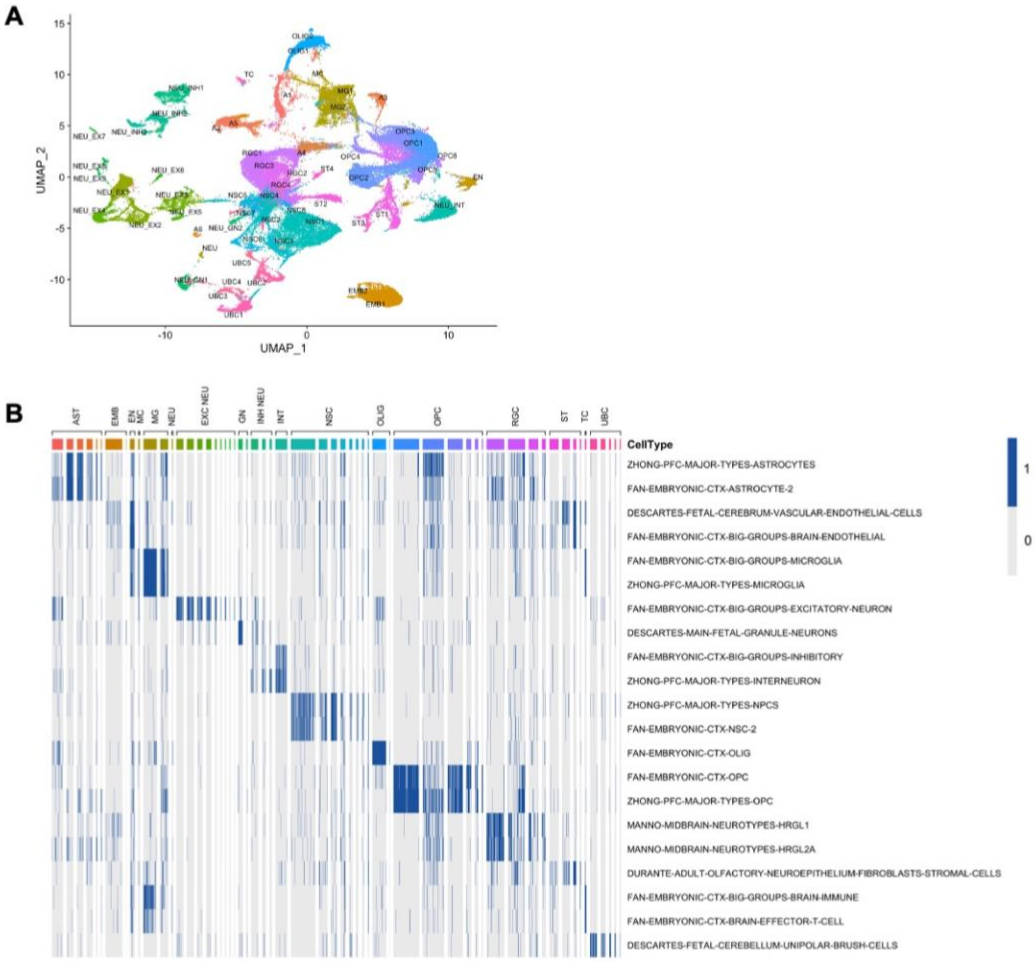
We would like to thank Dr. Matthew Havrda (Department of Molecular and Systems Biology, Geisel School of Medicine) for assistance in nuclei isolation protocol development. We would like to thank Ally Hawkins (Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation) for code review and Jaclyn Taroni (Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation) for assistance in data analysis. Research reported in this publication was supported through Geisel School of Medicine at Dartmouth's Center for Quantitative Biology through a grant from the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM130454

3.8. Funding sources

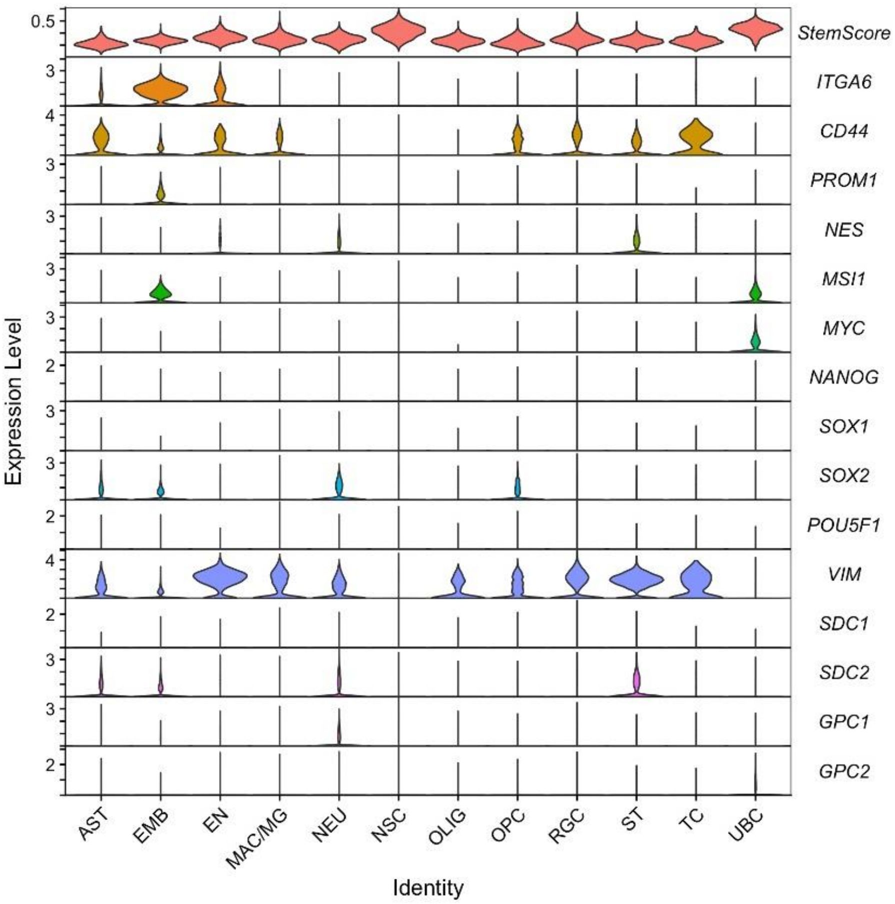
This work was supported by a Prouty Pilot award from the Dartmouth Cancer Center and a Single-cell Pediatric Cancer Atlas (ScPCA) grant from the Alex's Lemonade Stand Foundation. MKL was supported by the Burroughs-Wellcome Fund: Big Data in the Life Sciences at Dartmouth. NA was supported by the S.M. Tenney Fellowship at Dartmouth. This work was also supported by R01CA216265, R01CA253976, and P20GM104416 – 6369 to BCC and P20GM130454 – 7702 and R21CA253408 to HRF. Single nuclei RNA-seq experiments were conducted in the Genomics and Molecular Biology Shared

Resource (GMBSR) at Dartmouth, which is supported by NCI Cancer Center Support Grant 5P30CA023108 and NIH S10 (1S10OD030242) awards. Single nuclei RNA-experiments were also supported through the Dartmouth Center for Quantitative in collaboration with the GMBSR with support from NIGMS (P20GM130454) and NIH S10 (S10OD025235) awards.

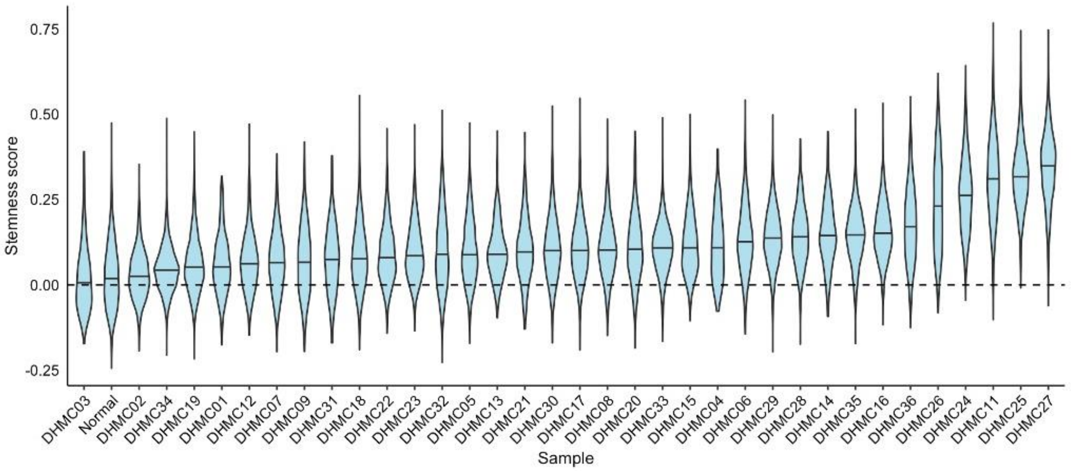
3.10. Supplemental materials



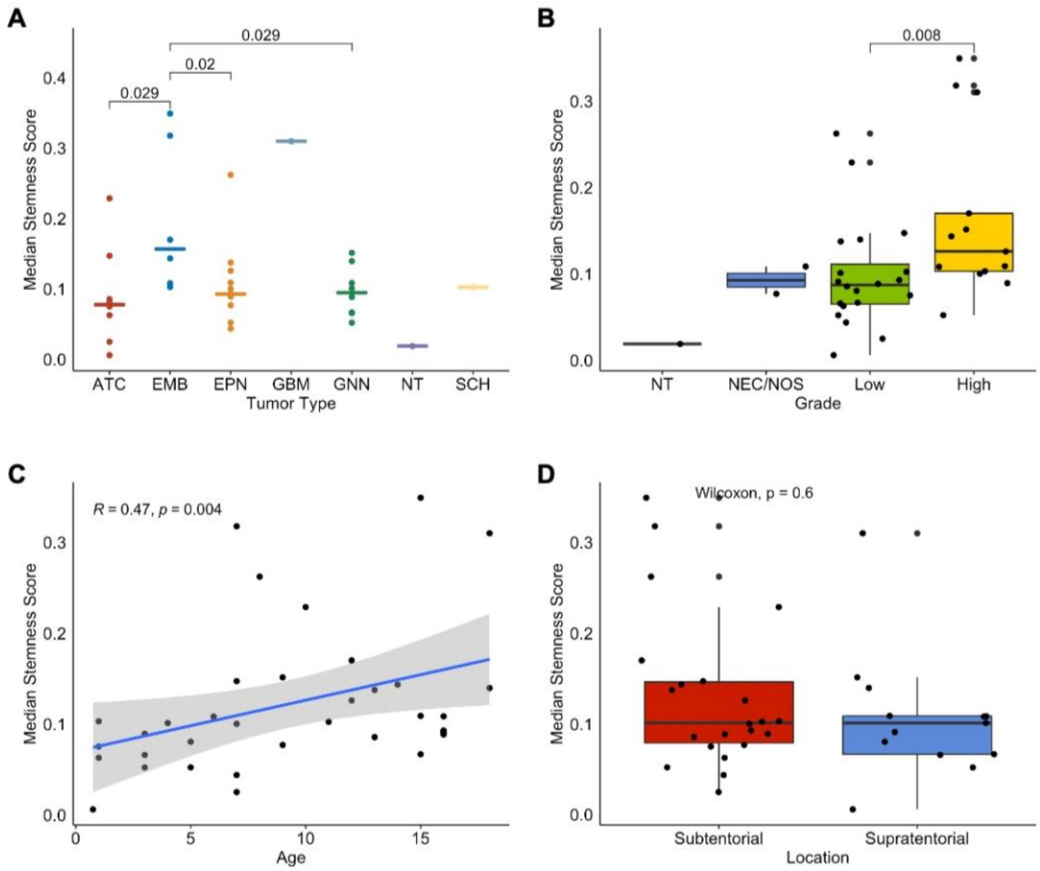
Supplementary Figure 3-1.
A) UMAP visualization of the 58 clusters identified through Seurat FindClusters. **B)** Heatmap of enrichment of cell type specific pathways for each nuclei. Tracking bar indicates cell cluster identity from the UMAP in **1A**.



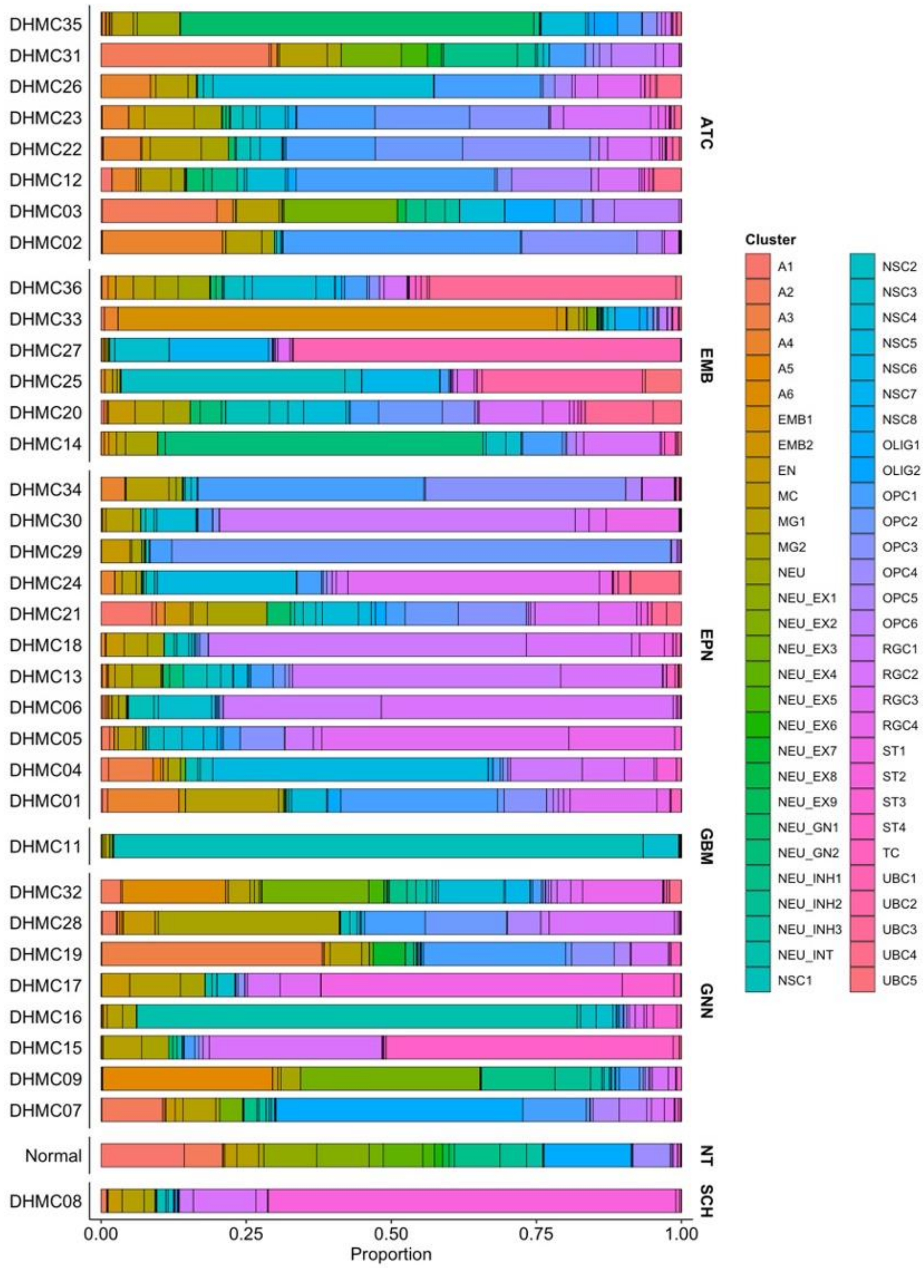
Supplementary Figure 3-2.
Expression levels of commonly used markers to isolate cancer stem cells and stemness score calculated from set of stem cell associated genes identified in Tirosh et al for each major cell type[468].



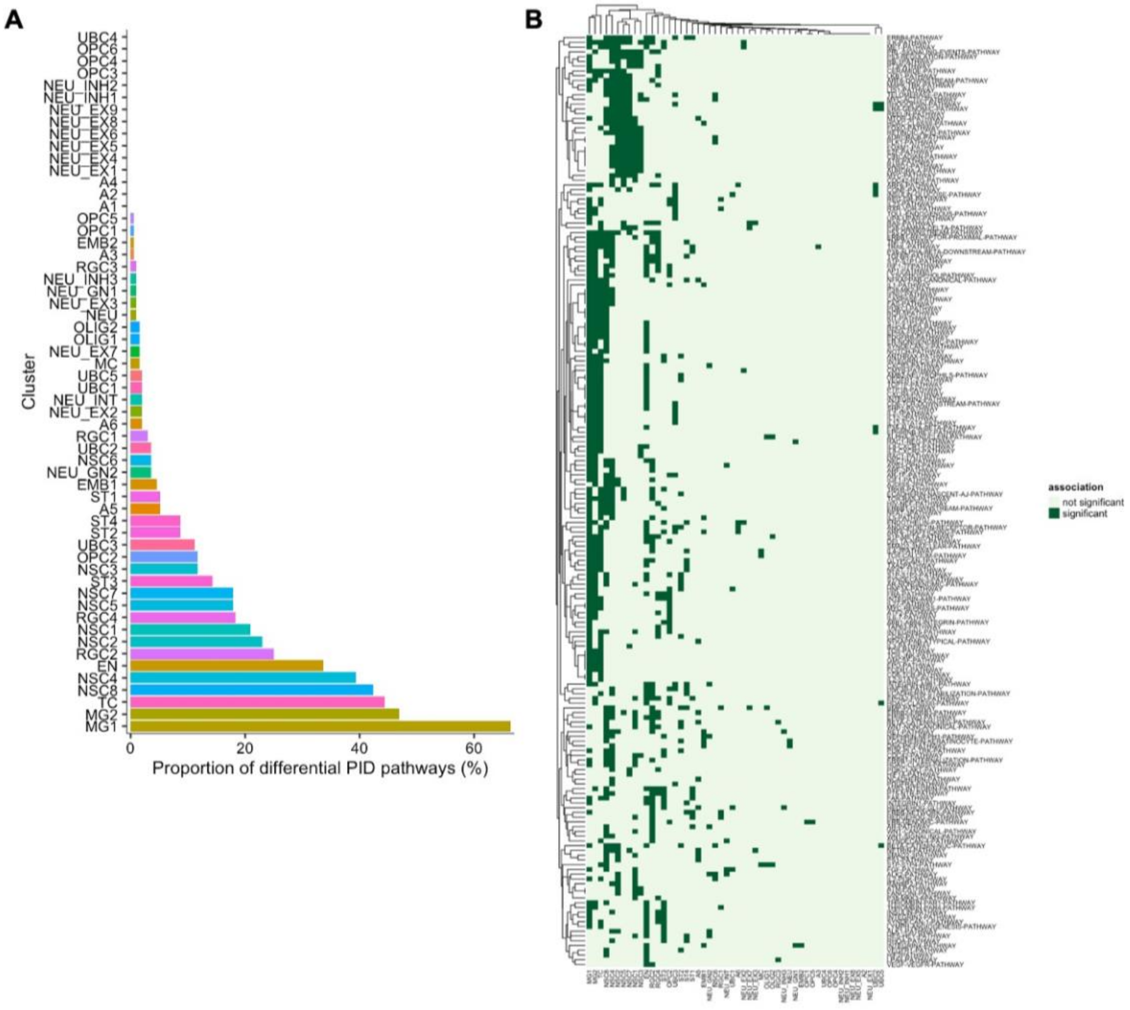
Supplementary Figure 3-3. Boxplot of the stemness scores of all nuclei for each sample.



Supplementary Figure 3-4.
A) Median stemness score distribution by tumor types. Horizontal line for each tumor type indicates median stemness score per tumor type. Comparisons between embryonal tumors and other tumor types were tested with Wilcoxon rank-sum test. **B)** Median stemness score distribution by grade. Comparison between median stemness scores of low and high grade conducted using Wilcoxon rank-sum test. **C)** Correlation between age at diagnosis and median stemness score of tumors. Correlation calculated using the Spearman rank method. Linear regression line and 95% confidence interval indicated by the blue line and gray band, respectively. **D)** Median stemness score distribution by tumor location class. Comparison between median stemness scores of subtentorial and supratentorial regions conducted using Wilcoxon rank-sum test.

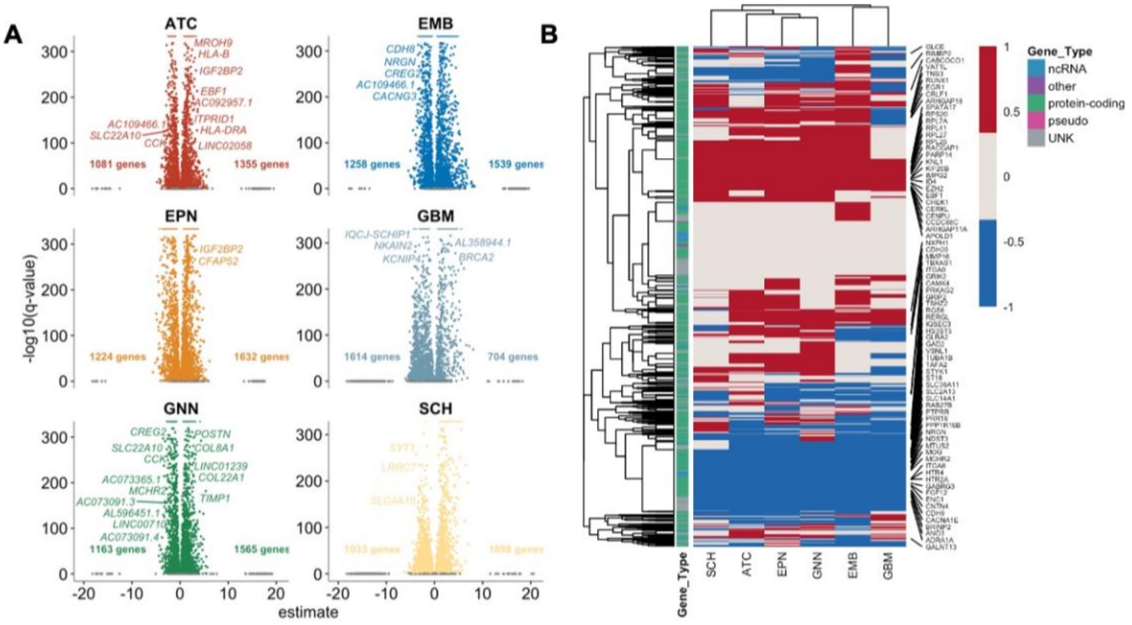


Supplementary Figure 3-5. Distribution of cell types present per sample, categorized by tumor types.
A: Astrocyte; **EMB**: Embryonal tumor cells; **EN**: Endothelial cells; **MC**: Macrophage; **MG**: Microglia; **NEU**: Neuron; **NEU_EX**: Excitatory neuron; **NEU_GN**: Granular neuron; **NEU_INH**: Inhibitory neuron; **NEU_INT**: Interneuron; **NSC**: Neural stem cell; **OLIG**: Oligodendrocyte; **OPC**: Oligodendrocyte precursor cell; **RGC**: Radial glial cell; **ST**: Stromal cell; **TC**: T cell; **UBC**: Unipolar brush cell.

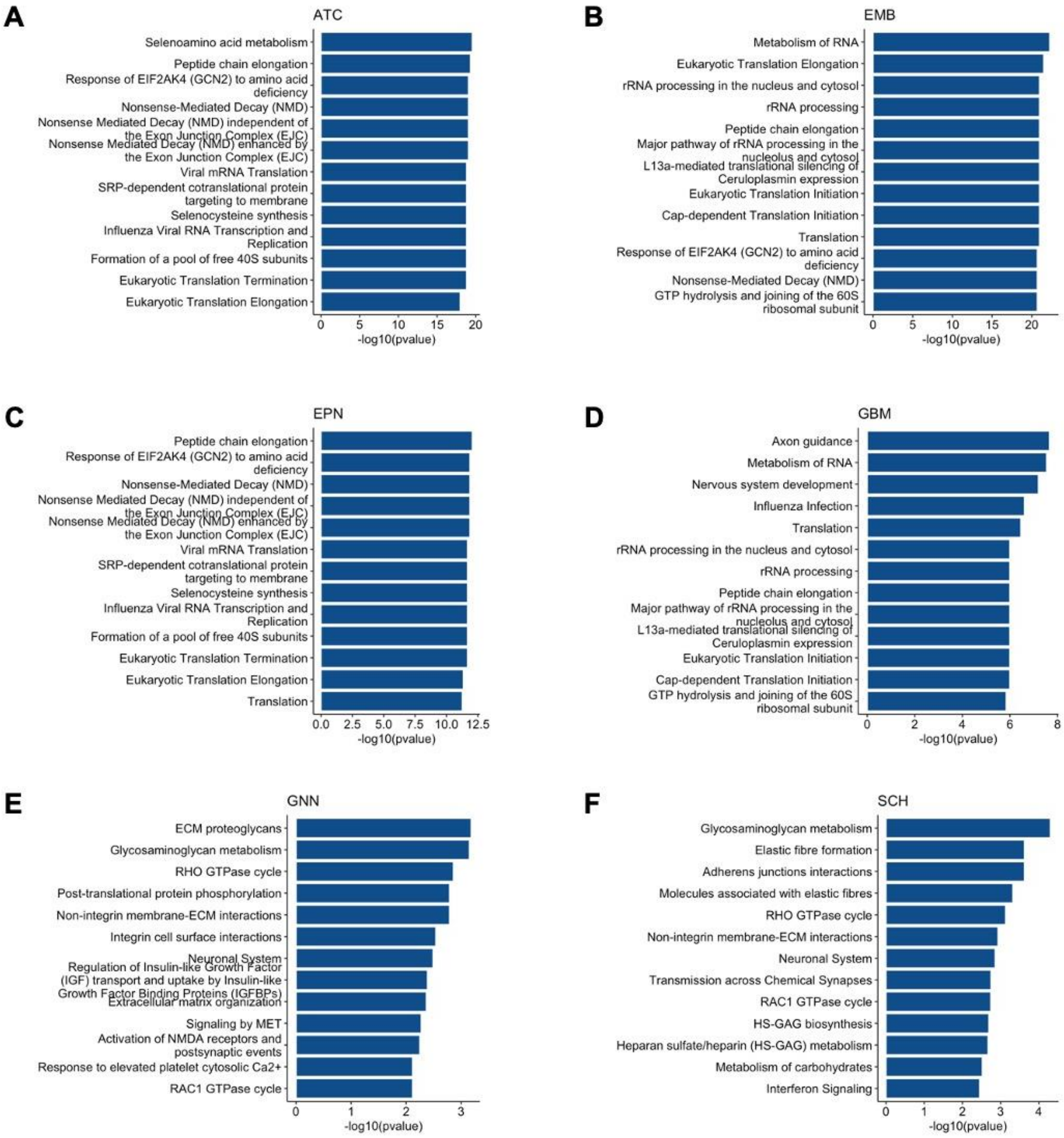


Supplementary Figure 3-6.

A) Proportion (out of 196 PID pathways tested) of pathways specific to cell types compared to all other nuclei. **B)** Hierarchical clustering of all 196 PID pathways tested for each cell type. Dark green indicates pathways relatively specific/important to the cell type.



Supplementary Figure 3-7.
A) Volcano plot of differentially expressed genes for each tumor type compared to non-tumor tissue, ***not*** adjusted for cell type. Number of genes on the left of the volcano plot indicate genes that are downregulated compared to non-tumor tissue. Number of genes on the right of the plot indicate genes that are upregulated compared to non-tumor tissue. **B)** Heatmap of differential expression direction and significance in all 4000 genes tested in the cell type unadjusted differential expression analyses. Red indicates significantly upregulated in the tumor type compared to non-tumor tissue. Blue indicates significantly downregulated in the tumor type compared to non-tumor tissue. Gray indicates the gene is not significantly differentially expressed. Tracking bar indicate the gene type.



Supplementary Figure 3-8.
Top 10 Reactome pathways associated with differentially expressed genes in **A)** astrocytoma, **B)** embryonal tumors, **C)** ependymoma, **D)** glioblastoma, **E)** glioneuronal/neuronal tumors, and **F)** Schwannoma.

Supplementary Table 3-1. Extended sample information

Sample Name	Sex	Age at diagnosis	LoLocation	Location Class	Tumor type	Grade	2021 WHO Diagnosis	Multi-seq Pool	# of Nuclei
DHMC01	M	5	Posterior Fossa	Subtentorial	Ependymoma	3	Posterior fossa ependymoma, NOS, CNS WHO grade 3	Pool18, Pool19	448
DHMC02	M	7	Posterior Fossa	Subtentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool18, Pool19	2819
DHMC03	M	0.75	Temporal Lobe	Supratentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool18, Pool19	476
DHMC04	F	15	Parietal Lobe	Supratentorial	Ependymoma	3	Supratentorial ependymoma, NOS, CNS WHO grade 3	Pool20, Pool21, Pool22	234
DHMC05	M	3	4th Ventricle	Subtentorial	Ependymoma	3	Posterior fossa ependymoma, NOS, CNS WHO grade 3	Pool20, Pool21, Pool22	1026
DHMC06	F	12	Posterior Fossa	Subtentorial	Ependymoma	3	Posterior fossa ependymoma, NOS, CNS WHO grade 3	Pool20, Pool21, Pool22	1761
DHMC07	M	3	Left Temporal Lobe	Supratentorial	Glioneuronal/Neuronal	1	Dysembryoplastic neuroepithelial tumor, CNS WHO grade 1	Pool20, Pool21, Pool22	886
DHMC08	M	11	Vestibular	Subtentorial	Schwannoma	1	Schwannoma, CNS WHO grade 1	Pool20, Pool21, Pool22	2501
DHMC09	M	15	Occipital Lobe	Supratentorial	Glioneuronal/Neuronal	1	Ganglioglioma, CNS WHO grade 1	Pool20, Pool21, Pool22	830
DHMC11	M	18	Frontal Lobe	Supratentorial	Glioblastoma	4	Pediatric-type diffuse high grade glioma, NOS, CNS WHO grade 4	NA	5795
DHMC12	F	1	Posterior Fossa	Subtentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool3, Pool4	482
DHMC13	F	16	Occipital Lobe	Supratentorial	Ependymoma	2	Supratentorial ependymoma, NOS, CNS WHO grade 2	Pool3, Pool4	543
DHMC14	M	14	Posterior Fossa	Subtentorial	Embryonal	4	Medulloblastoma, classic, CNS WHO grade 4	Pool3, Pool4	380
DHMC15	M	16	Occipital Lobe	Supratentorial	Glioneuronal/Neuronal	NEC	Desmoplastic ganglioglioma, NEC	Pool3, Pool4	558
DHMC16	F	9	Temporal Lobe	Supratentorial	Glioneuronal/Neuronal	3	Anaplastic ganglioglioma, NOS, CNS WHO grade 3	Pool5, Pool6	3514
DHMC17	M	4	Frontal Lobe	Supratentorial	Glioneuronal/Neuronal	1	Desmoplastic infantile ganglioglioma, CNS WHO grade 1	Pool5, Pool6	3865
DHMC18	M	9	Posterior Fossa	Subtentorial	Ependymoma	NOS	Posterior fossa ependymoma, NOS	Pool5, Pool6	3096
DHMC19	M	3	Temporal Lobe	Supratentorial	Glioneuronal/Neuronal	1	Dysembryoplastic neuroepithelial tumor, CNS WHO grade 1	NA	3610
DHMC20	F	1	Posterior Fossa	Subtentorial	Embryonal	4	Embryonal tumor with multilayered rosettes, NOS, CNS WHO grade 4	Pool9	410
DHMC21	F	16	4th Ventricle	Subtentorial	Ependymoma	1	Subependymoma, CNS WHO grade 1	Pool9	273
DHMC22	M	5	Suprasellar	Supratentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool9	851
DHMC23	F	13	Posterior Fossa	Subtentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool9	655
DHMC24	M	8	4th Ventricle	Subtentorial	Ependymoma	2	Posterior fossa ependymoma, NOS, CNS WHO grade 2	Pool10, Pool11	2001
DHMC25	M	7	4th Ventricle	Subtentorial	Embryonal	4	Medulloblastoma, classic, CNS WHO grade 4	Pool10, Pool11	3870
DHMC26	M	10	Posterior Fossa	Subtentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool10, Pool11	742
DHMC27	M	15	Posterior Fossa	Subtentorial	Embryonal	4	Medulloblastoma, desmoplastic/nodular, CNS WHO grade 4	Pool10, Pool11	2356
DHMC28	M	18	Lateral Ventricle	Supratentorial	Glioneuronal/Neuronal	1	Dysembryoplastic neuroepithelial tumor, CNS WHO grade 1	Pool12	1662
DHMC29	F	13	Spinal cord	Subtentorial	Ependymoma	2	Myxopapillary ependymoma, CNS WHO grade 2	Pool12	4963
DHMC30	M	7	Posterior Fossa	Subtentorial	Ependymoma	3	Posterior fossa ependymoma, NOS, CNS WHO grade 3	NA	2199
DHMC31	F	1	Posterior Fossa	Subtentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool15	290
DHMC32	M	16	Posterior Fossa	Subtentorial	Glioneuronal/Neuronal	1	Gangliocytoma, CNS WHO grade 1	NA	1091
DHMC33	F	6	Parieto-Temporal Lobe	Supratentorial	Embryonal	4	Embryonal tumor, NOS, CNS WHO grade 4	Pool1, Pool2	5167
DHMC34	F	7	Posterior Fossa	Subtentorial	Ependymoma	2	Posterior fossa ependymoma, NOS, CNS WHO grade 2	Pool1, Pool2	5743
DHMC35	M	7	Posterior Fossa	Subtentorial	Astrocytoma	1	Pilocytic astrocytoma, CNS WHO grade 1	Pool1, Pool2	1399
DHMC36	F	12	Posterior Fossa	Subtentorial	Embryonal	4	Medulloblastoma, classic, CNS WHO grade 4	Pool1, Pool2	753
Normal	B	NA	NA	Supratentorial	Non-Tumor		Non-Tumor		

Supplementary Table 3-2. Enriched pathways per cell types in tumor cells

Pathway	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster
PID-PI3KCI-AKT-PATHWAY	0.007658518	0.024156514	0.016	0.006	1	A2
PID-TRAIL-PATHWAY	3.74E-11	0.054566727	0.04	0.017	7.34E-09	A3
PID-EPHA2-FWD-PATHWAY	0.000715487	0.022841315	0.02	0.011	0.14023544	A3
PID-HEDGEHOG-GLI-PATHWAY	0.00171415	0.017702064	0.015	0.007	0.335973395	A3
PID-BETA-CATENIN-DEG-PATHWAY	2.65E-11	0.096605995	0.051	0.01	5.19E-09	A5
PID-MTOR-4PATHWAY	6.88E-09	0.086394285	0.047	0.01	1.35E-06	A5
PID-NFKAPPAB-CANONICAL-PATHWAY	1.66E-08	0.100912224	0.058	0.015	3.25E-06	A5
PID-HEDGEHOG-GLI-PATHWAY1	3.78E-07	0.065879318	0.035	0.007	7.42E-05	A5
PID-NFKAPPAB-ATYPICAL-PATHWAY	4.45E-07	0.07130021	0.039	0.009	8.72E-05	A5
PID-PS1-PATHWAY	8.13E-07	0.059421707	0.031	0.006	0.000159414	A5
PID-NCADHERIN-PATHWAY	5.67E-06	0.017647291	0.008	0.001	0.001111109	A5
PID-ERBB2-ERBB3-PATHWAY	1.37E-05	0.045207846	0.023	0.005	0.00268527	A5
PID-SMAD2-3PATHWAY	2.69E-05	0.017425519	0.008	0.001	0.005265678	A5
PID-NETRIN-PATHWAY	7.32E-05	0.02465997	0.012	0.002	0.014348044	A5
PID-HDAC-CLASSI-PATHWAY	0.000336305	0.056629323	0.035	0.011	0.065915788	A5
PID-NECTIN-PATHWAY	0.001393904	0.044683716	0.027	0.009	0.27320518	A5
PID-HEDGEHOG-2PATHWAY	0.003983186	0.037957699	0.023	0.008	0.780704363	A5
PID-PDGFRA-PATHWAY	0.007295236	0.015652566	0.008	0.001	1	A5
PID-ARF6-TRAFFICKING-PATHWAY	1.11E-38	0.222008333	0.111	0.012	2.18E-36	A6
PID-ARF6-PATHWAY	3.61E-13	0.121258297	0.062	0.011	7.08E-11	A6
PID-ENDOTHELIN-PATHWAY	1.63E-05	0.097877997	0.062	0.02	0.003192042	A6
PID-ANGIOPOIETIN-RECEPTOR-PATHWAY	5.89E-05	0.084290016	0.053	0.017	0.01154192	A6
PID-PDGFRA-PATHWAY1	0.001866883	0.020132835	0.01	0.001	0.365909031	A6
PID-HEDGEHOG-GLI-PATHWAY2	0.006149856	0.03983906	0.024	0.008	1	A6

Supplementary Table 3-3. Example of differential expression test results from both cell type-adjusted and unadjusted models.

Gene	Tumor Type	Adjusted Estimate	Adjusted q-value	Unadjusted estimate	Unadjusted q-value	Overlap in models
A2M	ATC	1.106065	8.04E-118	1.501315922	4.01E-176	Both
A2M	EMB	0.1749618	1	0.172147367	1	NA
A2M	EPN	0.5738485	1.57E-32	1.144053387	6.47E-123	Both
A2M	GBM	-2.01805	4.19E-37	-1.651417139	8.31E-18	Both
A2M	GNN	0.7766881	5.96E-66	1.659783074	2.70E-268	Both
A2M	SCH	-0.915399	3.56E-21	0.416544731	0.070211267	Adjusted
ABCA10	ATC	-0.09970278	1	-0.2752297	3.13E-06	Unadjusted
ABCA10	EMB	-0.1941893	1	-0.6816814	3.32E-51	Unadjusted
ABCA10	EPN	0.1890242	0.3655263	-0.002706531	1	NA
ABCA10	GBM	1.592022	4.37E-131	1.364205	0	Both
ABCA10	GNN	-0.1871171	0.2973362	-0.01963009	1	NA
ABCA10	SCH	-1.730696	5.15E-45	-1.100408	2.78E-21	Both
ABCA12	ATC	0.9356376	2.14E-09	0.5135202	0.002618384	Both
ABCA12	EMB	0.394355	1	0.7119486	4.00E-12	Unadjusted
ABCA12	EPN	0.01071014	1	-0.7342798	1.27E-07	Unadjusted
ABCA12	GBM	0.5144688	1	-0.3152585	1	NA
ABCA12	GNN	0.06792107	1	-0.4845534	0.03697754	Unadjusted
ABCA12	SCH	-0.4401558	1	-1.327338	0.4535809	NA
ABCA13	ATC	1.964646	3.85E-13	1.809611854	4.57E-12	Both
ABCA13	EMB	1.781864	1.53E-07	1.588221844	8.89E-10	Both
ABCA13	EPN	1.996186	3.89E-14	1.624995348	1.39E-11	Both
ABCA13	GBM	0.3813485	1	0.072782054	1	NA
ABCA13	GNN	2.365599	5.69E-25	2.150861523	5.26E-22	Both
ABCA13	SCH	2.093733	8.15E-06	1.357772693	0.055815271	Adjusted

Supplementary Table 3-4. Top 20 pathways associated with differentially expressed genes in astrocytoma in the cell type adjusted model

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
R-HSA-2408522	Selenoamino acid metabolism	82/932	82/1584	2.78E-20	4.93E-18	4.29E-18	82
R-HSA-156902	Peptide chain elongation	81/932	81/1584	4.91E-20	4.93E-18	4.29E-18	81
R-HSA-927802	Nonsense-Mediated Decay (NMD)	80/932	80/1584	8.67E-20	4.93E-18	4.29E-18	80
R-HSA-9633012	Response of EIF2AK4 (GCN2) to amino acid deficiency	80/932	80/1584	8.67E-20	4.93E-18	4.29E-18	80
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	80/932	80/1584	8.67E-20	4.93E-18	4.29E-18	80
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	80/932	80/1584	8.67E-20	4.93E-18	4.29E-18	80
R-HSA-168273	Influenza Viral RNA Transcription and Replication	79/932	79/1584	1.53E-19	4.93E-18	4.29E-18	79
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	79/932	79/1584	1.53E-19	4.93E-18	4.29E-18	79
R-HSA-192823	Viral mRNA Translation	79/932	79/1584	1.53E-19	4.93E-18	4.29E-18	79
R-HSA-2408557	Selenocysteine synthesis	79/932	79/1584	1.53E-19	4.93E-18	4.29E-18	79
R-HSA-72689	Formation of a pool of free 40S subunits	79/932	79/1584	1.53E-19	4.93E-18	4.29E-18	79
R-HSA-72764	Eukaryotic Translation Termination	79/932	79/1584	1.53E-19	4.93E-18	4.29E-18	79
R-HSA-156842	Eukaryotic Translation Elongation	82/932	83/1584	1.02E-18	3.03E-17	2.64E-17	82
R-HSA-72766	Translation	86/932	88/1584	2.12E-18	5.87E-17	5.11E-17	86
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	80/932	81/1584	3.09E-18	5.98E-17	5.21E-17	80
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	80/932	81/1584	3.09E-18	5.98E-17	5.21E-17	80
R-HSA-72312	rRNA processing	80/932	81/1584	3.09E-18	5.98E-17	5.21E-17	80
R-HSA-72613	Eukaryotic Translation Initiation	80/932	81/1584	3.09E-18	5.98E-17	5.21E-17	80
R-HSA-72737	Cap-dependent Translation Initiation	80/932	81/1584	3.09E-18	5.98E-17	5.21E-17	80

Supplementary Table 3-5. Top 20 pathways associated with differentially expressed genes in embryonal tumors in the cell type adjusted model

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
R-HSA-8953854	Metabolism of RNA	97/891	99/1584	5.52E-23	2.13E-20	1.90E-20	97
R-HSA-156842	Eukaryotic Translation Elongation	83/891	83/1584	3.14E-22	4.20E-20	3.75E-20	83
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	81/891	81/1584	1.08E-21	4.20E-20	3.75E-20	81
R-HSA-156902	Peptide chain elongation	81/891	81/1584	1.08E-21	4.20E-20	3.75E-20	81
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	81/891	81/1584	1.08E-21	4.20E-20	3.75E-20	81
R-HSA-72312	rRNA processing	81/891	81/1584	1.08E-21	4.20E-20	3.75E-20	81
R-HSA-72613	Eukaryotic Translation Initiation	81/891	81/1584	1.08E-21	4.20E-20	3.75E-20	81
R-HSA-72737	Cap-dependent Translation Initiation	81/891	81/1584	1.08E-21	4.20E-20	3.75E-20	81
R-HSA-8868773	rRNA processing in the nucleus and cytosol	81/891	81/1584	1.08E-21	4.20E-20	3.75E-20	81
R-HSA-72766	Translation	87/891	88/1584	1.09E-21	4.20E-20	3.75E-20	87
R-HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	80/891	80/1584	2.01E-21	5.18E-20	4.62E-20	80
R-HSA-927802	Nonsense-Mediated Decay (NMD)	80/891	80/1584	2.01E-21	5.18E-20	4.62E-20	80
R-HSA-9633012	Response of EIF2AK4 (GCN2) to amino acid deficiency	80/891	80/1584	2.01E-21	5.18E-20	4.62E-20	80
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	80/891	80/1584	2.01E-21	5.18E-20	4.62E-20	80
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	80/891	80/1584	2.01E-21	5.18E-20	4.62E-20	80
R-HSA-168273	Influenza Viral RNA Transcription and Replication	79/891	79/1584	3.72E-21	6.86E-20	6.12E-20	79
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	79/891	79/1584	3.72E-21	6.86E-20	6.12E-20	79
R-HSA-192823	Viral mRNA Translation	79/891	79/1584	3.72E-21	6.86E-20	6.12E-20	79
R-HSA-2408557	Selenocysteine synthesis	79/891	79/1584	3.72E-21	6.86E-20	6.12E-20	79

Supplementary Table 3-6. Top 20 pathways associated with differentially expressed genes in ependymomas in the cell type adjusted model

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count
R-HSA-156902	Peptide chain elongation	78/1009	81/1584	8.61E-13	7.32E-11	6.25E-11	78
R-HSA-927802	Nonsense-Mediated Decay (NMD)	77/1009	80/1584	1.34E-12	7.32E-11	6.25E-11	77
R-HSA-9633012	Response of EIF2AK4 (GCN2) to amino acid deficiency	77/1009	80/1584	1.34E-12	7.32E-11	6.25E-11	77
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	77/1009	80/1584	1.34E-12	7.32E-11	6.25E-11	77
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	77/1009	80/1584	1.34E-12	7.32E-11	6.25E-11	77
R-HSA-168273	Influenza Viral RNA Transcription and Replication	76/1009	79/1584	2.08E-12	7.32E-11	6.25E-11	76
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	76/1009	79/1584	2.08E-12	7.32E-11	6.25E-11	76
R-HSA-192823	Viral mRNA Translation	76/1009	79/1584	2.08E-12	7.32E-11	6.25E-11	76
R-HSA-2408557	Selenocysteine synthesis	76/1009	79/1584	2.08E-12	7.32E-11	6.25E-11	76
R-HSA-72689	Formation of a pool of free 40S subunits	76/1009	79/1584	2.08E-12	7.32E-11	6.25E-11	76
R-HSA-72764	Eukaryotic Translation Termination	76/1009	79/1584	2.08E-12	7.32E-11	6.25E-11	76
R-HSA-156842	Eukaryotic Translation Elongation	79/1009	83/1584	4.46E-12	1.44E-10	1.23E-10	79
R-HSA-72766	Translation	83/1009	88/1584	5.43E-12	1.62E-10	1.38E-10	83
R-HSA-2408522	Selenoamino acid metabolism	78/1009	82/1584	6.86E-12	1.90E-10	1.62E-10	78
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	77/1009	81/1584	1.05E-11	2.04E-10	1.74E-10	77
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	77/1009	81/1584	1.05E-11	2.04E-10	1.74E-10	77
R-HSA-72312	rRNA processing	77/1009	81/1584	1.05E-11	2.04E-10	1.74E-10	77
R-HSA-72613	Eukaryotic Translation Initiation	77/1009	81/1584	1.05E-11	2.04E-10	1.74E-10	77
R-HSA-72737	Cap-dependent Translation Initiation	77/1009	81/1584	1.05E-11	2.04E-10	1.74E-10	77

Supplementary Table 3-7. Top 20 pathways associated with differentially expressed genes in glioneuronal/neuronal tumors in the cell type adjusted model

ID	Description	Gene Ratio	BgRatio	pvalue	p.adjust	qvalue	Count
R-HSA-3000178	ECM proteoglycans	31/955	36/1584	0.00065832	0.12592627	0.12056588	31
R-HSA-1630316	Glycosaminoglycan metabolism	28/955	32/1584	0.00070802	0.12592627	0.12056588	28
R-HSA-9012999	RHO GTPase cycle	55/955	71/1584	0.00138515	0.12592627	0.12056588	55
R-HSA-3000171	Non-integrin membrane-ECM interactions	23/955	26/1584	0.00162695	0.12592627	0.12056588	23
R-HSA-8957275	Post-translational protein phosphorylation	23/955	26/1584	0.00162695	0.12592627	0.12056588	23
R-HSA-216083	Integrin cell surface interactions	32/955	39/1584	0.00288836	0.17919434	0.17156645	32
R-HSA-112316	Neuronal System	84/955	116/1584	0.00324124	0.17919434	0.17156645	84
R-HSA-381426	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	26/955	31/1584	0.00408903	0.18469701	0.17683489	26
R-HSA-1474244	Extracellular matrix organization	73/955	100/1584	0.00429528	0.18469701	0.17683489	73
R-HSA-6806834	Signaling by MET	20/955	23/1584	0.00531138	0.19815172	0.18971686	20
R-HSA-442755	Activation of NMDA receptors and postsynaptic events	17/955	19/1584	0.00563222	0.19815172	0.18971686	17
R-HSA-76005	Response to elevated platelet cytosolic Ca2+	29/955	36/1584	0.00762987	0.22713544	0.21746682	29
R-HSA-9013149	RAC1 GTPase cycle	29/955	36/1584	0.00762987	0.22713544	0.21746682	29
R-HSA-8875878	MET promotes cell motility	16/955	18/1584	0.00848635	0.23458695	0.22460114	16
R-HSA-71387	Metabolism of carbohydrates	35/955	45/1584	0.00955902	0.23601299	0.22596647	35
R-HSA-373760	L1CAM interactions	21/955	25/1584	0.00975764	0.23601299	0.22596647	21
R-HSA-114608	Platelet degranulation	28/955	35/1584	0.01042774	0.23685043	0.22676826	28
R-HSA-112315	Transmission across Chemical Synapses	57/955	78/1584	0.0110163	0.23685043	0.22676826	57
R-HSA-3000157	Laminin interactions	15/955	17/1584	0.01271289	0.24599446	0.23552305	15

Supplementary Table 3-8. Top 20 pathways associated with differentially expressed genes in glioblastoma in the cell type adjusted model

ID	DESCRIPTION	GENERATIO	BGRATIO	PVALUE	P.ADJUST	QVALUE	COUNT
R-HSA-422475	Axon guidance	143/990	177/1584	2.18E-08	5.58E-06	4.62E-06	143
R-HSA-8953854	Metabolism of RNA	86/990	99/1584	2.89E-08	5.58E-06	4.62E-06	86
R-HSA-9675108	Nervous system development	149/990	187/1584	6.36E-08	8.21E-06	6.79E-06	149
R-HSA-168255	Influenza Infection	74/990	85/1584	2.45E-07	2.37E-05	1.96E-05	74
R-HSA-72766	Translation	76/990	88/1584	3.49E-07	2.70E-05	2.23E-05	76
R-HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression	70/990	81/1584	1.01E-06	3.26E-05	2.70E-05	70
R-HSA-156902	Peptide chain elongation	70/990	81/1584	1.01E-06	3.26E-05	2.70E-05	70
R-HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol	70/990	81/1584	1.01E-06	3.26E-05	2.70E-05	70
R-HSA-72312	rRNA processing	70/990	81/1584	1.01E-06	3.26E-05	2.70E-05	70
R-HSA-72613	Eukaryotic Translation Initiation	70/990	81/1584	1.01E-06	3.26E-05	2.70E-05	70
R-HSA-72737	Cap-dependent Translation Initiation	70/990	81/1584	1.01E-06	3.26E-05	2.70E-05	70
R-HSA-8868773	rRNA processing in the nucleus and cytosol	70/990	81/1584	1.01E-06	3.26E-05	2.70E-05	70
R-HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit	69/990	80/1584	1.43E-06	3.41E-05	2.82E-05	69
R-HSA-927802	Nonsense-Mediated Decay (NMD)	69/990	80/1584	1.43E-06	3.41E-05	2.82E-05	69
R-HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	69/990	80/1584	1.43E-06	3.41E-05	2.82E-05	69
R-HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	69/990	80/1584	1.43E-06	3.41E-05	2.82E-05	69
R-HSA-156842	Eukaryotic Translation Elongation	71/990	83/1584	1.97E-06	3.41E-05	2.82E-05	71
R-HSA-168273	Influenza Viral RNA Transcription and Replication	68/990	79/1584	2.03E-06	3.41E-05	2.82E-05	68
R-HSA-1799339	SRP-dependent cotranslational protein targeting to membrane	68/990	79/1584	2.03E-06	3.41E-05	2.82E-05	68

Supplementary Table 3-9. Pathways associated with differentially expressed genes in schwannomas in the cell type adjusted model

ID	DESCRIPTION	GENERATIO	BGRATIO	PVALUE	P.ADJUST	QVALUE	COUNT
R-HSA-1630316	Glycosaminoglycan metabolism	27/793	32/1584	5.00E-05	0.01935523	0.01842603	27
R-HSA-1566948	Elastic fibre formation	12/793	12/1584	0.0002377	0.03066365	0.02919156	12
R-HSA-418990	Adherens junctions interactions	12/793	12/1584	0.0002377	0.03066365	0.02919156	12
R-HSA-2129379	Molecules associated with elastic fibres	11/793	11/1584	0.00047814	0.04626015	0.04403931	11
R-HSA-9012999	RHO GTPase cycle	49/793	71/1584	0.00074078	0.05733614	0.05458357	49
R-HSA-3000171	Non-integrin membrane-ECM interactions	21/793	26/1584	0.00117638	0.07522604	0.07161461	21
R-HSA-112316	Neuronal System	74/793	116/1584	0.00139126	0.07522604	0.07161461	74
R-HSA-112315	Transmission across Chemical Synapses	52/793	78/1584	0.00178807	0.07522604	0.07161461	52
R-HSA-9013149	RAC1 GTPase cycle	27/793	36/1584	0.00181729	0.07522604	0.07161461	27
R-HSA-2022928	HS-GAG biosynthesis	14/793	16/1584	0.002034	0.07522604	0.07161461	14
R-HSA-1638091	Heparan sulfate/heparin (HS-GAG) metabolism	16/793	19/1584	0.00213821	0.07522604	0.07161461	16
R-HSA-71387	Metabolism of carbohydrates	32/793	45/1584	0.00301595	0.09458205	0.09004139	32
R-HSA-913531	Interferon Signaling	27/793	37/1584	0.00355974	0.09458205	0.09004139	27
R-HSA-421270	Cell-cell junction organization	15/793	18/1584	0.00366597	0.09458205	0.09004139	15
R-HSA-8986944	Transcriptional Regulation by MECP2	15/793	18/1584	0.00366597	0.09458205	0.09004139	15
R-HSA-1474244	Extracellular matrix organization	63/793	100/1584	0.0049286	0.11921041	0.1134874	63
R-HSA-9013404	RAC2 GTPase cycle	12/793	14/1584	0.0063644	0.13817936	0.13154569	12
R-HSA-8980692	RHOA GTPase cycle	24/793	33/1584	0.00642695	0.13817936	0.13154569	24
R-HSA-442755	Activation of NMDA receptors and postsynaptic events	15/793	19/1584	0.00939508	0.191363	0.18217612	15

Chapter 4

4. Hydroxymethylation alterations in progenitor-like cell types of pediatric central nervous system tumors are associated with cell type-specific transcriptional changes

The following authors contributed to the work:

Min Kyung Lee, Nasim Azizgolshani, Ze Zhang, Laurent Perreard , Fred W. Kolling, Lananh N. Nguyen, George J. Zanazzi, Lucas A. Salas, Brock C. Christensen

4.1. Abstract

Although intratumoral heterogeneity has been established in pediatric central nervous system tumors, epigenomic alterations at the cell type level have largely remained unresolved. To identify cell type-specific alterations to cytosine modifications in pediatric central nervous system tumors we utilized a multi-omic approach that integrated bulk DNA cytosine modification data (methylation and hydroxymethylation) with both bulk and single-cell RNA-sequencing data. We demonstrate a large reduction in the scope of significantly differentially modified cytosines in tumors when accounting for tumor cell type composition. In the progenitor-like cell types of tumors, we identified a preponderance differential CpG hydroxymethylation rather than methylation. Genes with differential hydroxymethylation, like *HDAC4* and *IGF1R,* were associated with cell type-specific changes in gene expression in tumors. Our results highlight the importance of epigenomic alterations in the progenitor-like cell types and its role in cell type-specific transcriptional regulation in pediatric CNS tumors.

4.2. Introduction

Central nervous system (CNS) tumors are the leading cause of cancer death in the pediatric population[334]. While major progress has been made in reducing the mortality in pediatric cancers in the past few decades, the magnitude of reduction in the mortality rate of CNS tumors have not been as substantial[346]. Even among patients who survive childhood cancers, those who have survived CNS tumors have the highest cumulative burden of disease post-survival[337]. Craniospinal radiation and neuro-toxic therapy are major risk factors for the future burden on quality of life with late effects including neurocognitive impairments such as academic and memory decline, and adverse health outcomes like abnormal hearing and growth hormone deficiency[338,339,342,497–499]. Efforts to address discrepancies in the reduction of mortality rates and extensive chronic health burdens later in life have been made with the recent advances in technology that have allowed for better insight into the molecular characterization of pediatric CNS tumors[305,351–354,436,437,443,444,486,500–502]. Molecular biomarkers are progressively being incorporated into the diagnosis and management of certain pediatric CNS tumor types[336].

One method to supplementally diagnose and subtype CNS tumors is DNA methylation[503]. Capper et al. developed a classification method to address previous issues in inter-observer variability for histopathological diagnosis of many CNS tumors[503]. Since the development of this method, DNA methylation classification is now used regularly for certain pediatric CNS tumor types, like ependymomas, to understand the prognosis and manage treatment decisions[353,354]. This method utilizes bisulfite-treated DNA, which does not distinguish between 5-methylcytosine and 5-hydroxymethylcytosine, although it has been indicated only 5-methylcytosine signal from oxidative bisulfite-treated DNA alters the classification from this method[183,243]. Moreover, while advancements have improved management strategies for some tumor types, many other pediatric CNS tumor types remain underexplored.

DNA methylation is one of the most well-studied epigenomic marks, primarily known for its role in regulating gene expression. DNA methylation occurs when a methyl group is added to the 5-carbon position of a cytosine in the context of a Cytosine-phosphate-Guanine (CpG) dinucleotides by DNA methyltransferases (DNMTs)[50,504–508]. Methylation of CpG island promoters is associated with repression of gene expression while methylation of gene bodies is associated with activation of gene expression[47,114,509]. 5-methylcytosine (5-mC) many times co-exist with H3K9me3 marks and do not overlap with H3K4me3 marks and H2A.Z[47,510,511]. In addition, DNA methylation marks function as genome stabilizers by silencing transposable elements[47,104]. The main ways DNA methylation is altered in cancer include genome-wide hypomethylation in repetitive elements like retrotransposable elements[220,512], hypermethylation of promoters[220–223], and propensity for cytosines in CpG contexts to be mutated[194,224–226].

Cytosines can also remain in a hydroxymethylated state (5-hydroxymethylcytosine, 5-hmC). 5-hmC is formed when 5-mC is actively being demethylated by ten-eleven translocation (TET) enzymes[143,174,513]. TET enzymes add a hydroxyl group onto the methyl group to become 5-hydroxymethylcytosine, then add the hydroxyl group again to become 5-formylcytosine, then again to become 5-carboxylcytosine, which is excised to become unmethylated[143,174,513,514]. While 5-hmC is an intermediate, it has been shown to have functional roles and be stable in the genome. Like 5-mC, 5-hmC has been associated with regulating transcription. It is enriched in gene bodies of active genes and in transcription start sites in which promoters are marked with H3K27me3 and H3K4me4[158,175]. 5-hmC has also been shown to play roles in maintaining pluripotency and tumorigenesis[175,515]. While generally 5-hmC levels are relatively much lower than 5-mC levels, higher levels of 5-hmC are found in the brain tissue compared to other tissue and in embryonal stem cells developmentally programmed neuronal cells[144,157,159,175,516–519]. Although progress has been made since the discovery of TET enzymes producing 5-hmC[143,513,514], more investigation is needed to understand the functional roles of 5-hmC. While alterations in hydroxymethylation patterns have not been as well examined, studies have indicated decreased hydroxymethylation across the genome in a variety of tumor types including adult and pediatric CNS tumors[178,230,231,233,239,241,243,515,520–522], and mutations in hydroxymethylation-associated genes such as *IDH1/2* and *TET1/2/3* have been associated with certain tumor types like gliomas and acute myeloid leukemia[178,201,523–525].

Numerous studies have established that brain tumors display intratumoral cellular heterogeneity[302–305,437,443,526–533]. While it is known that both DNA methylation and hydroxymethylation patterns are tissue type and cell type dependent[70,89,158,175,534–536], limited research has addressed cell type-specific DNA cytosine modification alterations in these tumors. This gap exists largely due to the high cost and limitations in technologies to profile cytosine modifications at the cell type-specific scale[36]. While the importance of cell type composition effects in epigenome-wide association studies has been well documented[84,537–540], single-cell methylation profiling strategies[17,21,541,542] are slowly developing in comparison to more accessible and commercially available genome profiling technologies focused on gene expression or chromatin accessibility. To address these shortcomings, computational methods have been developed to deconvolute cell type composition using DNA methylation for certain tissue types[31,32,34,36,37,543–547]. While these methods have greatly improved our understanding of the cell type composition effects on many epigenome-wide association studies, they have not been utilized in investigating cell type composition effects on brain tumors due to some limited applicability in brain tissue.

In this study, we use a multi-omic approach to study cell type-level epigenomic alterations in pediatric CNS tumors to maximize the applicability of currently available methods. By integrating single nuclei RNA-seq and cytosine modification data, we provide a more complete picture of the cytosine modification alterations associated with pediatric CNS types and cytosine modifications that are associated with changes in transcription at the cell type level in pediatric CNS tumors.

4.3. Methods

Sample information

Cytosine modifications, bulk tissue gene expression, and single nuclei gene expression were measured in 32 pediatric CNS tumors of various types and 2 non-tumor pediatric brain tissue (**Table 4-1, Supplementary Table 4-1**). This study was approved by the Institutional Review Board Study #00030211. Only samples with all four molecular measurements were included in downstream analyses. The samples were collected from patients being treated at Dartmouth-Hitchcock Medical Center and the Dartmouth Cancer Center from 1993 to 2017. For each tumor type, the number of samples was distributed evenly with 8 samples for astrocytoma, 6 for embryonal tumors, 10 for ependymoma, and 8 for glioneuronal/neuronal tumors. Pathological re-review for the histopathologic tumor type and grade were done according to the 2021 World Health Organization CNS tumor classification system, then categorized into broader tumor types. The non-tumor pediatric brain tissues were obtained from patients who underwent surgical resection for epilepsy.

Table 4-1. Subject demographics.

		Tumor types				
	Total (N=34)	**Astrocytoma (N=8)**	**Embryonal (N=6)**	**Ependymoma (N=10)**	**Glioneuronal/ neuronal (N=8)**	**Non-Tumor (N=2)**
Sex						
F	14 (41 %)	3 (38 %)	3 (50 %)	5 (50 %)	1 (12 %)	2 (100 %)
M	20 (59 %)	5 (62 %)	3 (50 %)	5 (50 %)	7 (88 %)	0 (0 %)
Age (years)						
Mean (SD)	8.5 (±5.3)	5.6 (±4.5)	9.2 (±5.4)	9.5 (±4.3)	11 (±6.5)	5.8 (±7.4)
Grade						
Low	18 (53 %)	8 (100 %)	0 (0 %)	4 (40 %)	6 (75 %)	0 (0 %)
High	12 (35 %)	0 (0 %)	6 (100 %)	5 (50 %)	1 (12 %)	0 (0 %)
NEC/NOS	2 (6 %)	0 (0 %)	0 (0 %)	1 (10 %)	1 (12 %)	0 (0 %)
Missing	2 (5.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)
Location						
Metastasis	1 (3 %)	1 (12 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Subtentorial	19 (56 %)	5 (62 %)	5 (83 %)	8 (80 %)	1 (12 %)	0 (0 %)
Supratentorial	14 (41 %)	2 (25 %)	1 (17 %)	2 (20 %)	7 (88 %)	2 (100 %)

Data collection and pre-processing

Single nuclei RNA-sequencing

The protocol to obtain single nuclei RNA-sequencing data and initial pre-processing steps were described in Chapter 3. To summarize briefly, nuclei were isolated from fresh frozen tissue samples following the Nuclei Pure Prep nuclei isolation kit (Sigma-Aldrich, St. Louis, MO). Each sample was multiplexed with lipid-tagged oligonucleotides following the MULTI-seq protocol[448]. Libraries for single nuclei RNA-seq were prepared following the 10X Genomics Single Cell Gene Expression workflows (10X Genomics, Pleasanton, CA). Libraries were pooled and sequenced using the Illumina NextSeq500 instrument. 10X Cell Ranger software was used to align sequences to the GRCh38 pre-mRNA reference genome.

Low-quality nuclei, as defined as having greater than 10,000 and less than 2,000 features and more than 5% of reads that map to mitochondrial genes, were removed for analyses. Samples were demultiplexed using an integrative approach, combining barcode based demultiplexing and genotype-based demultiplex method[450,453]. Downstream analyses for single nuclei-RNA seq were done with the Seurat package v4 in R[449–452].

Bulk RNA-sequencing

Unused nuclei from our single nuclei RNA-seq experiment were used for bulk RNA-sequencing. RNA was isolated following the RNeasy Plus kit (Qiagen, Hilden, Germany). Libraries for bulk RNA-seq were prepared following the Takara Pico v3 low-input protocol (Takara Bio, Kusatsu, Japan).

Quality control for raw single-end RNA-seq data was checked using FastQC v0.11.8[548]. Reads were trimmed of polyA sequences and low-quality bases using Cutadapt v2.4[414]. Reads were aligned to the human pre-mRNA genome GRCh38 with STAR v2.7.7a[425]. Quality control of aligned reads was confirmed with *CollectRNASeqMetrics* in the Picard software v2.18.29[416]. Duplicate reads were identified with *MarkDuplicates* function in the Picard software[416]. One sample with an extremely high duplicate read percentage was removed from downstream analyses. Counts per gene were estimated using the *htseq-count* function in the HTseq software v0.11.2[549].

DNA methylation and hydroxymethylation

In total, DNA from 33 paired pediatric brain tumor samples was treated with tandem bisulfite and oxidative bisulfite conversion followed by hybridization to Infinium HumanMethylationEPIC BeadChips to measure DNA methylation (5-mC) and hydroxymethylation (5-hmC). Raw BeadArray data were preprocessed using the *SeSAMe* pipeline from Bioconductor, including data normalization and quality control[550]. Cross-reactive probes, SNP-related probes, sex chromosome probes, non-CpG probes, and low-quality probes (pOOBHA > 0.05) were masked in the analysis[409]. The *oxBS.MLE* function was used to infer 5-mC and 5-hmC levels[185].

Tumor purity estimates

Tumor purity for the tissue samples with DNA cytosine modifications was estimated using the *getPurity* function with the non-tumor pediatric tumor tissue as our non-tumor reference and the low-grade glioma (LGG) option as our cancer type in the InifiniumPurify package v1.3.1 in R[551].

Statistical analyses

Epigenome-wide association studies

Linear regression models, adjusting for sex, age at diagnosis, and tumor purity in all models, were used to identify differentially methylated and hydroxymethylated CpGs associated with each tumor type compared to the non-tumor tissue. Multiple linear regression models, with adjustments for different cell type proportions identified from the single nuclei RNA-seq data, were added to the models. Linear regression models were fit by using *lmFit* and *eBayes* functions in the limma package in R[410]. CpGs were considered differentially methylated or hydroxymethylated under the q-value threshold of 0.05.

Cell type-specific differential hydroxymethylation and methylation for each tumor type were identified using CellDMC[37]. Proportions of cell types of interest (neurons and progenitor-like cell types) were pulled from the single nuclei RNA-seq dataset. To limit overfitting the model in our relatively smaller sample size, we aggregated the progenitor-like cell types into a single cell type category. The progenitor-like cell types included neural stem cells (NSC), radial glial cells (RGC), oligodendrocyte precursor cells (OPC), and unipolar brush cells (UBC). UBCs were included due to the high levels of stemness score in the cell types identified previously.

Differential gene expression testing

Negative binomial regression models were used to identify the differential expressed genes in each tumor type compared to non-tumor tissue. One model was fit adjusting for age at diagnosis and sex. One model was fit adjusting for age at diagnosis, sex, and the proportions for cell types of interest (NEU, NSC, RGC, OPC, UBC), Negative binomial models were fit by using *DESeq* function in the DESeq2 package v1.36.0 in R[552]. Genes were considered as differentially expressed under the adjusted p-value threshold of 0.05.

Pathways enrichment testing

Reactome pathways enrichment associated with differentially expressed genes in each tumor type were identified using the *enrichPathway* function in the ReactomePA package v1.40.0 in R[421].

Genomic context enrichment test

Enrichment tests for genomic context for differentially hydroxymethylated CpGs were conducted using the Mantel-Haenszel test. The MH test was adjusted for the type of probe (Type I or Type II) used for the CpG in the Illumina Methylation EPIC array.

4.4. Results

To assess the potential normal tissue margin in our tissues that may confound downstream analyses, we first determined the tumor purity of our pediatric CNS tumor samples that were used to measure DNA cytosine modifications. Tumor purity in our samples varied but did not significantly differ based on tumor type or grade (**Supplementary Figure 4-1**).

Genomic burden altered cytosine modifications

To determine the global epigenomic burden of altered cytosine modifications in pediatric CNS tumors compared to non-tumor pediatric brain tissue, we compared median beta values for both 5-hmC and 5-mC across samples at each CpG and determined the methylation dysregulation index (MDI). MDI is a summary measure of the epigenome-wide alteration of tumors compared to non-tumor tissue[553]. Tumor tissues displayed a decrease in 5-hmC and a slight increase in 5-mC compared to non-tumor tissue (**Figure 4-1A**). The 5-hmC MDI values were not significantly different by tumor type or by tumor grade (**Figure 4-1B**), whereas 5-mC MDI values varied by tumor type. Embryonal tumors had the greatest extent of epigenome-wide alteration burden compared to non-tumor tissue, astrocytomas had the lowest burden of 5-mC MDI compared to non-tumor tissue, and we observed increasing 5-mC MDI with increasing tumor grade. 5-hmC MDI and 5-mC MDI were positively correlated (R = 0.44, p-value = 0.013, **Figure 4-1C**). We repeated our analysis after removing one astrocytoma sample with an outlier 5-hmC MDI value and observed consistent results (**Supplementary Figure 4-2**). We tested and confirmed that the burden of observed epigenomic alterations was not due to differences in tumor purity, (**Supplementary Figure 4-3**, **Supplementary Table 4-2A**). However, we did observe significant differences in 5-mC MDI by tumor grade (**Supplementary Table 4-2B**). While 5-hmC is prevalent at only 6% of 5-mC, the level of dysregulation of the hydroxymethylome is comparable to the level of dysregulation of the methylome with 5-hmC MDI being 49% of 5-mC MDI (**Table 4-2**). Our results suggest that while 5-hmC may not be as prevalent, epigenome-wide alterations of 5-hmC in tumors are occurring at comparable levels to altered 5-mC.

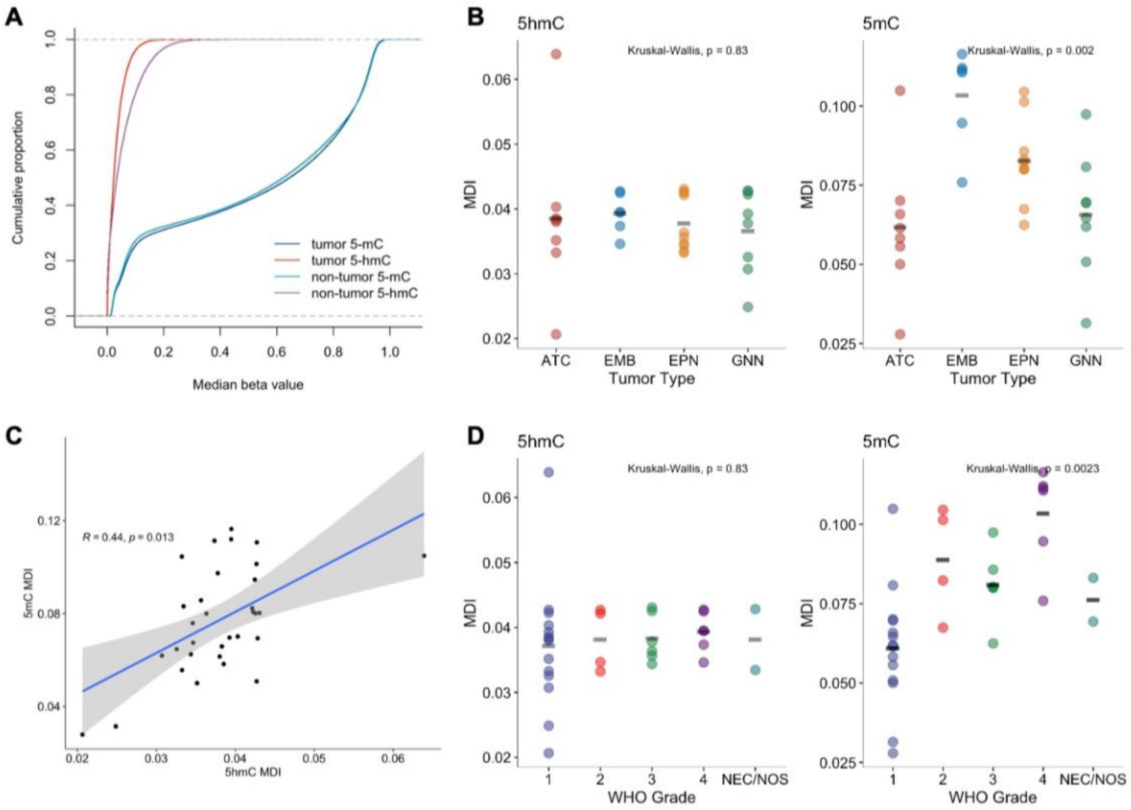


Figure 4-1. Global methylation dysregulation, but not global hydroxymethylation dysregulation, is associated with tumor type and grade. A) Cumulative proportion of 5-hmC and 5-mC in tumors and non-tumor tissue. **B)** Methylation dysregulation index of 5-hmC and 5-mC by tumor type and **D)** grade. Gray segments indicate median MDI values. Differences in MDI calculated using Kruskal-Wallis test. **C)** Correlation between 5-hmC MDI and 5-mC MDI calculated using Spearman rank correlation. Linear regression line indicated by the blue line. 95% confidence interval indicated by gray bands.

Table 4-2. Summary measure of global averages and MDI of 5-hmC and 5-mC

Measure	Modification	Minimum	Median	Mean	Max	Mean ratio (5-hmC/5-mC)
Global average	**5-hmC**	0.016	0.027	0.032	0.102	0.060
	5-mC	0.448	0.535	0.530	0.579	
MDI	**5-hmC**	0.021	0.038	0.038	0.064	0.494
	5-mC	0.028	0.078	0.077	0.116	

Cell type composition influences bulk-omics comparisons between pediatric CNS tumors and non-tumor pediatric brain tissue

We utilized our single nuclei RNA-seq data to identify the cell type composition of pediatric CNS tumor tissue and non-tumor pediatric brain tissue. Based on the cell type proportion distributions for all of our samples, we identified neuronal-like cells (NEU), neural stem cells (NSC), oligodendrocyte precursor cells (OPC), radial glial cells (RGC), and unipolar brush cells (UBC) as having the most variance (**Supplementary Figure 4-4A**). For each tumor type we compared proportions of cell types with non-tumor pediatric brain tissue. Supporting our principal component analysis, the cell types with the greatest differences were NEU, NSC, OPC, RGC, and UBC (**Supplementary Figure 4-4B**).

We conducted an epigenome-wide association study to determine the differential hydroxymethylated and methylated CpGs associated with each tumor type compared to non-tumor pediatric brain tissue. To reduce potential confounding by cell type composition, we incorporated cell type proportions as covariates in a stepwise manner to each series of linear models. Importantly, as the number of cell type proportion covariates included in the models increased, the scope of differentially hydroxymethylated and differentially methylated CpGs associated with each tumor type decreased (**Figure 4-2A – 2D, Supplementary Figure 4-5-5 – 4-8**). In addition, across our models in different tumor types, the extent of differentially hydroxymethylated CpGs (dhmCpGs) was far greater than that of differentially methylated CpGs (dmCpGs). When all five cell types (NEU, NSC, OPC, RGC, and UBC) were incorporated into the model, we observed low number of dmCpGs associated with each tumor type. Embryonal tumors had the greatest number of dhmCpGs, and the 83.1% were specific to the embryonal tumors (**Figure 4-2E**). In the model with all five cell types included, 87 dhmCpGs were associated with astrocytoma, 850 dhmCpGs were associated with embryonal tumors, 31 dhmCpGs were associated with ependymoma, and 126 dhmCpGs were associated with glioneuronal/neuronal tumors. We identified 90 dhmCpGs (10.4%) that were shared across two or three of the tumor types and 28 dhmCpGs (3.2%) that were shared across all tumor types (**Figure 4-2E**). Our results suggest that epigenome-wide association studies comparing bulk pediatric CNS tumor tissue to non-tumor pediatric tissue are considerably influenced by the cell type composition. Moreover, it was quite unexpected that the observed differences were almost solely in hydroxymethylation and not in methylation.

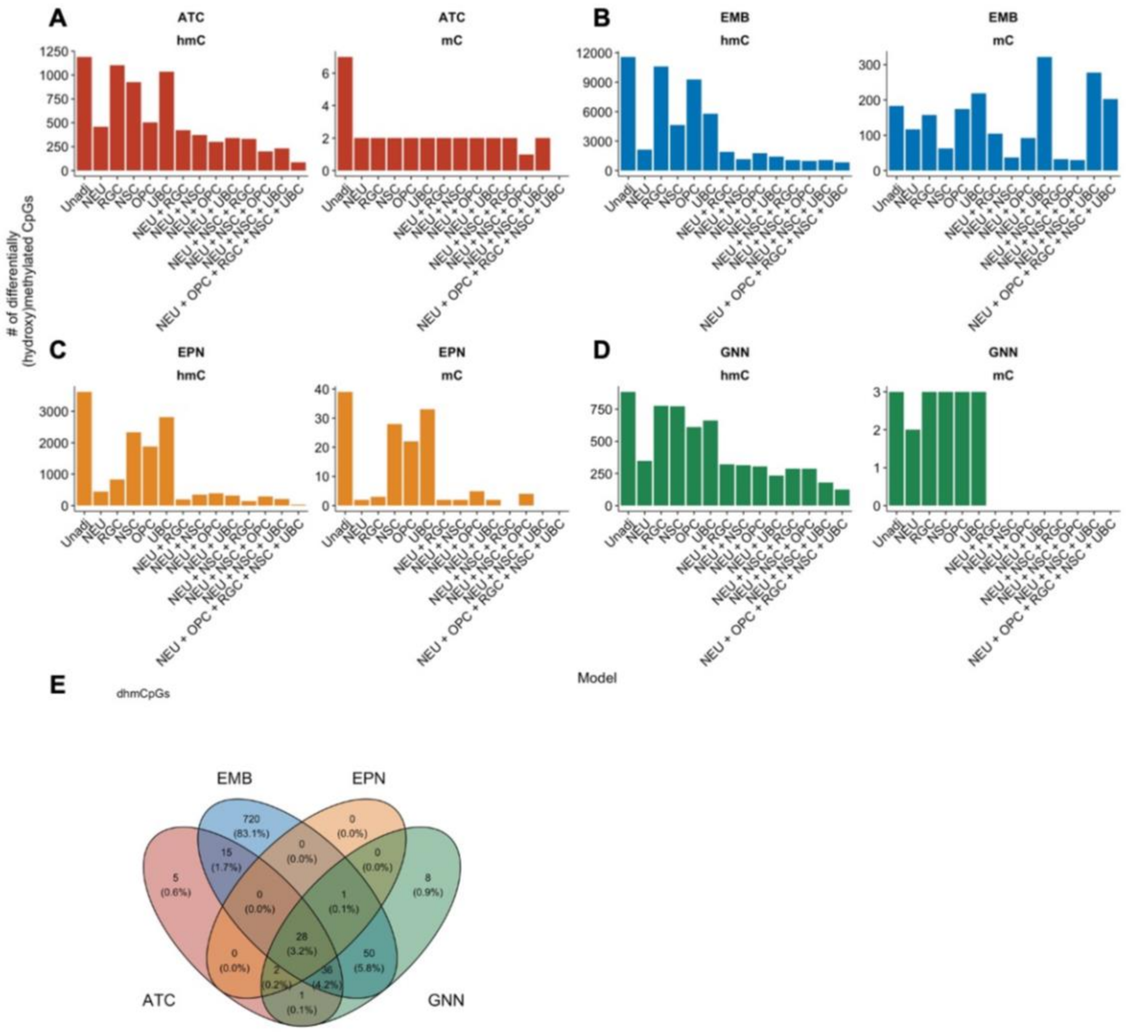


Figure 4-2. Adjusting for proportions of cell types of interest reduce the number of differentially hydroxymethylated and methylated CpGs across tumor types compared to non-tumor pediatric brain tissue.
Number of differentially hydroxymethylated and methylated CpGs under q-value < 0.05 threshold in **A)** astrocytoma (ATC), **B)** embryonal tumors (EMB), **C)** ependymoma (EPN), and **D)** glioneuronal/neuronal tumors (GNN) compared to non-tumor pediatric brain tissue. X-axis indicates each cell type proportion included in the model. Each model, even 'unadjusted' model includes sex and age at diagnosis in the linear model. **E)** Venn diagram of the differentially hydroxymethylated CpGs among the different tumor types.

We then compared transcriptome data from bulk RNA-seq in each of the tumor types with non-tumor pediatric brain tissue. The differential expression testing model included the same covariates (sex, age at diagnosis, and tumor purity) and the same five cell type proportions used for the EWAS analysis. Including proportions of major cell types of interest led to differences in an average of around 702 genes (range: 536 – 892) detected as significantly differentially expressed. In astrocytoma and glioneuronal/neuronal tumors, the adjusted model identified more genes that were significantly differentially expressed. In embryonal tumors and ependymomas, the adjusted model identified fewer genes that were significantly differentially expressed. Some key tumor progression-associated genes like *PTEN* in astrocytoma and in embryonal tumors, *MYCN* in ependymoma, and *BRCA2* in glioneuronal/neuronal tumors would not otherwise have been identified as significantly differentially expressed in the tumors had the cell type proportions not been adjusted for.

Across all tumor types, the majority of differentially expressed genes were increased in expression compared to the non-tumor pediatric brain tissue (**Supplementary Figure 4-9A, Supplementary Figure 4-10– 4-13**). Almost half (43%, 3020 genes) of all genes with increased expression were shared across all tumor types (**Supplementary Figure 4-9B**). Among the genes with shared increases in expression in tumors were *IRX5*, *MYOSLID*, *CWH43, ITGA2,* and *HOXA3*. Genes with increased expression across all tumor types were associated with biological oxidations and keratinization among other pathways (**Supplementary Figure 4-9D**). There were 253 genes (13.6%) that had decreased expression shared across tumor types (**Supplementary Figure 4-9C**), including *NPTXR, SCG2 , B4GAT1,* and *ATRN.* Genes that were decreased in expression across all tumor types were associated with the insulin receptor signaling and ion channel transport among other pathways (**Supplementary Figure 4-9E**).

To identify potentially important gene regulation by differential hydroxymethylation we compared changes in hydroxymethylation in dhmCpGs from the five-cell type-adjusted model with gene expression in each tumor type. Generally, genes with decreased hydroxymethylation levels had increased gene expression across tumor types compared to non-tumor pediatric brain tissue (**Figure 4-3**). Only one dhmCpGs associated with ependymoma had significant decreased expression. The dhmCpGs with differential expression did not generally favor promoters or gene body regions (**Figure 4-3, Supplementary Table 4-3**). Only embryonal tumors displayed slightly varying associations. While many of the dhmCpGs associated with embryonal tumors followed similar patterns of decreased 5-hmC levels and increased gene expression, there were some CpGs with decreased 5-hmC and decreased gene expression, as well as CpGs with increased 5-hmC with increased or decreased gene expression levels. Embryonal tumor associated dhmCpGs with significantly increased gene expression were less likely to be in promoter regions compared to dhmCpGs with significantly decreased gene expression (OR (95%CI) = 0.23 (0.064 – 0.78), p-value = 0.01). On the contrary, embryonal tumor associated dhmCpGs with significant increased expression were marginally more likely to be in gene body regions (OR (95%CI) = 2.81 (0.84 – 10.34), p-value = 0.06). We could not test for associations between promoter or gene body regions for other tumor types due to the limited number of dhmCpGs.

Interestingly, there were two CpGs with decreased 5-hmC levels and increased gene expression in astrocytoma, ependymoma, and glioneuronal/neuronal tumors: cg18280362 located in the promoter region of *CWH43* and cg08278401 located in the promoter region of *LRRC72*. In addition, we investigated the association between changes in 5-mC methylation and gene expression in the embryonal tumors where there were 24 dmCpGs associated with significant changes in gene expression (**Supplementary Figure 4-14**). While we could not conduct statistical tests to test for an enrichment of promoter/gene body regions for shared dhmCpGs with increased gene expression, there were 18 dhmCpGs with increased gene expression in non-promoter regions and 3 dhmCpGs with increased gene expression in promoter regions. Moreover, there were 9 dhmCpGs with increased gene expression not in gene body regions and 12 dhmCpGs in gene body regions (**Supplementary Table 4-3**). Our results indicate that hydroxymethylation may be associated with changes in gene expression for certain genes in pediatric CNS tumors.

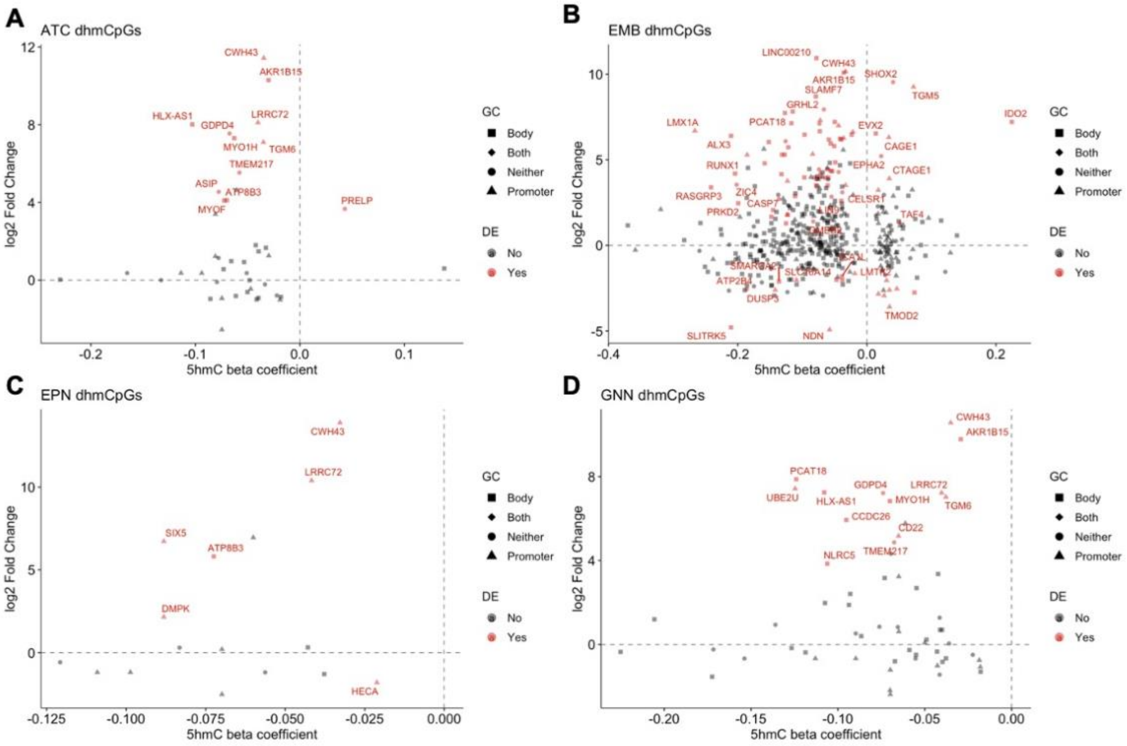


Figure 4-3. Hypo-hydroxymethylation of CpGs are associated with changes in gene expression.
Association between differentially hydroxymethylated CpG beta coefficients and log2 fold changes in gene expression for **A)** astrocytoma, **B)** embryonal tumors, **C)** ependymoma, and **D)** glioneuronal/neuronal tumors. Red points indicate significantly differentially expressed genes. Shapes indicate genomic context of CpGs.

Molecular alterations in pediatric CNS tumors occur in a cell type-specific and tumor type-specific manner

One of the major questions that remains unanswered in many epigenome-wide association studies is whether altered cytosine modification can be ascribed to a specific cell type. With data from single nuclei RNA-seq for these pediatric CNS tumors and non-tumor pediatric brain tissues, we sought to identify epigenomic alterations at a cell type-specific level. To reduce the number of covariates in our analysis we focused on neuronal-like and progenitor-like cell types (**Supplementary Table 4-4**). The progenitor-like cells were an aggregation of neural stem cells, radial glial cells, oligodendrocyte precursor cells, and unipolar brush cells. We used an approach developed by Zheng et al[37] called CellDMC to identify cell-type-specific differentially hydroxymethylated and methylated CpGs. Using CellDMC we identified abundant dhmCpGs for each cell type and tumor type, far greater than the scope of CpGs identified with bulk tissue EWAS (**Figure 4-4A, Supplementary Figure 4-15– 4-19, Supplementary Table 4-5**). While there were a relatively lower number of dmCpGs compared to the dhmCpGs, there were some dmCpGs detected in the cell type-specific model (**Figure 4-4B**). Majority of the cell type-specific dhmCpGs were tumor-type-specific (**Figure 4-4C – 4-4D, Supplementary Figure 4-19**). However, 128 dhmCpGs were observed in the neuronal-like cell types and 534 dhmCpGs were observed in the progenitor-like cell types across all four tumor types. While some neuronal-like cell-specific dhmCpGs were acting on the same genes as the progenitor-like cell-specific dhmCpGs, genes that had decreased 5-hmC in the progenitor-like cells were exclusive (**Supplementary Figure 4-20**).

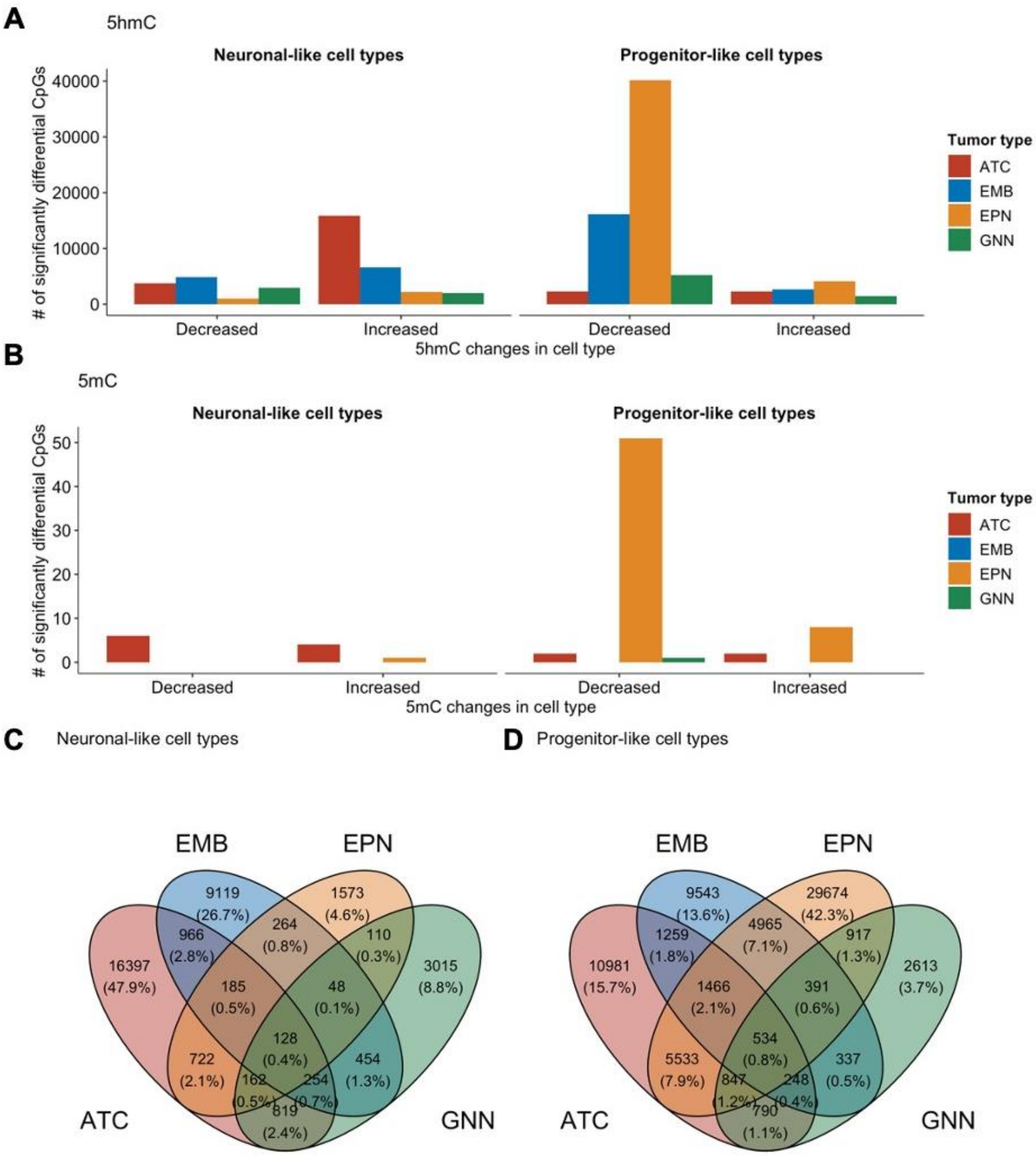


Figure 4-4. 5-hmC is altered in cell type-specific and tumor type-specific manner.
Cell type associated differentially **A)** hydroxymethylated and **B)** methylated CpGs in each tumor type. Venn diagram of shared differentially hydroxymethylated CpGs in **C)** neuronal-like cell types and **D)** progenitor-like cell types across the four tumor types.

We then assessed the genomic context of cell type-specific dhmCpGs and tested for enrichment to various genomic contexts stratified by the direction of differential hydroxymethylation. Interestingly, both increased and decreased dhmCpGs in neuronal-like and progenitor-like cell types of astrocytoma and glioneuronal/neuronal tumors were enriched in similar contexts at Dnase hypersensitive sites (DHS), 1st exons, promoter regions (TSS200, TSS1500), and 5’ UTR regions (**Figure 4-5**). dhmCpGs in ependymoma were dependent on the cell type in which it was occurring. Ependymoma associated dhmCpGs in the neuronal-like cells and CpGs with increased 5-hmC in progenitor-like cells were enriched in similar regions as the astrocytoma and glioneuronal/neuronal tumors. On the contrary, ependymoma associated CpGs with decreased 5-hmC in the progenitor-like cells were enriched in transcription factor binding sites (TFBS), 3’ UTR, gene body, and exon regions. The dhmCpGs, especially for those occurring in the progenitor-like cell types, in embryonal tumors were enriched in distinct genomic contexts compared to the other tumor types. Progenitor-like cell type-specific dhmCpGs were enriched in the transcription factor binding sites, 3’ UTR, gene body, exons, and enhancers.

Our findings indicate that most of the hydroxymethylation alterations occur in the progenitor-like cell types and are tumor-type-specific.

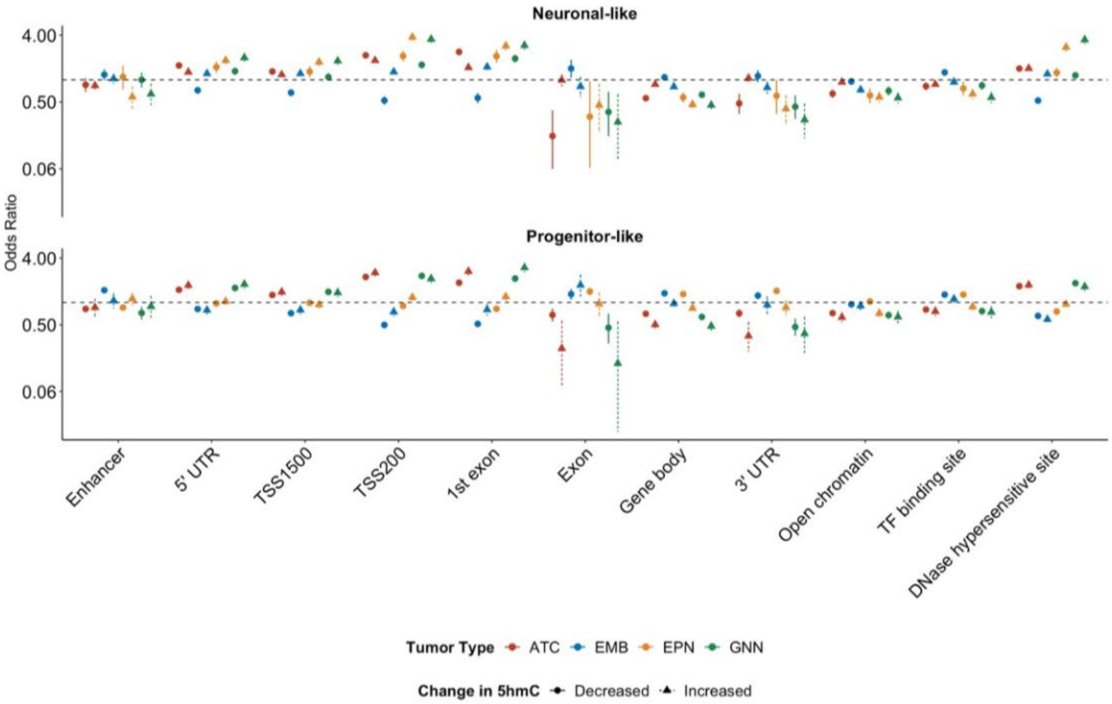


Figure 4-5. Cell type-specific differential hydroxymethylation tumor type-specific.
Enrichment of differentially hydroxymethylated CpGs at specific genomic contexts by tumor type and direction of differential methylation.

Cell type-specific gene expression changes associated with changes in hydroxymethylation

We next evaluated cell-specific gene expression changes for genes with cell-type-specific changes in hydroxymethylation. We calculated gene expression scores for genes associated with CpGs with differentially hydroxymethylated CpGs in the neuronal-like cells and progenitor-like cells for each granular cell types incorporated in our analysis for each tumor type (**Supplementary Figure 4-21– 4-24**). Interestingly, for all tumor types, the expression scores for genes associated with CpGs with increased or decreased hydroxymethylation were increased in the oligodendrocyte precursor cells (OPCs) of the tumors compared to non-tumor pediatric brain tissue (**Figure 4-6A**). Only the OPCs in embryonal tumors did not show a statistically significant increase in the expression of genes with increased 5-hmC in the progenitor-like cells. On the contrary, gene expression levels for each of the gene sets with cell type-specific alterations in 5-hmC were decreased in each of the cell types for all tumors compared to the non-tumor pediatric brain tissue.

HDAC4, established as associated with cancer progression and poor prognosis in a variety of tumor types[554–562], was one gene with cell type-specific dhmCpGs across all four tumor types. Interestingly, the majority of the CpGs with decreased 5-hmC were associated with progenitor-like cell types, while the majority of the CpGs with increased 5-hmC were associated with the neuronal-like cell types in the tumor tissue (**Figure 4-6B**). More than 50% of the dhmCpGs in *HDAC4* for each tumor type were in the gene body (**Table 4-3**). There were few dhmCpGs in the 5' UTR, TSS200, and DNase hypersensitive sites (DHS). The neuronal-like cell types had lower expression of *HDAC4* across all tumor types compared to the non-tumor tissue (**Figure 4-6D**). On the contrary, the progenitor-like cell types had higher levels of *HDAC4* expression.

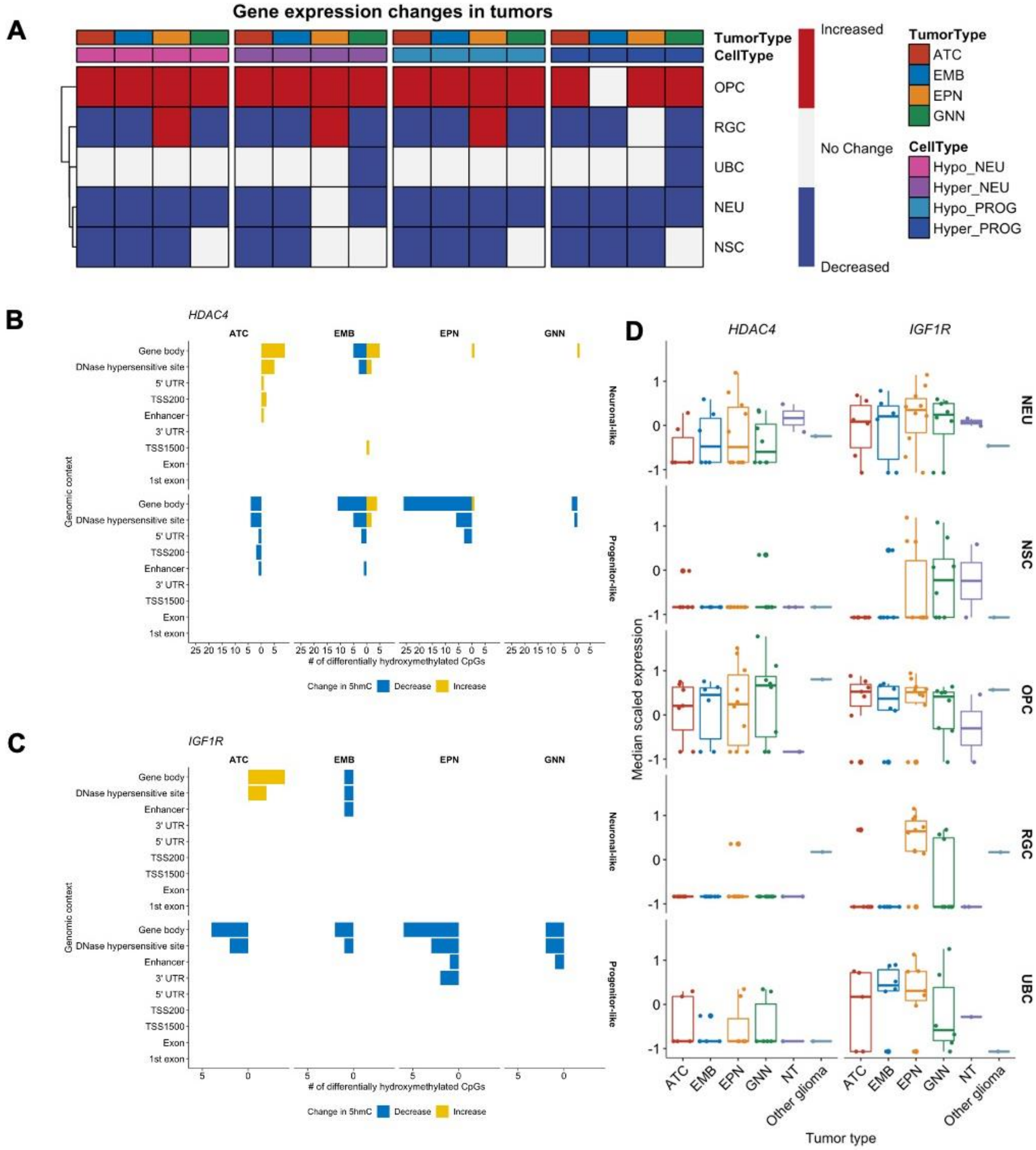


Figure 4-6. Alterations in hydroxymethylation is associated with cell type specific changes in gene expression.

A) Summary heatmap of changes in gene expression in the gene sets with differentially hydroxymethylated CpGs per cell type. Number of differentially hydroxymethylated CpGs associated with **B)** *HDAC4* and **C)** *IGF1R* in each genomic context across the different tumor types in neuronal-like cell types and progenitor-like cell types. Blue bars indicate the number of hydroxymethylated CpGs that are decreased in the tumors. Yellow bars indicate the number of hydroxymethylated CpGs that are increased in the tumors. **D)** Gene expression levels of *HDAC4* and *IGF1R* for each cell type across the tumor types and non-tumor tissue.

Table 4-3. Genomic context of dhmCpGs in *HDAC4* and *IGF1R* for each tumor type.

HDAC4	**TSS200**	**TSS1500**	**Gene body**	**1st exon**	**5' UTR**	**3' UTR**	**Exon bound**	**Enhancer**	**DHS**	**dhmCpG total**
ATC	2 (15%)	0 (0%)	10 (77%)	0 (0%)	1 (8%)	0 (0%)	0 (0%)	1 (8%)	5 (38%)	13
EMB	0 (0%)	1 (5%)	16 (84%)	0 (0%)	2 (11%)	0 (0%)	0 (0%)	1 (5%)	9 (47%)	19
EPN	0 (0%)	0 (0%)	27 (90%)	0 (0%)	3 (10%)	0 (0%)	0 (0%)	0 (0%)	6 (20%)	30
GNN	0 (0%)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (50%)	2
IGF1R										
ATC	0 (0%)	0 (0%)	4 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (50%)	4
EMB	0 (0%)	0 (0%)	3 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (33%)	2 (67%)	3
EPN	0 (0%)	0 (0%)	6 (75%)	0 (0%)	0 (0%)	2 (25%)	0 (0%)	1 (13%)	3 (38%)	8
GNN	0 (0%)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (50%)	2 (100%)	2

IGF1R had dhmCpGs across all tumor types and is associated with tumorigenesis, therapy resistance, and poor survival in different cancer types, including in some pediatric CNS tumor types[563–573]. Most of the dhmCpGs with decreased 5-hmC were associated with the progenitor-like cell types in the tumor tissue while only a couple dhmCpGs were in the neuronal-like cell types of the tumor tissue (**Figure 4-6C**). Like *HDAC4,* the dhmCpGs in *IGF1R* were mostly located in the gene body and DNase hypersensitive sites, with a few scattered in the enhancer and 3' UTR regions (**Table 4-4)**. Consistent with the lack of changes in hydroxymethylation in the neuronal-like cell types of the tumors, gene expression levels of *IGF1R* did not differ between tumors and the non-tumor tissue among neuronal-like cell types (**Figure 4-6D**). However, following the decreases in hydroxymethylation, *IGF1R* gene expression levels were higher in the progenitor-like cell types, particularly the OPCs, in the tumors than in the progenitor-like cell types of non-tumor tissue. EWAS results from bulk tumor tissue identified only one or two CpGs in *HDAC4* and *IGF1R* as differentially hydroxymethylated in either cell type-adjusted or unadjusted model (**Table 4-4**).

Table 4-4. Comparison of number of differentially hydroxymethylated CpGs in *HDAC4* and *IGF1R* identified by bulk tissue EWAS and CellDMC for each tumor type.

	Tumor type	**Bulk EWAS (CT unadjusted)**	**Bulk EWAS (CT adjusted)**	**CellDMC (Neuronal-like)**	**CellDMC (Progenitor-like)**
HDAC4	**ATC**	0	0	12	7
	EMB	1	1	11	17
	EPN	1	0	1	30
	GNN	0	0	1	2
IGF1R	**ATC**	0	0	4	4
	EMB	2	0	1	2
	EPN	1	0	0	8
	GNN	0	0	0	2

Our results suggest potentially critical roles of hydroxymethylation of CpGs located within the gene body regions in regulating the gene expression of critical cancer genes, like *HDAC4* and *IGF1R*.

4.5. Discussion

In this study, we investigated the cell type-specific cytosine modification alterations in pediatric central nervous system tumors with a multi-omic approach. We described the cell type composition effects that occur in epigenome-wide association studies using bulk pediatric central nervous system tumors and non-tumor pediatric brain tissue. We identified that there were more differentially hydroxymethylated CpGs associated with each tumor type, particularly in the progenitor-like cell types, rather than differentially methylated CpGs. Lastly, we show that the cell type-specific changes in hydroxymethylation are associated with cell type-specific gene expression changes in pediatric central nervous system tumors.

Based on methods to classify tumor subtypes and the predominant focus on DNA methylation, it was unexpected that there were very few differentially methylated CpGs associated with each tumor type. One possible explanation for this phenomenon may be that as these are pediatric tissues, there is still ongoing development with which 5-hmC is associated. As our results suggest the epigenome-wide alterations of 5-hmC in these tumors, it may be critical to distinguish between 5-mC and 5-hmC to better understand the molecular underpinnings of these pediatric CNS tumors. Furthermore, it may be beneficial to incorporate 5-hmC into cytosine modification-based classification methods to improve performance.

Pediatric tumors are known not to have substantial genetic alterations. Our results suggest that pediatric CNS tumors may be characterized by non-mutational epigenomic reprogramming more so than genomic aberrations[40,188]. We identified a substantial number of differentially hydroxymethylated CpGs associated with progenitor-like cell types of each tumor type. Additionally, even among the shared differentially hydroxymethylated CpGs in the progenitor-like cell types, numerous differentially hydroxymethylated CpGs were located within different genes that regulate epigenetic patterns, such as *DNMT3A, HDAC4, MLLT3,* and *KAT2B.* Furthermore, pediatric brain cancers have been shown to contain somatic mutations in epigenetic regulator genes such as *H3F3A*, *KDM6A*, and *MLL3*[574–576]. Considering the dysregulation of the epigenome may be important when developing new therapeutic strategies for these tumors.

While much more investigation has been conducted into how DNA methylation regulates gene expression, less is known about how DNA hydroxymethylation can also be associated with changes in gene expression. We identified relationships between cell type-specific hydroxymethylation patterns and cell type-specific gene expression in our pediatric CNS tumors. Our findings indicate that hydroxymethylation changes in the gene body regions can alter gene expression. Previous studies have found positive associations between DNA methylation in gene body regions and gene expression changes[114,194]. However, many genome-wide DNA methylation studies use the traditional bisulfite treatment approach to measure 5-mC. Because bisulfite treatment alone cannot distinguish between 5-mC and 5-hmC[183], some methylation signals may have been from 5-hmC. Further studies that explicitly distinguish between 5-hmC and 5-mC are needed to gain a clearer understanding of the effects of DNA cytosine modifications on gene expression.

We identified two genes, *HDAC4* and *IGF1R*, in our pediatric CNS tumors that were both epigenetically and transcriptionally altered in comparison to non-tumor pediatric brain tissue. *HDAC4* and *IGF1R* had differentially hydroxymethylated CpGs and increased expression in oligodendrocyte precursor cells across all four of our tumor types. Our results suggest a potential role of hydroxymethylation regulating genes associated with tumorigenesis. With these targets already having been studied in adult cancers, there are pharmacological inhibitors that already exist for these targets. Our study expands previously suggested ideas of targeting *HDAC4* and *IGF1R* in certain pediatric CNS tumor types[568,577,578].

Accruing a large sample size for pediatric CNS tumors is extremely difficult as they are very rare in the general population. While our study does incorporate a decent sample size for these rare tumors, the smaller sample size limited the inclusion of other variables and cell types that may affect methylation and transcription into our models. Future studies with an expanded cohort of pediatric CNS patients will allow us to assess the epigenomic alterations in additional cell types of interest, such as glial cells. Moreover, following our findings of cell type-specific changes in DNA cytosine modifications in these pediatric CNS tumors, other tumor types may also have cell type-specific that have yet to be detected. Tools to understand the cell type composition of tissues should be incorporated in bulk epigenome-wide association studies to discriminate the cell type composition effects.

4.6. Conclusion

Our study addresses gaps that currently exist in understanding epigenomic alterations at the cell type level in pediatric central nervous system tumors. Changes in hydroxymethylation were particularly drastic in progenitor-like cells and were associated with cell type level alterations in transcription. We highlight the relevance of epigenome dysregulation in pediatric central nervous system tumors that may lead us to more effective therapeutic targets.

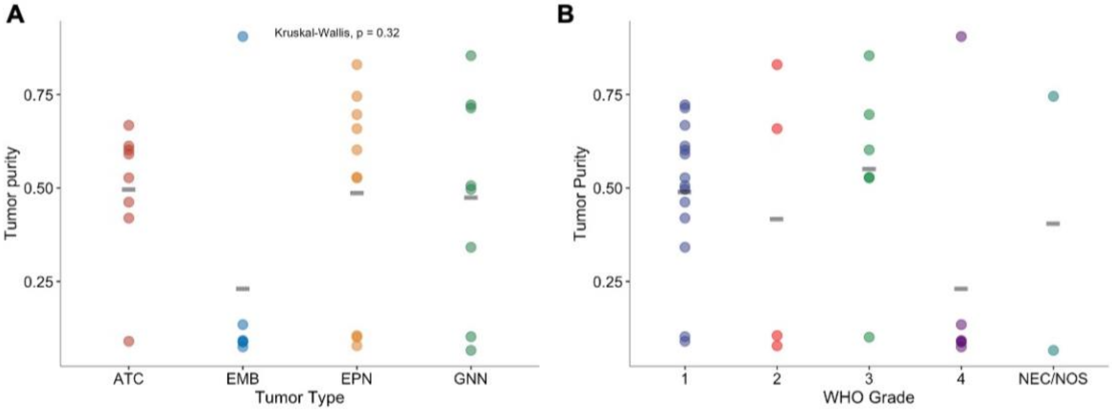
4.7. Author contributions

MKL, NA, and BCC designed the study. NA, GJZ, and LN identified subject populations and collected tissue samples. MKL, NA, LP, and FWK performed experiments to collect cytosine modification and gene expression data. MKL and ZZ processed data for downstream analyses. MKL performed statistical analyses under the supervision of LAS and BCC. BCC supervised the project. All authors reviewed the manuscript.

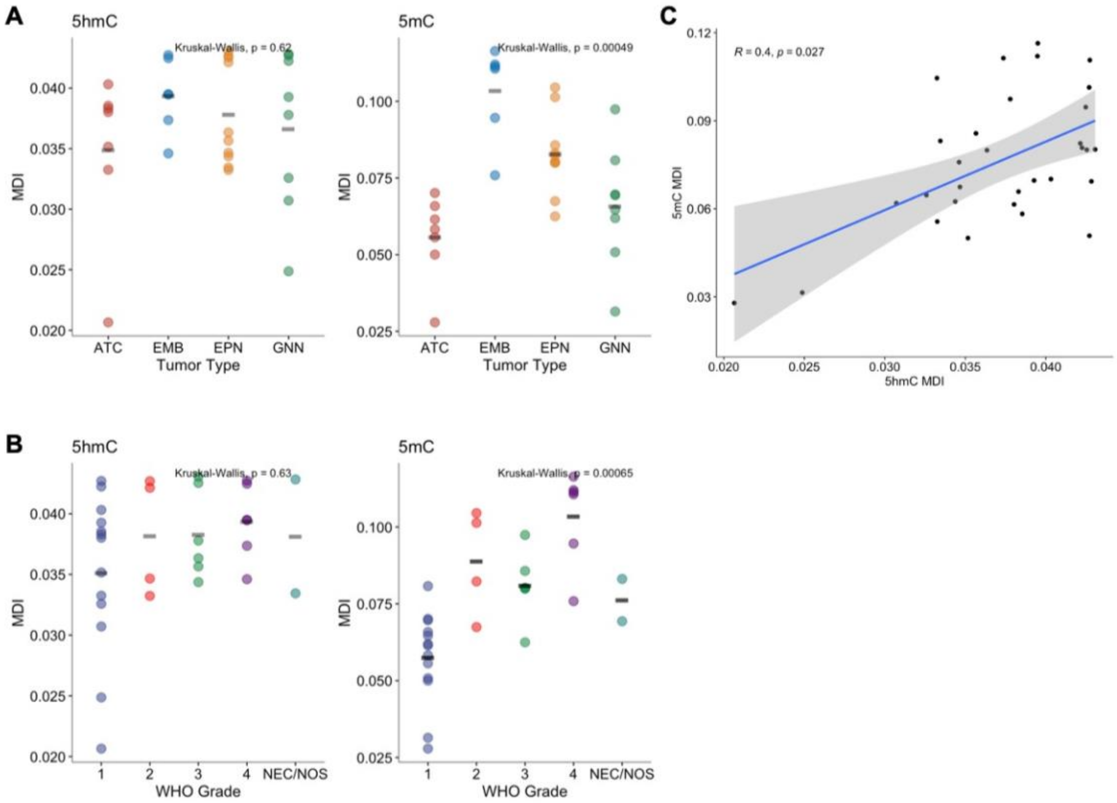
4.8. Funding sources

This work was supported by a Prouty Pilot award from the Dartmouth Cancer Center and a Single-cell Pediatric Cancer Atlas (ScPCA) grant from the Alex’s Lemonade Stand Foundation. MKL was supported by the Burroughs-Wellcome Fund: Big Data in the Life Sciences at Dartmouth. NA was supported by the S.M. Tenney Fellowship at Dartmouth. This work was also supported by R01CA216265, R01CA253976, and P20GM104416 – 6369 to BCC and CDMRP/Department of Defense (W81XWH-20-1-0778) and P20 GM104416-09/8299 to LAS. Single nuclei RNA-seq experiments were conducted in the Genomics and Molecular Biology Shared Resource (GMBSR) at Dartmouth, which is supported by NCI Cancer Center Support Grant 5P30CA023108 and NIH S10 (1S10OD030242) awards. Single-nuclei RNA experiments were also supported through the Dartmouth Center for Quantitative in collaboration with the GMBSR with support from NIGMS (P20GM130454) and NIH S10 (S10OD025235) awards.

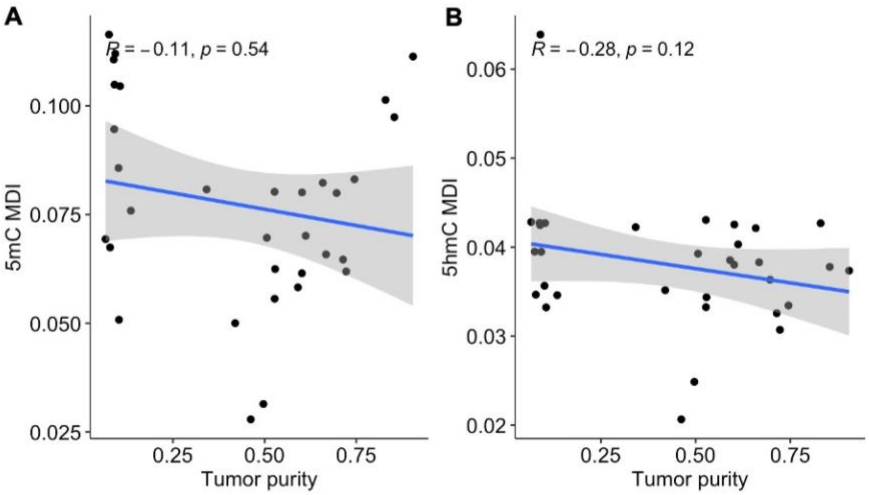
4.9. Supplemental materials



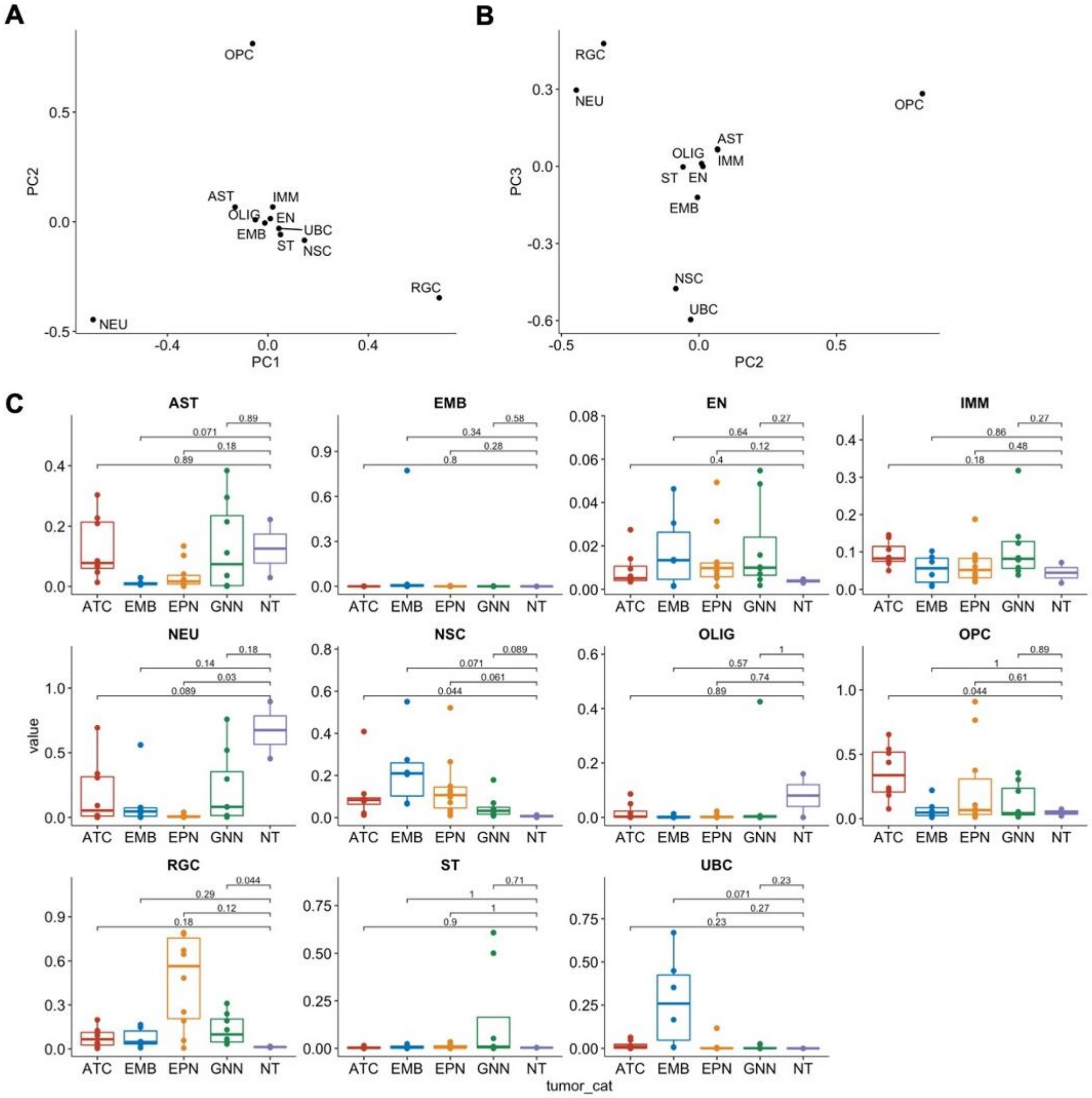
Supplementary Figure 4-1.
Tumor purity differences by **A)** tumor type and **B)** grade.



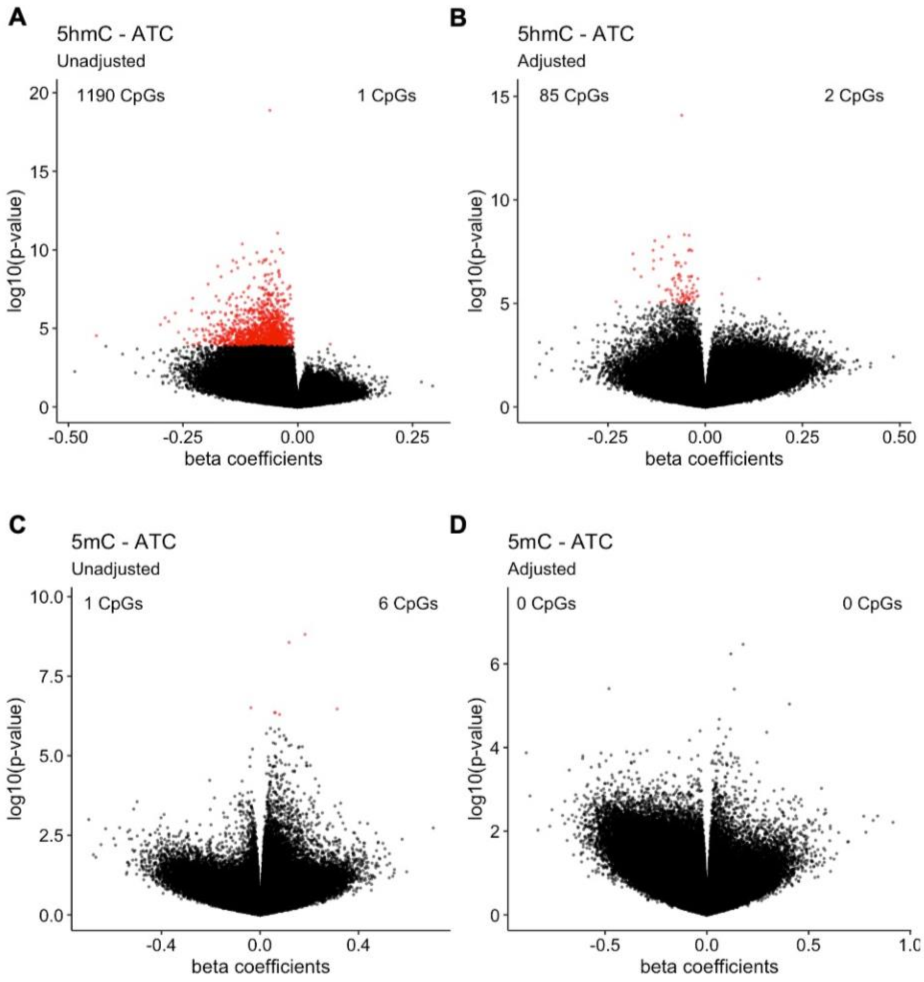
Supplementary Figure 4-2.
Methylation dysregulation index from 5-hmC and 5-mC by **A)** tumor type and **B)** grade without outliers. Gray segments indicate median MDI values. Differences in MDI calculated using Kruskal-Wallis test. **C)** Correlation between 5-hmC MDI and 5-mC MDI calculated using Spearman rank correlation. Linear regression line indicated by the blue line. 95% confidence interval indicated by gray bands.



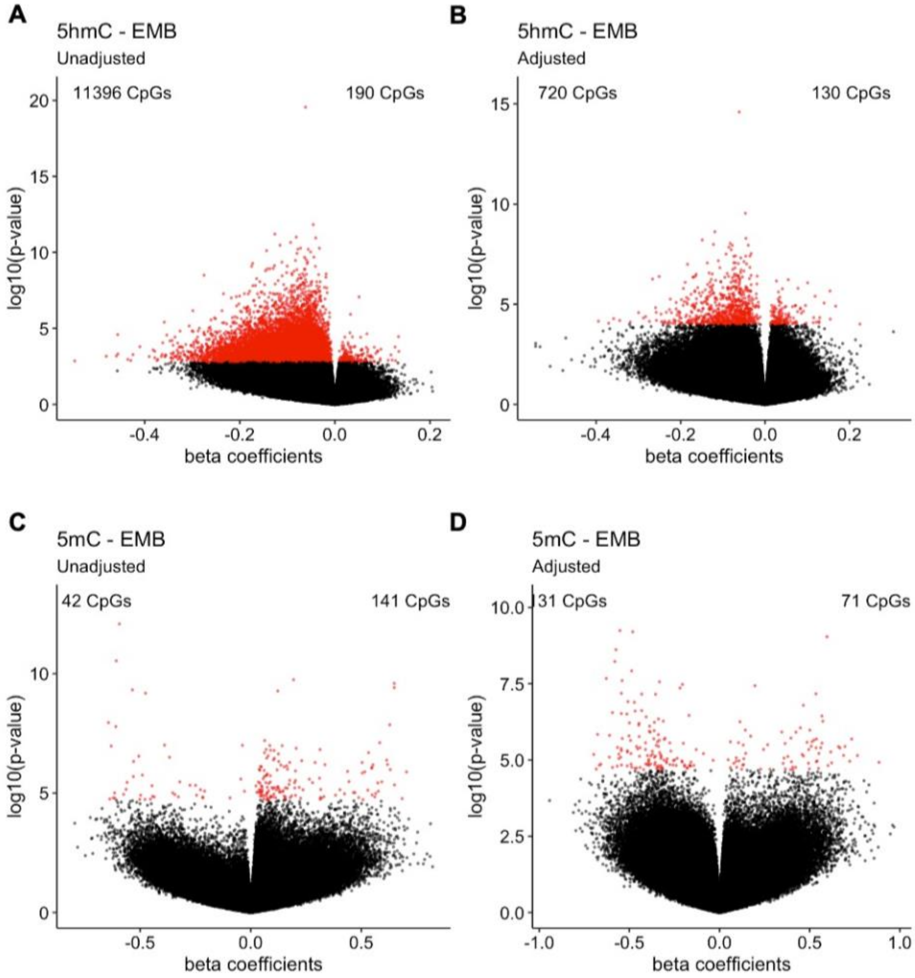
Supplementary Figure 4-3.
Tumor purity is not associated with MDI. Correlation between tumor purity and **A)** 5-mC MDI and **B)** 5-hmC MDI calculated using Spearman rank correlation. Linear regression line indicated by the blue line. 95% confidence interval indicated by gray bands.



Supplementary Figure 4-4.
A) Results from principal component analysis of cell type proportions from single nuclei RNA-seq of pediatric central nervous system tumors and non-tumor pediatric brain tissue. **B)** Comparison of each tumor type's proportions per cell type against the proportions found in non-tumor pediatric brain tissue. Each point indicates a sample. Differences in proportions calculated using Wilcoxon signed-rank test.

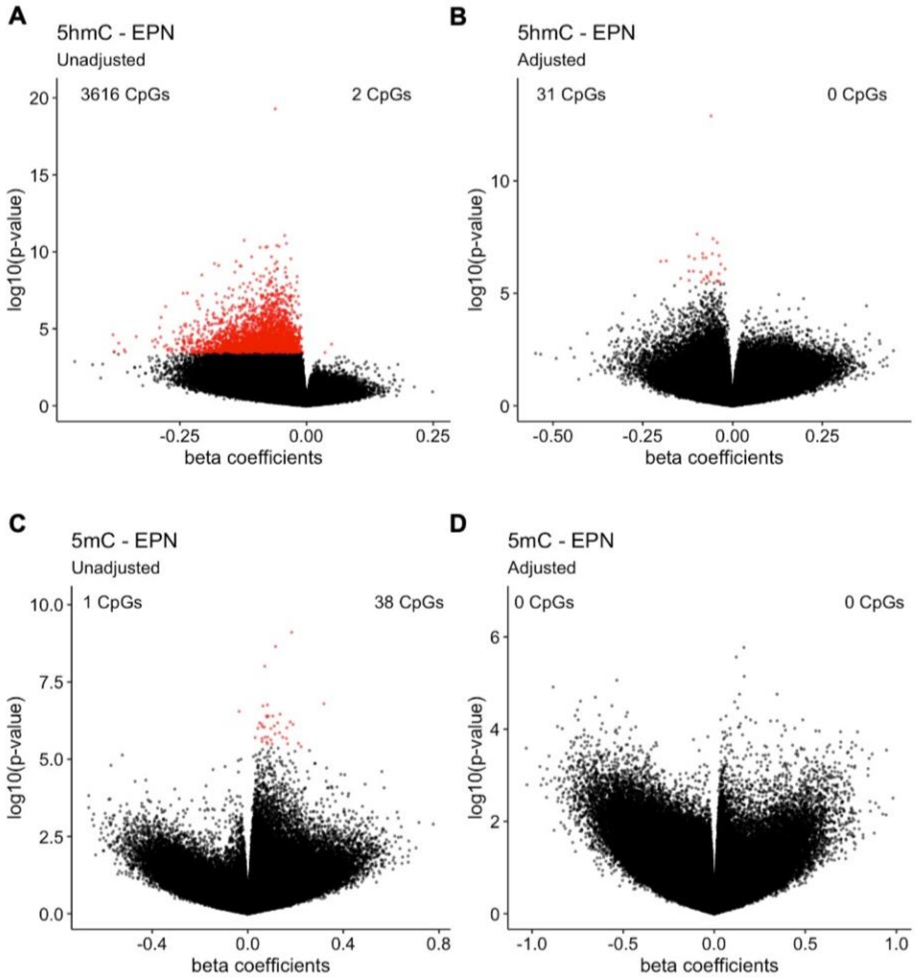


Supplementary Figure 4-5.
Volcano plots of differential **A)** 5-hmC CpGs and **C)** 5-mC CpGs in astrocytoma in the cell type proportion unadjusted model. Volcano plots of differential **B)** 5-hmC CpGs and **D)** 5-mC CpGs in astrocytoma in the cell type proportion adjusted model. Labeled # of CpGs on the left of each plot are CpGs with decreased methylation in tumors compared to non-tumor tissue. Labeled # of CpGs on the right of each volcano plot are CpGs with increased methylation in tumors compared to non-tumor tissue. Red points indicate statistically significant differential CpGs under the q-value < 0.05 threshold.

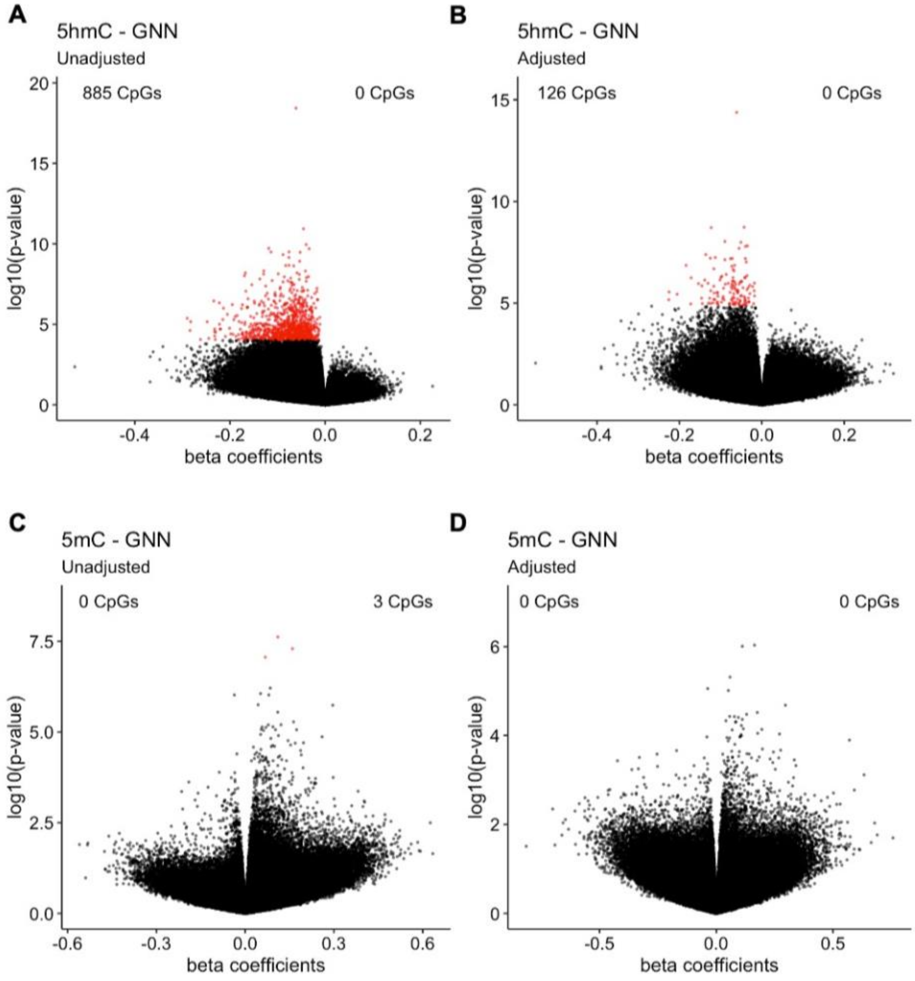


Supplementary Figure 4-6.

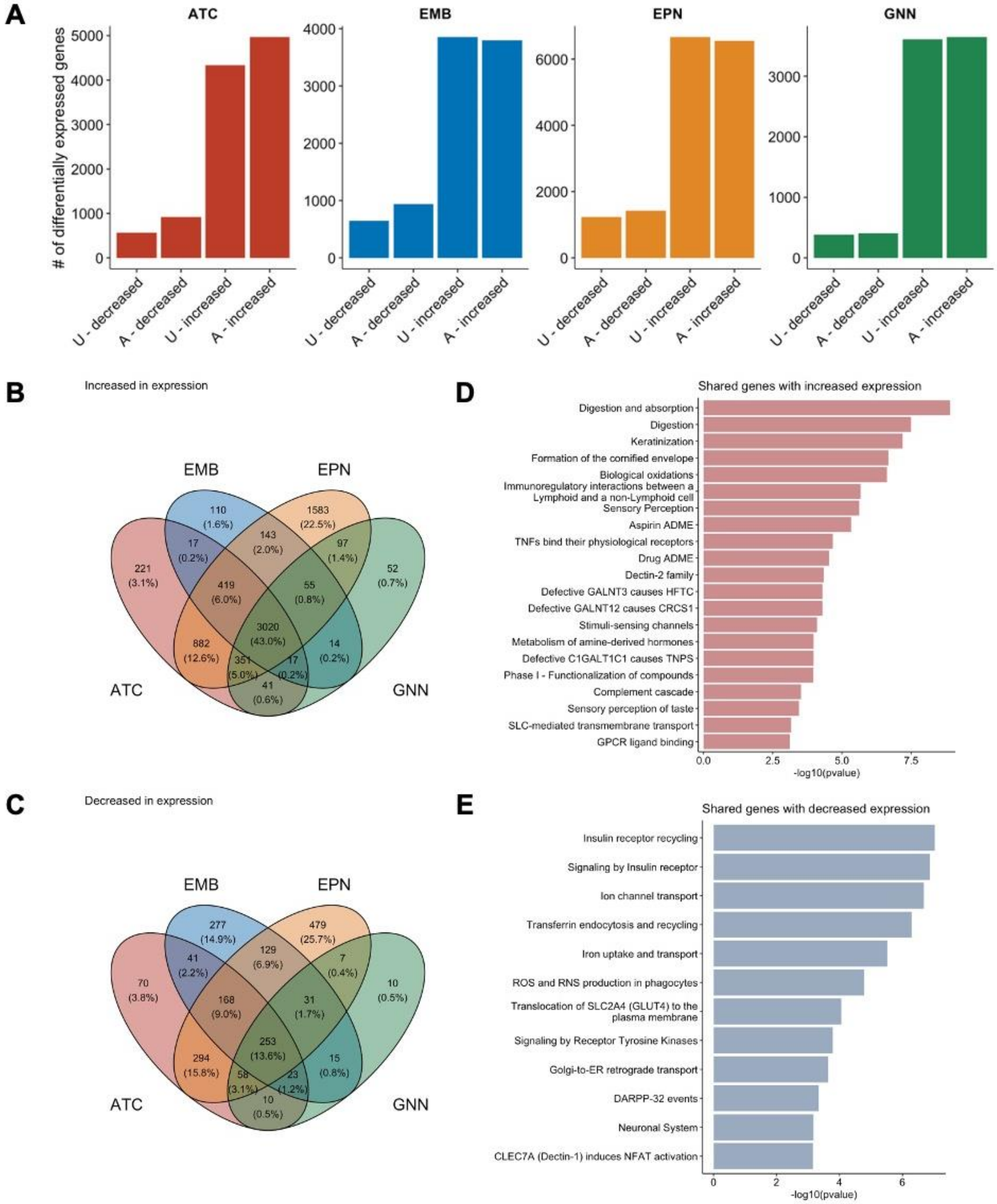
Volcano plots of differential **A)** 5-hmC CpGs and **C)** 5-mC CpGs in embryonal tumors in the cell type proportion unadjusted model. Volcano plots of differential **B)** 5-hmC CpGs and **D)** 5-mC CpGs in astrocytoma in the cell type proportion adjusted model. Labeled # of CpGs on the left of each plot are CpGs with decreased methylation in tumors compared to non-tumor tissue. Labeled # of CpGs on the right of each volcano plot are CpGs with increased methylation in tumors compared to non-tumor tissue. Red points indicate statistically significant differential CpGs under the q-value < 0.05 threshold.



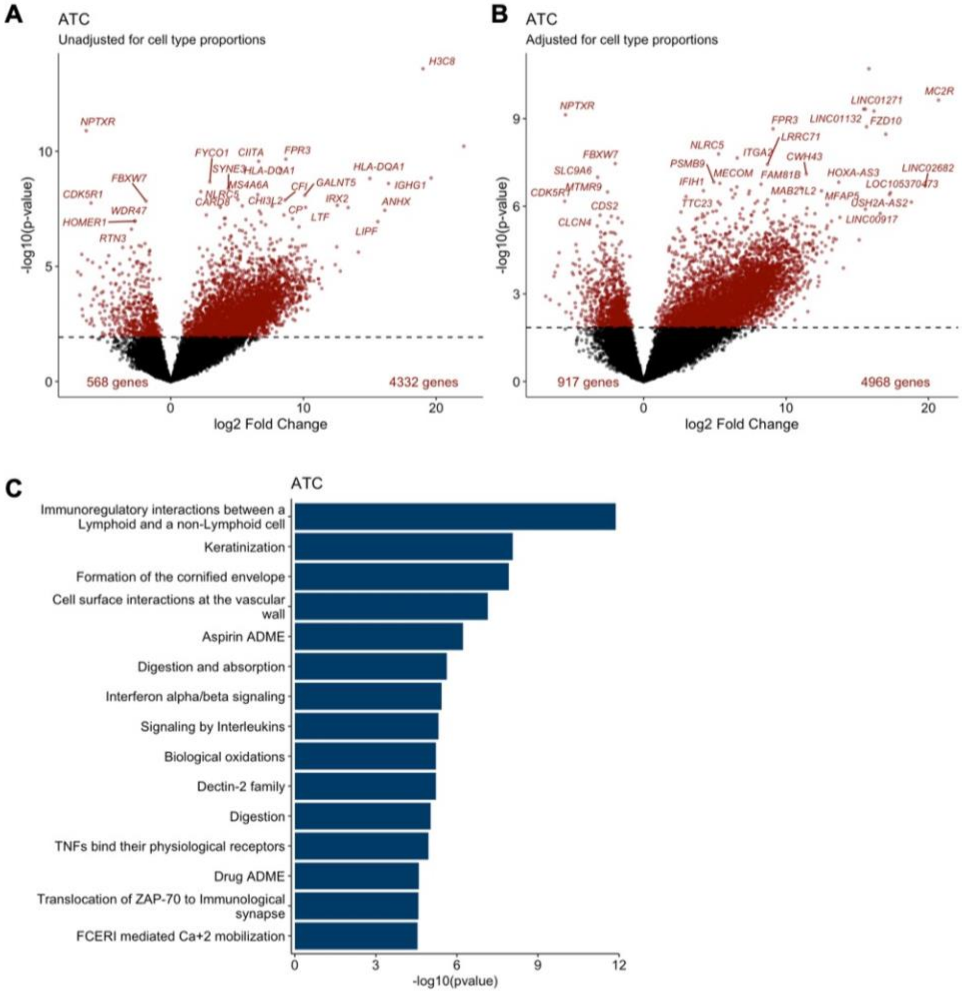
Supplementary Figure 4-7.
Volcano plots of differential **A)** 5-hmC CpGs and **C)** 5-mC CpGs in ependymoma in the cell type proportion unadjusted model. Volcano plots of differential **B)** 5-hmC CpGs and **D)** 5-mC CpGs in astrocytoma in the cell type proportion adjusted model. Labeled # of CpGs on the left of each plot are CpGs with decreased methylation in tumors compared to non-tumor tissue. Labeled # of CpGs on the right of each volcano plot are CpGs with increased methylation in tumors compared to non-tumor tissue. Red points indicate statistically significant differential CpGs under the q-value < 0.05 threshold.



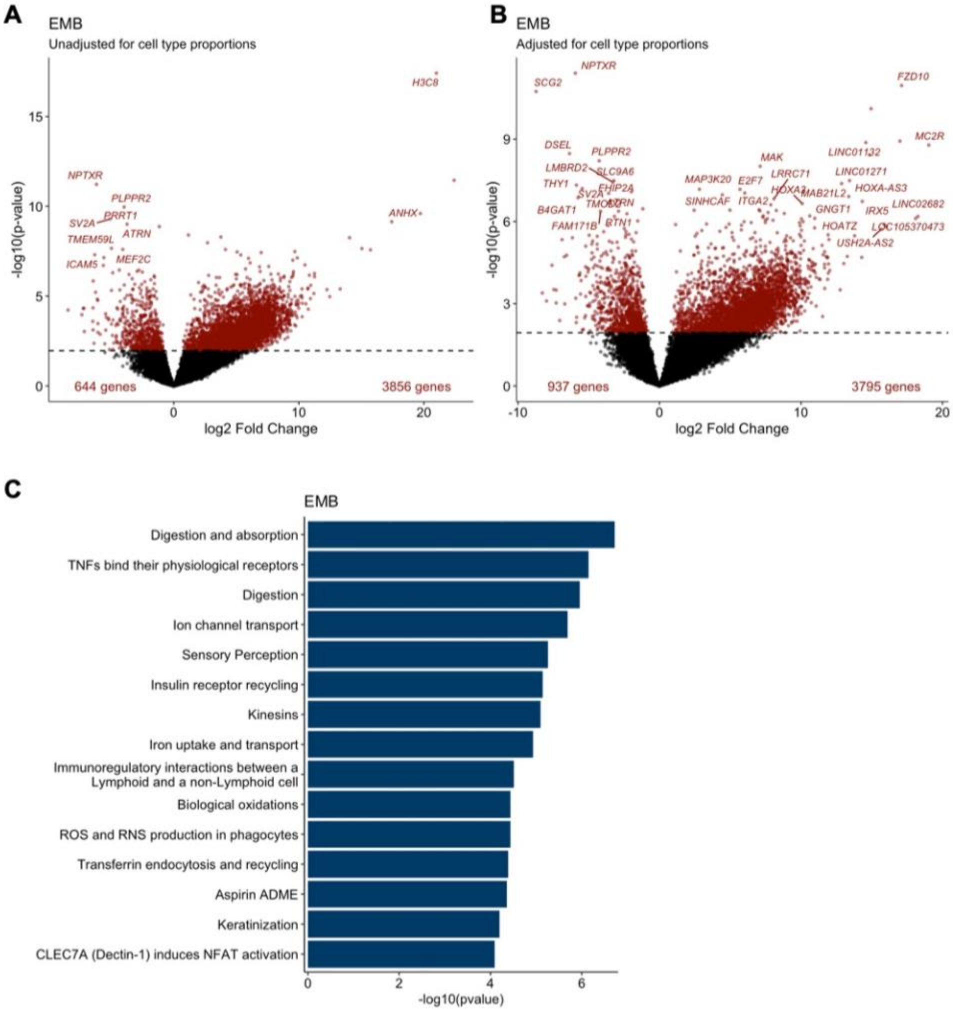
Supplementary Figure 4-8.
Volcano plots of differential **A)** 5-hmC CpGs and **C)** 5-mC CpGs in glioneuronal/neuronal tumors in the cell type proportion unadjusted model. Volcano plots of differential **B)** 5-hmC CpGs and **D)** 5-mC CpGs in astrocytoma in the cell type proportion adjusted model. Labeled # of CpGs on the left of each plot are CpGs with decreased methylation in tumors compared to non-tumor tissue. Labeled # of CpGs on the right of each volcano plot are CpGs with increased methylation in tumors compared to non-tumor tissue. Red points indicate statistically significant differential CpGs under the q-value < 0.05 threshold.



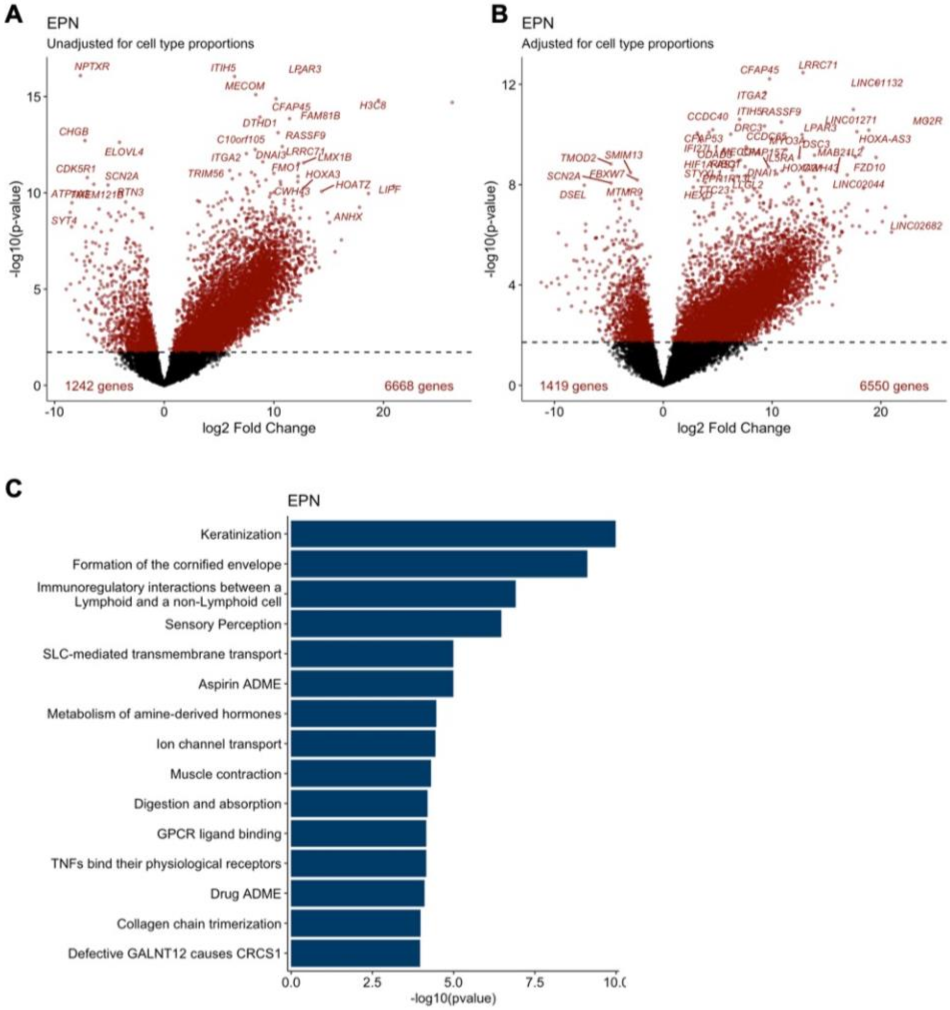
Supplementary Figure 4-9.
A) Number of differentially expressed genes unadjusted and adjusted for cell type proportions for each tumor type. Venn diagram of the genes with significant **B)** increased and **C)** decreased expression the tumor types. **D)** Pathways associated with shared genes with increased expression in the tumors. Only pathways under q-value < 0.05 are shown. **E)** Pathways associated with shared genes with decreased expression in the tumors. Only pathways under q-value < 0.05 are shown.



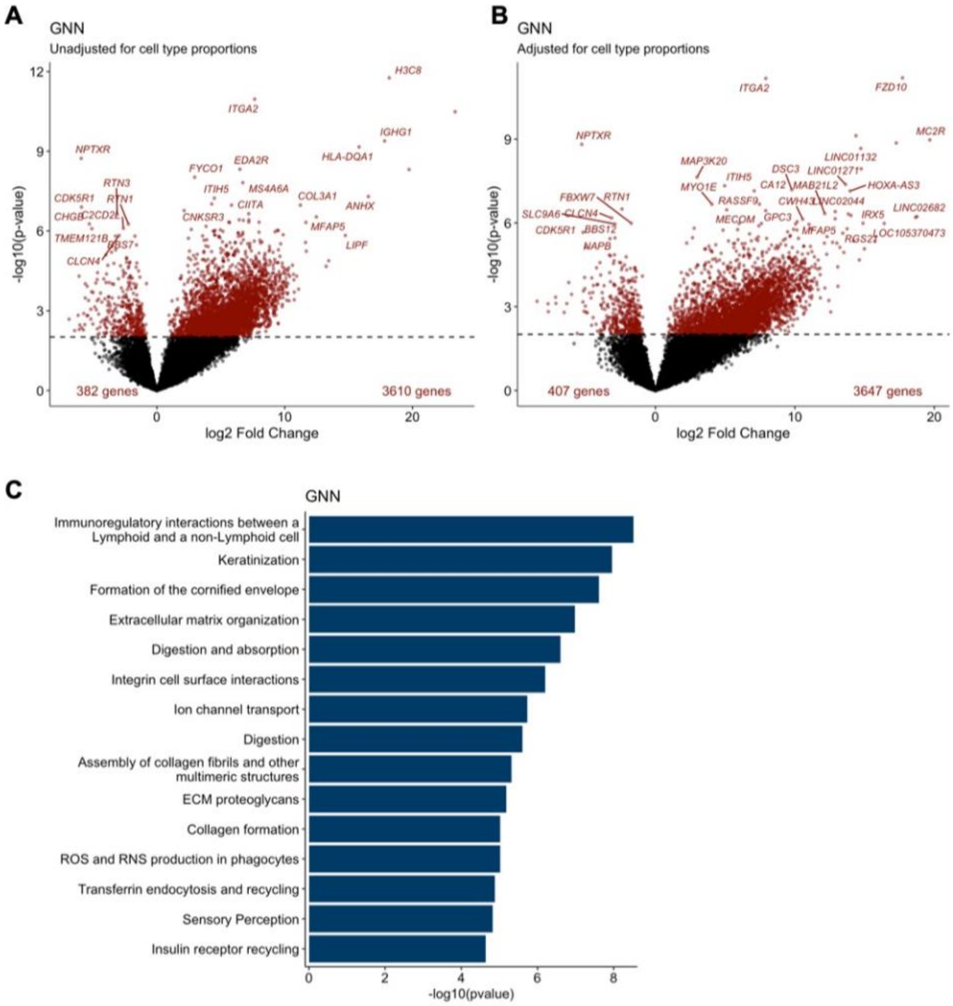
Supplementary Figure 4-10.
Volcano plot of differential expression test in the **A)** cell type proportion unadjusted and **B)** cell type proportion adjusted model comparing astrocytoma and non-tumor brain tissue. **C)** Pathways associated with the differential expression in astrocytoma.



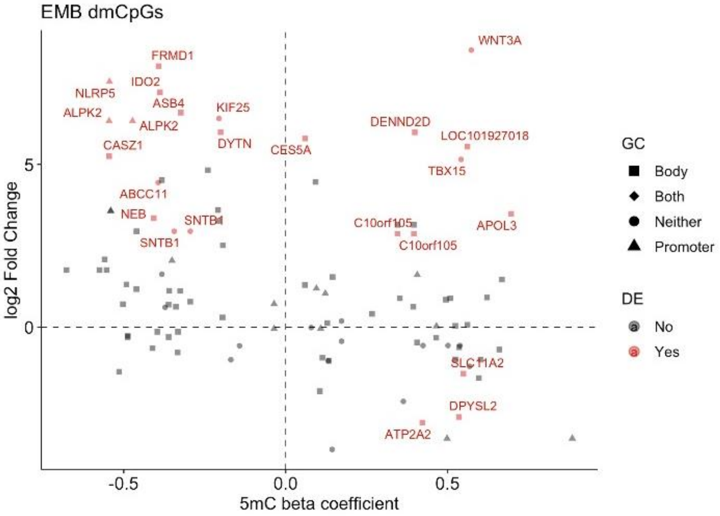
Supplementary Figure 4-11.
Volcano plot of differential expression test in the **A)** cell type proportion unadjusted and **B)** cell type proportion adjusted model comparing embryonal tumors and non-tumor brain tissue. **C)** Pathways associated with the differential expression in embryonal tumors.



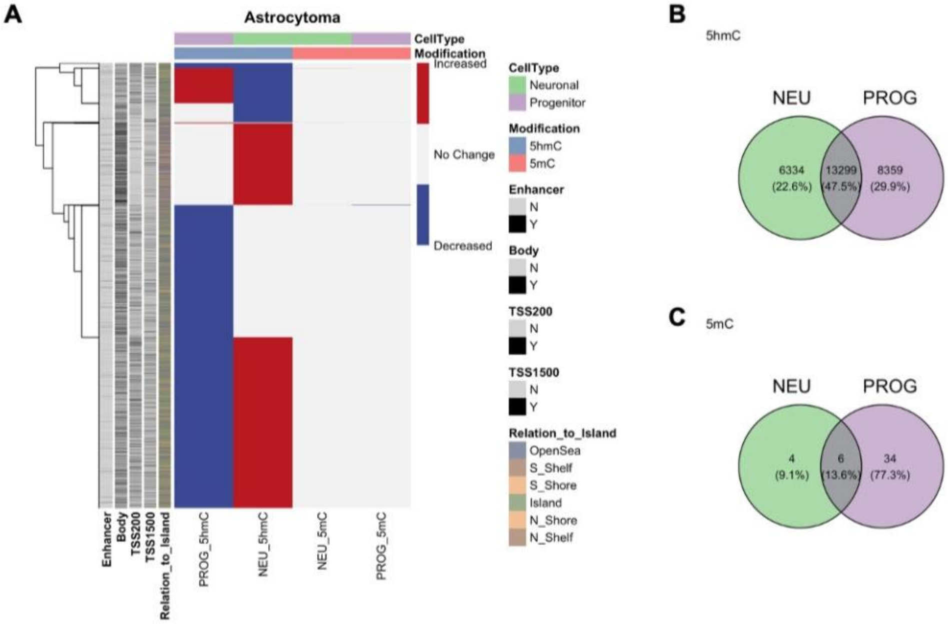
Supplementary Figure 4-12.
Volcano plot of differential expression test in the **A)** cell type proportion unadjusted and **B)** cell type proportion adjusted model comparing ependymoma and non-tumor brain tissue. **C)** Pathways associated with the differential expression in ependymoma.



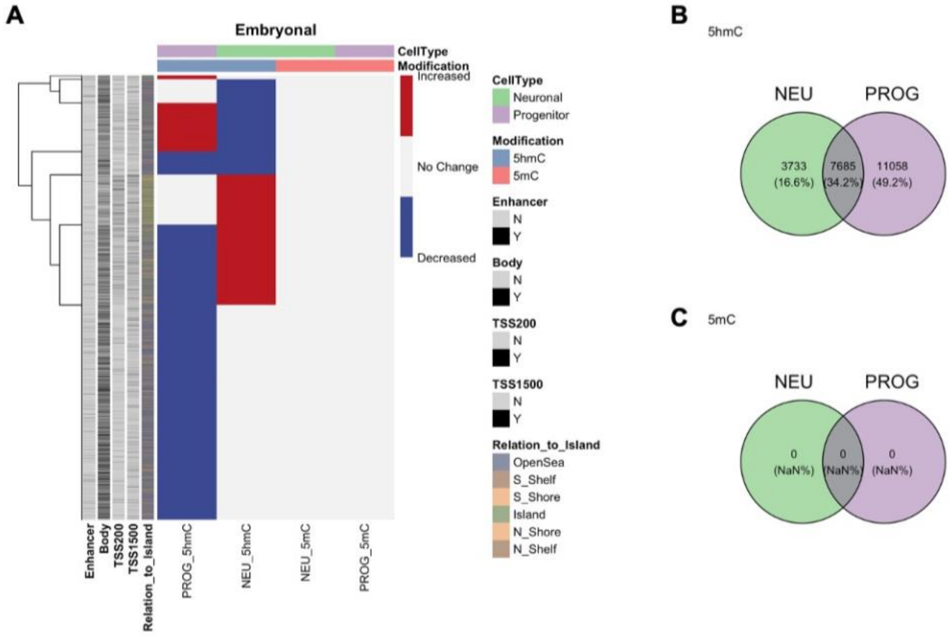
Supplementary Figure 4-13.
Volcano plot of differential expression test in the **A)** cell type proportion unadjusted and **B)** cell type proportion adjusted model comparing glioneuronal/neuronal tumors and non-tumor brain tissue. **C)** Pathways associated with the differential expression in glioneuronal/neuronal tumors.



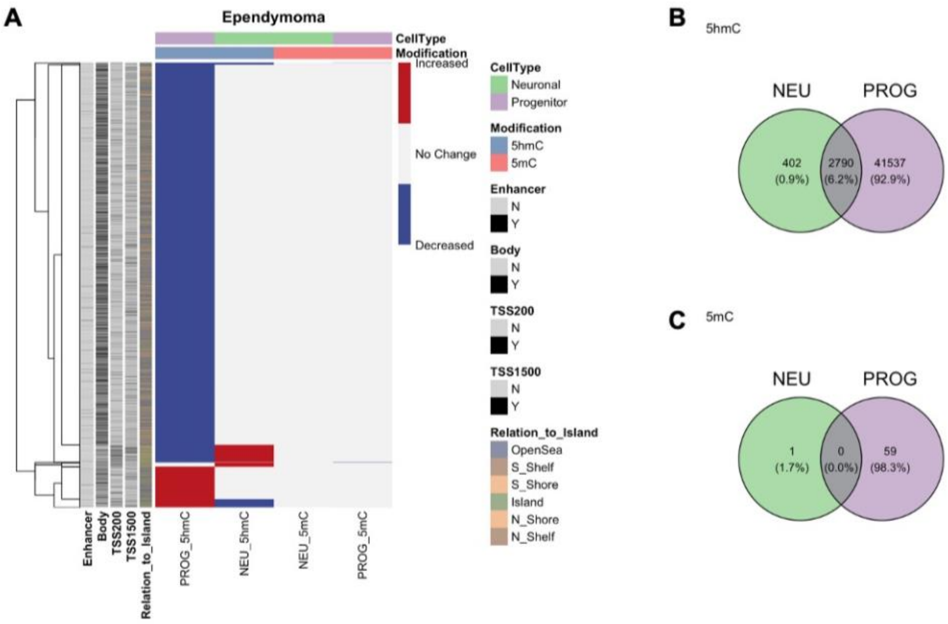
Supplementary Figure 4-14.
Association between changes in 5-mC and gene expression for embryonal tumors. Red points indicate significantly differentially expressed genes. Shapes indicate the genomic context of each CpG.



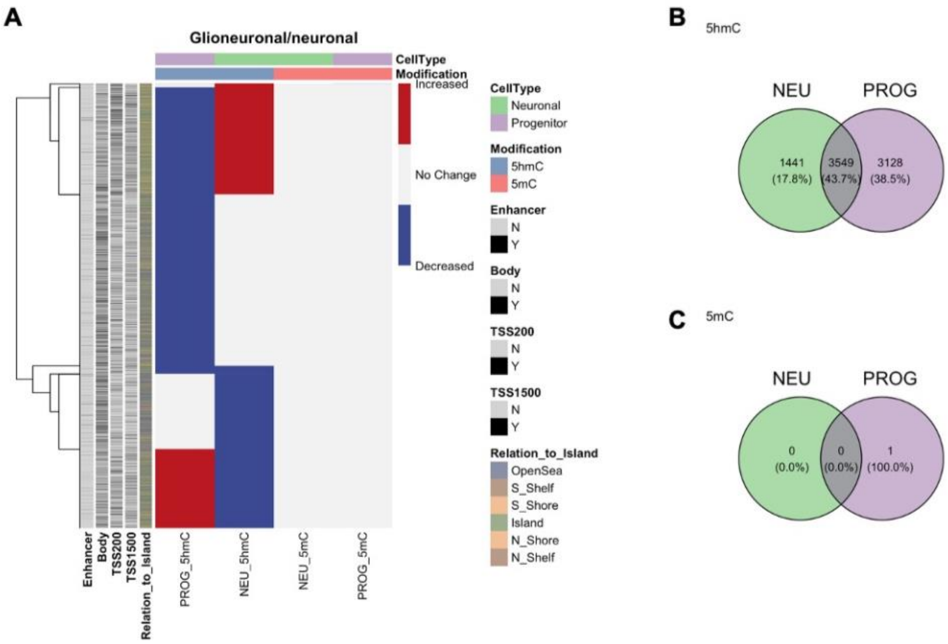
Supplementary Figure 4-15.
A) Cell type specific differentially hydroxymethylated and methylated CpGs in astrocytoma. Venn diagram of differentially **B)** hydroxymethylated and **C)** methylated CpGs in neuronal-like cell types (NEU) and progenitor-like cell types (PROG).



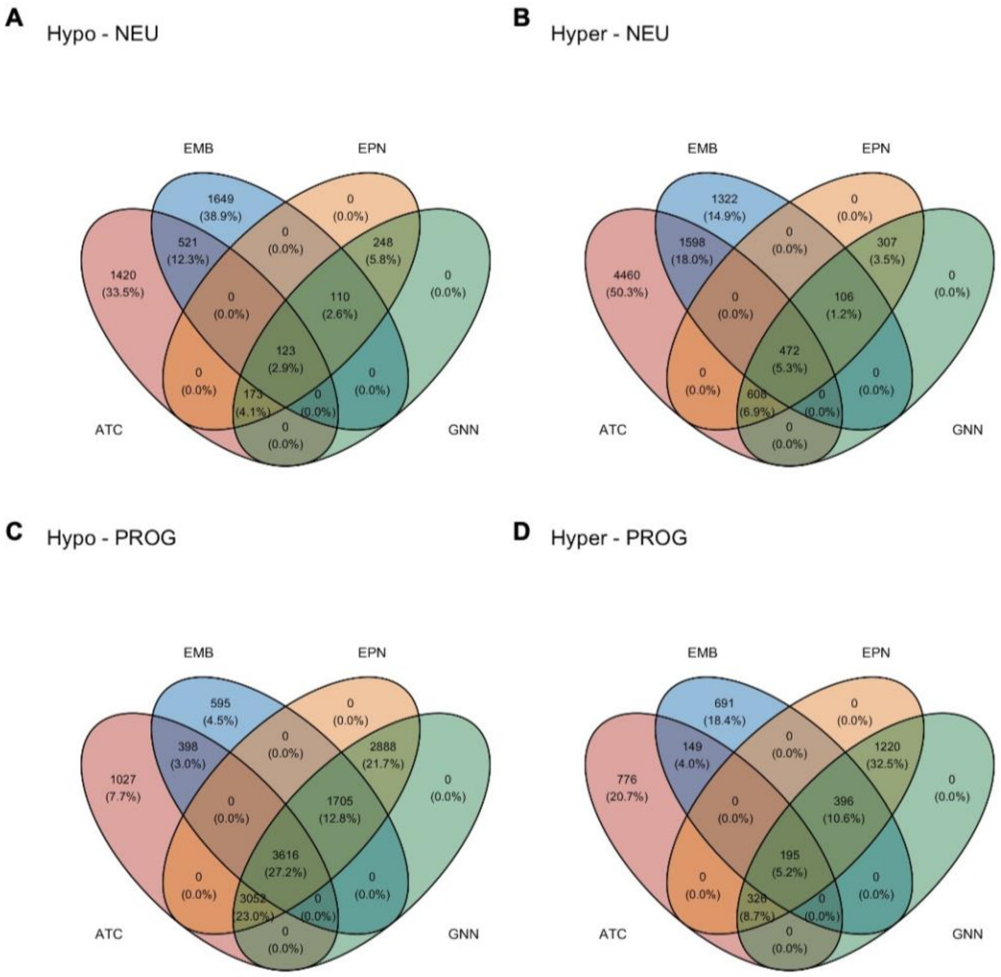
Supplementary Figure 4-16.
A) Cell type specific differentially hydroxymethylated and methylated CpGs in embryonal tumors. Venn diagram of differentially **B)** hydroxymethylated and **C)** methylated CpGs in neuronal-like cell types (NEU) and progenitor-like cell types (PROG).



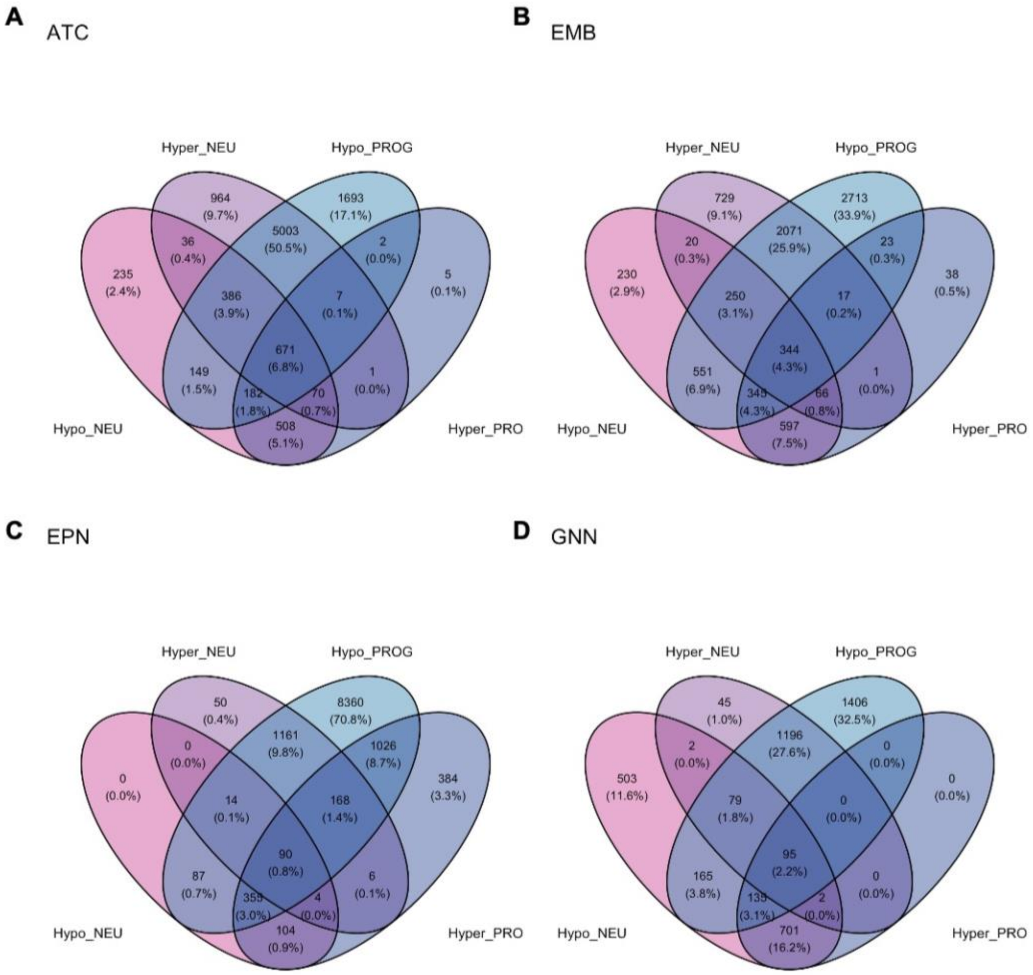
Supplementary Figure 4-17.
A) Cell type specific differentially hydroxymethylated and methylated CpGs in ependymoma. Venn diagram of differentially **B)** hydroxymethylated and **C)** methylated CpGs in neuronal-like cell types (NEU) and progenitor-like cell types (PROG).



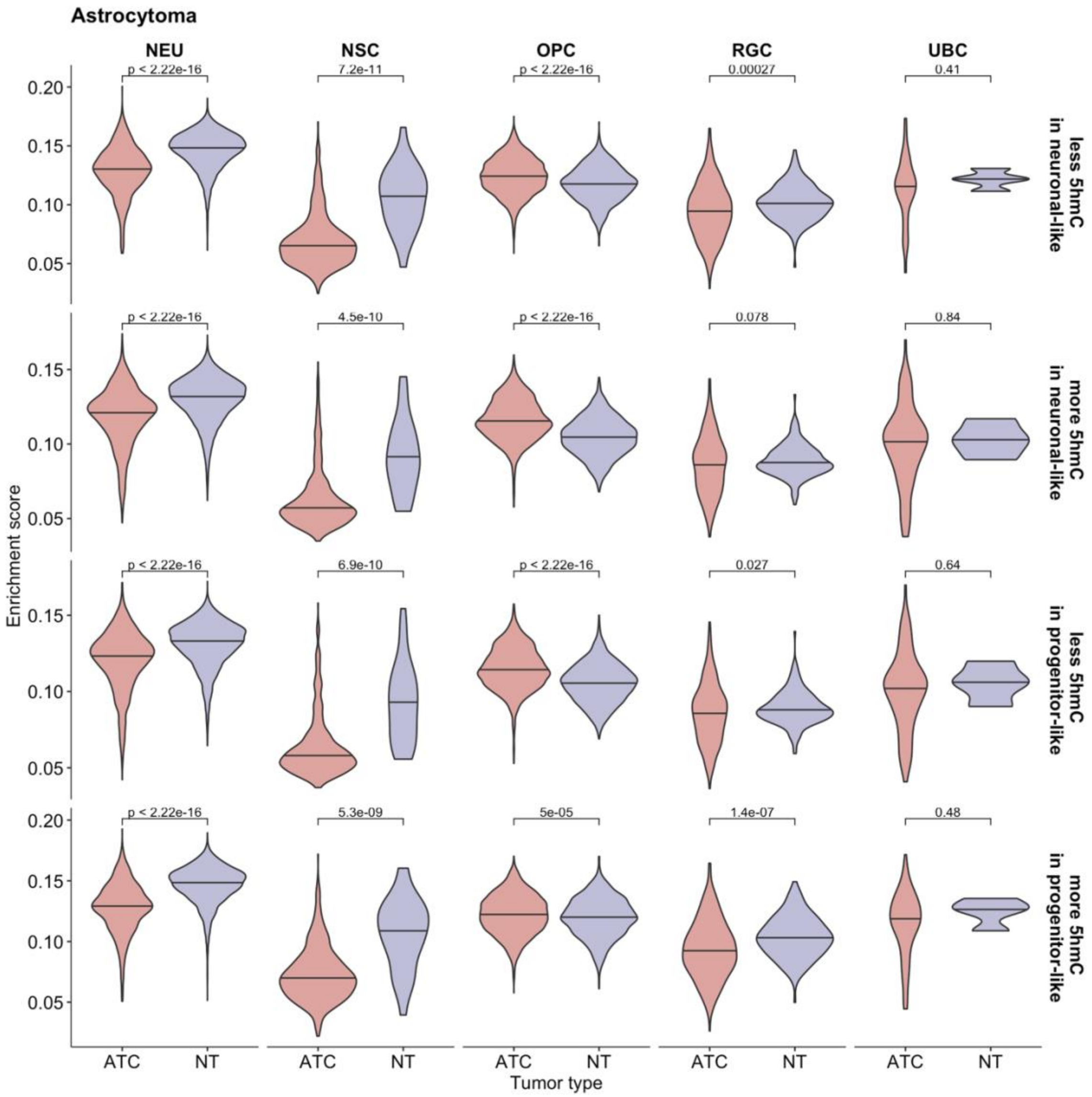
Supplementary Figure 4-18.
A) Cell type specific differentially hydroxymethylated and methylated CpGs in glioneuronal/neuronal tumors. Venn diagram of differentially **B)** hydroxymethylated and **C)** methylated CpGs in neuronal-like cell types (NEU) and progenitor-like cell types (PROG).



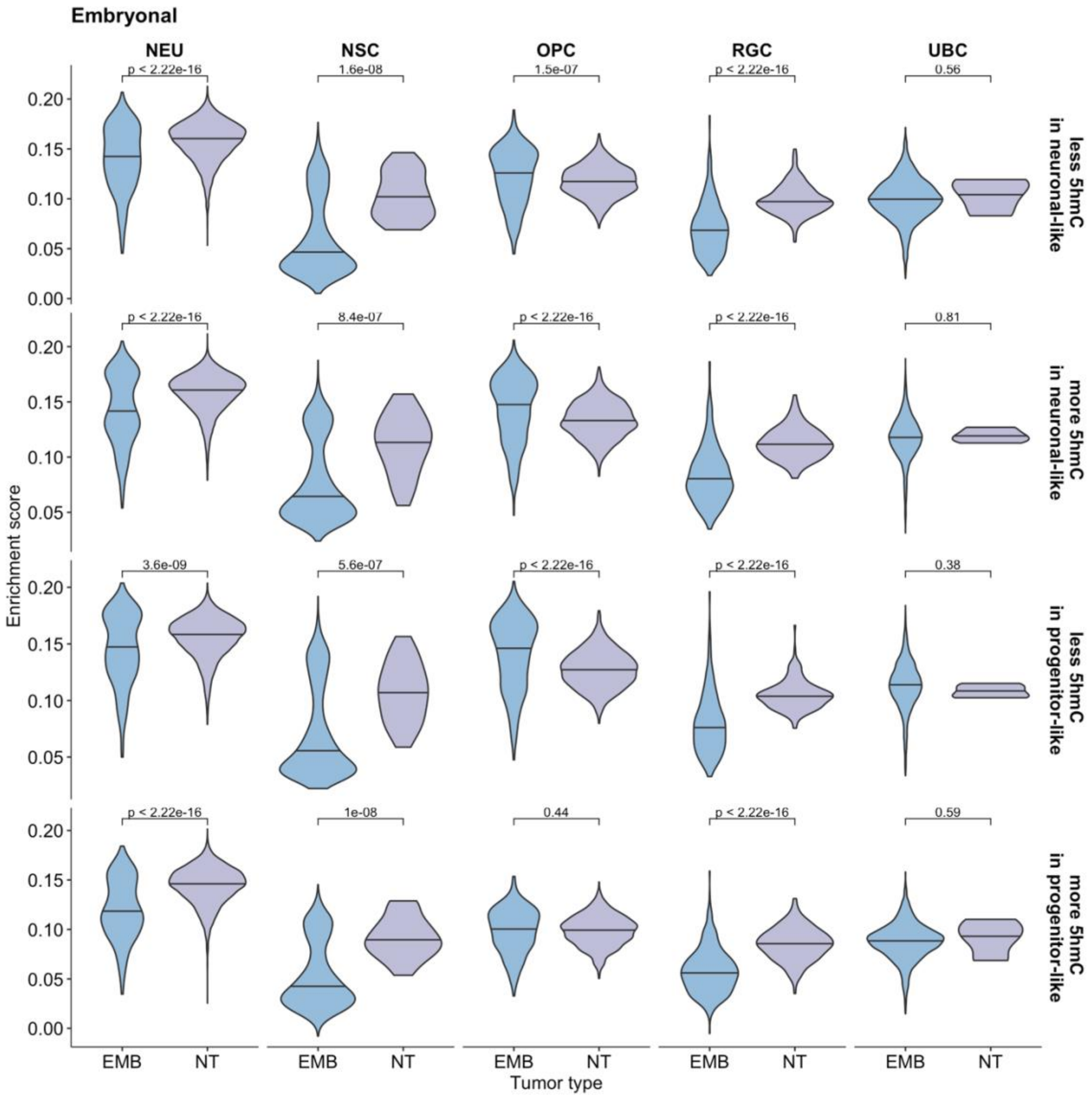
Supplementary Figure 4-19.
Venn diagram of **A)** hypomethylated and **B)** hypermethylated CpGs in neuronal-like cell types across tumor types. Venn diagram of **C)** hypomethylated and **D)** hypermethylated CpGs in progenitor-like cell types across tumor types.



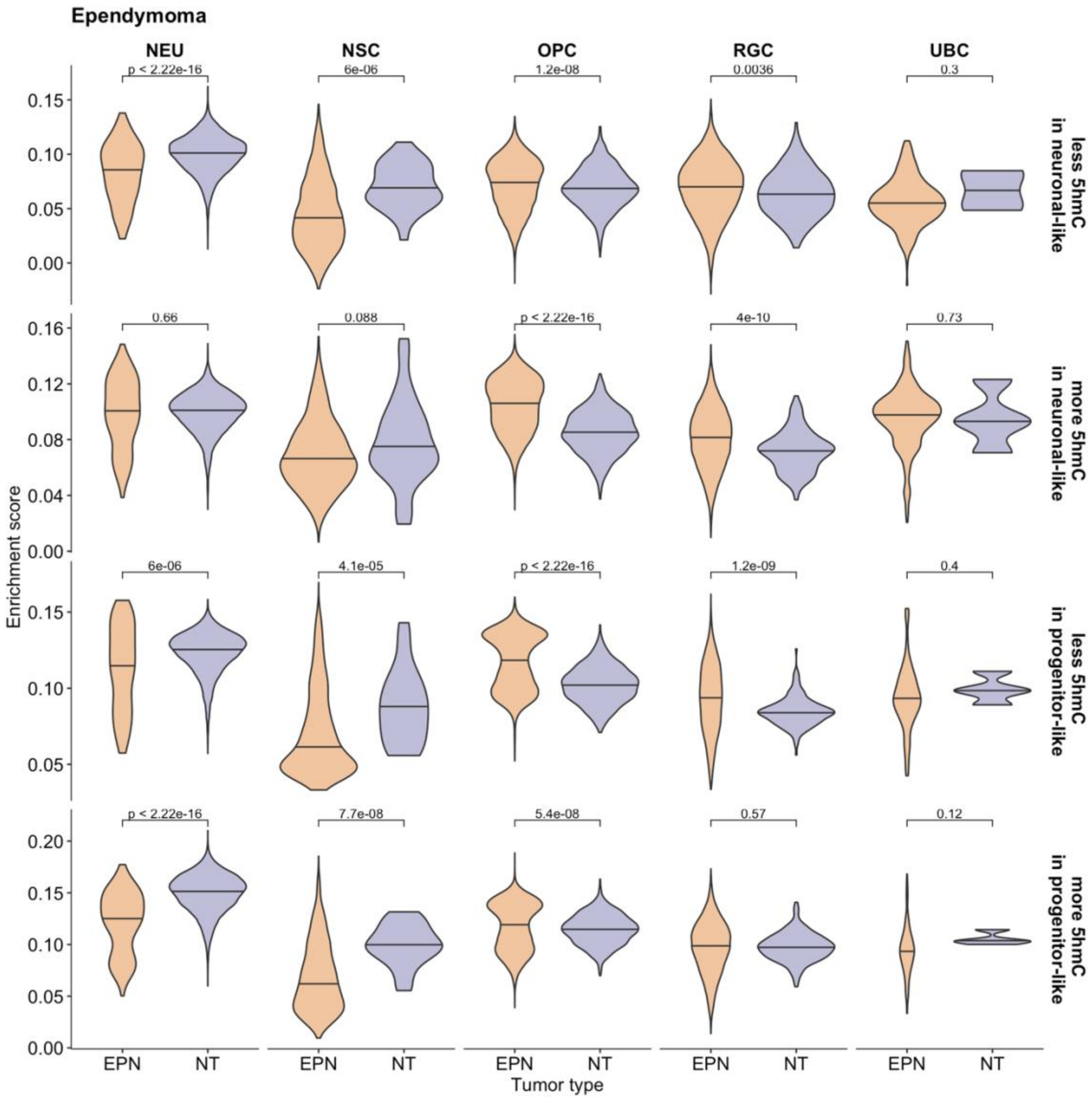
Supplementary Figure 4-20.
Venn diagram of genes with differentially hydroxymethylated CpGs in **A)** astrocytomas, **B)** embryonal tumors, **C)** ependymomas, and **D)** glioneuronal/neuronal tumors.



Supplementary Figure 4-21.
Boxplot of enrichment scores of genes with differentially hydroxymethylated CpGs per cell type in astrocytoma and non-tumor brain tissue. Comparison between tumor and non-tumor made with Wilcoxon rank test.

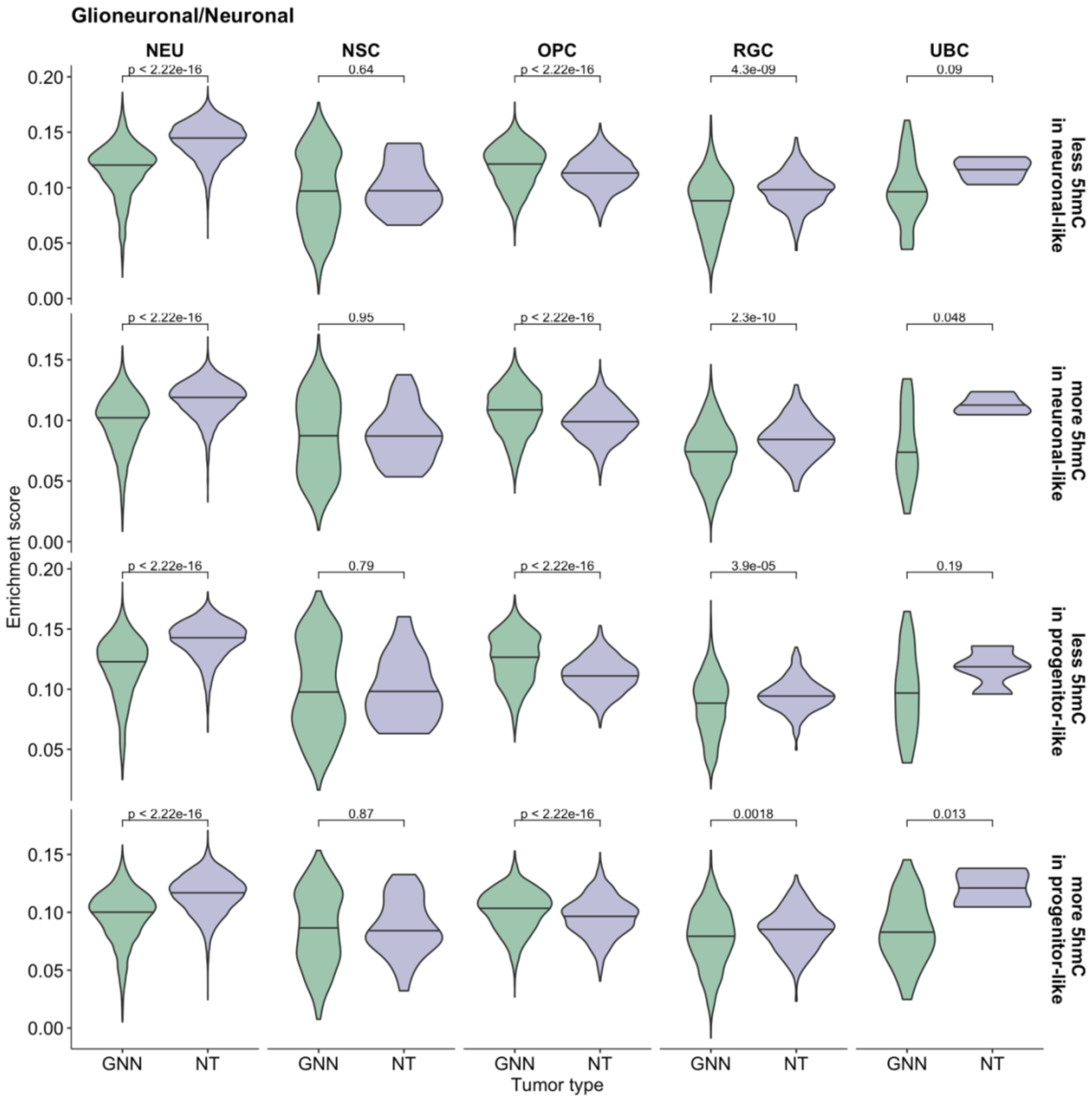


Supplementary Figure 4-22.
Boxplot of enrichment scores of genes with differentially hydroxymethylated CpGs per cell type in embryonal and non-tumor brain tissue. Comparison between tumor and non-tumor made with Wilcoxon rank test.



Supplementary Figure 4-23.

Boxplot of enrichment scores of genes with differentially hydroxymethylated CpGs per cell type in ependymoma and non-tumor brain tissue. Comparison between tumor and non-tumor made with Wilcoxon rank test.



Supplementary Figure 4-24.
Boxplot of enrichment scores of genes with differentially hydroxymethylated CpGs per cell type in glioneuronal/neuronal and non-tumor brain tissue. Comparison between tumor and non-tumor made with Wilcoxon rank test.

Supplementary Table 4-1. Distribution of samples with the varying molecular characterization available for analysis.

	Cytosine modifications	Bulk RNAseq	Single cell RNAseq	Used in manuscript
Non-tumor	4	2	3	2
Astrocytoma	7	7	7	7
Embryonal	6	6	6	6
Ependymoma	12	10	12	10
Glioneuronal/Neuronal	8	8	8	8
Total	37	34	35	33

Supplementary Table 4-2A. Association between tumor purity and grade with 5-hmC MDI

	Estimate	Std Error	P-value
Tumor purity	-0.007	0.005	0.22
G1	Referent		
G2	0.001	0.004	0.90
G3	0.002	0.004	0.68
G4	0.0005	0.004	0.90
NEC/NOS	0.0004	0.006	0.94

Supplementary Table 4-2B. Association between tumor purity and grade with 5-mC MDI

	Estimate	Std Error	P-value
Tumor purity	0.003	0.012	0.80
G1	Referent		
G2	0.028	0.010	0.0074
G3	0.020	0.008	0.025
G4	0.043	0.009	4.5E-5
NEC/NOS	0.016	0.013	0.24

Supplementary Table 4-3. The number of differentially expressed genes with differentially hydroxymethylated CpGs identified by bulk tissue epigenome wide association study.
Categorized by the direction of gene expression change and the genomic context of the differentially hydroxymethylated CpGs for each tumor type.

Astrocytoma					
	Body	Both	Neither	Promoter	Total
Decrease	0	0	0	0	0
Increase	5	0	4	3	12
Total	5	0	4	3	12
Embryonal tumors					
	Body	Both	Neither	Promoter	Total
Decrease	6	0	1	10	17
Increase	44	0	13	17	74
Total	50	0	14	27	91
Ependymoma					
	Body	Both	Neither	Promoter	Total
Decrease	0	0	0	1	1
Increase	1	0	0	4	5
Total	1	0	0	5	6
Glioneuronal/neuronal tumors					
	Body	Both	Neither	Promoter	Total
Decrease	0	0	0	0	0
Increase	6	0	2	5	13
Total	6	0	2	5	13

Supplementary Table 4-4. The number of nuclei from single nuclei RNA-seq that were included in the CellDMC analysis.

Tumor type	**Nuclei N**	**Neuronal-like**	**Progenitor-like**	**NEU**	**NSC**	**OPC**	**RGC**	**UBC**
Astrocytoma	7714	1280	4543	1280	685	3269	485	104
Embryonal	12936	462	7649	462	3380	446	448	3375
Ependymoma	22287	204	19679	204	1695	9582	8120	282
Glioneuronal/Neuronal	16016	3848	4694	3848	731	2189	1730	44
Non-Tumor	17451	8394	1431	8394	29	1224	174	4

Supplementary Table 4-5. The number of differentially hydroxymethylated CpGs per tumor type identified by CellDMC. Categorized by the change in level of hydroxymethylation and cell type of association.

	ATC	EMB	EPN	GNN
Hypo-hydroxymethylated in neuronal-like cells	3741	4829	1031	2963
Hyper-hydroxymethylated in neuronal-like cells	15892	6589	2161	2027
Hypo-hydroxymethylated in progenitor-like cells	2270	16099	40216	5233
Hyper-hydroxymethylated in progenitor-like cells	2270	2644	4111	1444

Chapter 5

5. Discussion

5.1. Overview of findings

5.1.1. Chapter 2: Distinct cytosine modification profiles define epithelial-to-mesenchymal cell-state transitions

Epithelial-to-mesenchymal transition (EMT), a cellular program important in normal embryogenesis and wound healing, is one of the mechanisms that contributes to intratumoral heterogeneity and leads to tumor progression and metastasis[322]. Cells do not switch from an epithelial cell type to a mesenchymal cell type like a binary switch in phenotype. Instead cells gradually transition by shifting to various intermediate cell states between the two fully differentiated cell types[322]. Due to challenges in isolating intermediate EMT cell states, understanding of the molecular underpinnings of intermediate EMT cell states is still limited. While studies have begun to characterize the transcriptome and chromatin structures of the different intermediary states in EMT, limited data exist on DNA cytosine modifications profiles across EMT states.

As DNA cytosine modifications are critical in normal developmental processes like EMT, we aimed to investigate the DNA cytosine modifications during EMT in cancer. In a previously developed model of single cell clones from heterogeneous ER/PR-negative breast cancer cell lines, we utilized a multi-omic approach which included measures of DNA cytosine modifications, chromatin accessibility and gene expression. From the start, we observed more drastic differences in the hydroxymethylation profiles of more intermediate cell states compared to the more differentiated cell states, rather than in the methylation profiles. We identified 17,862 CpGs with increasing 5-hmC and 7,903 CpGs which were mostly decreasing in 5-mC in the intermediate clones. The CpGs with increasing 5-hmC levels included CpGs that tracked to key EMT associated transcription factors like *SNAI1* and *TWIST1*, and epithelial or mesenchymal cell type markers like *CDH1* and *MMP19.* The open chromatin regions containing CpGs with increased 5-hmC were associated with Rho family of GTPases which have been shown to function as cellular switches in coordinating cell polarity and migration by regulating the cytoskeleton. Furthermore, open chromatin regions with CpGs with increased 5-hmC were enriched in motifs of EMT transcription factors like ZEB1 and SNAI2.

Chapter 2 addresses the gap in understanding of the role of DNA cytosine modification marks, particularly hydroxymethylation marks, in regulating EMT. The results from this chapter also highlight the utility of a multi-omic approach to gain better understanding of how the different epigenetic systems coordinate to regulate dynamic processes like EMT.

5.1.2. Chapter 3: Tumor type and cell type-specific gene expression alterations in diverse pediatric central nervous system tumors identified using single nuclei RNA-seq

In Chapter 3, we switched our focus to focus on appreciating the intratumoral heterogeneity in primary tumors of pediatric central nervous systems (CNS), a relatively understudied tumor type. Pediatric CNS tumors are difficult to study as they occur very rarely in the general population, with an incidence rate of 3.57 per 100,000 for malignant types and 2.65 for non-malignant types[334]. Difficulty in sample accrual to characterize and understand the different types of pediatric CNS tumors has led to slower progress in developing targeted therapies for these tumors. The survival rates vary among tumor types, with a 96.8% 5-year relative survival rate in pilocytic astrocytoma and 19.8% 5-year relative survival rate in glioblastoma[335]. Even pediatric CNS tumor patients who go on to be in remission and survive their primary tumors are at risk of higher disabling conditions from the harsh treatments and the tumor itself[337]. To improve poor survival rates and to reduce the extremely high burden of disabling conditions post-tumor, better treatment options and management strategies are needed for pediatric CNS tumors.

In this chapter, we focused on characterizing the heterogeneity and determining the transcriptomic alterations in the pediatric CNS tumors by performing single nuclei RNA-seq on 84,700 nuclei from 35 tumors and 3 non-tumor pediatric brain tissue. Major cell subpopulations were associated with specific tumor types. For example, we identified significant proportions of oligodendrocyte precursor cell populations in astrocytomas and sizeable proportions of unipolar brush-like cells with high stemness in embryonal tumors. Our results delineated clear transcriptomic alterations between tumors and non-tumor cells within the same cell types. Furthermore, we distinguished pathways enriched in cell types of interest for therapy resistance and tumor progression, like Aurora-B kinase and retinoic acid pathways in the neural stem cells.

Additionally, this chapter highlighted the importance of considering cell type composition effects on transcriptomic alterations when comparing tumors to non-tumor tissue. Although the number of genes that were significantly differentially expressed in the cell type identity adjusted model were less than that of the unadjusted model, we identified novel genes and pathways that would not have been determined to be associated with each tumor type. We distinguished pathways that would otherwise have been obscured that were associated with differentially expressed genes in the tumors by adjusting for cellular identity. These pathways included translation associated processes and interferon gamma signaling.

We expanded on previously published bodies of work that demonstrated the heterogeneous nature of pediatric CNS tumors by adding to the small patient tumors that had been published and by adding novel tumor types that have not yet been characterized using single cell genomics technologies. We also compared cell type populations in various pediatric CNS tumor types to non-tumor pediatric brain tissue that has not been explored previously to our knowledge. The results from this chapter suggest cell type-specific and tumor type-specific targets for potential therapies. Lastly, we advocate for the consideration of differences in cell type composition when comparing tumors to non-tumor tissues.

5.1.3. Chapter 4: Hydroxymethylation alterations in progenitor-like cell types of pediatric central nervous system tumors are associated with transcriptional changes

While recent single cell genomics technologies have significantly improved the way we can characterize single cell types and cell states as shown in Chapter 3, due to the barriers in cost and challenges in computational analysis, bulk tissue characterization and comparisons remain very consistently used. While bulk tissue molecular measures may not provide the level of granularity as in single cell genomics technologies, cell type composition effects can be accounted for using computational methods.

As we observed in Chapter 3, pediatric CNS tumors are composed of heterogeneous cell types. Most studies investigating DNA cytosine modifications in pediatric CNS tumors have mainly applied bulk tissue approaches. While most other published studies only utilized bulk tissue datasets, we were able to complement data our bulk tissue cytosine modification data with matching single nuclei RNA-seq data. We again applied a multi-omic approach by integrating cytosine modification profiles, bulk tissue and single cell gene expression profiles to investigate the epigenomic alterations at the cell type level and its effects on the transcriptome.

Like the cell type composition effects on transcriptomic alterations found in Chapter 3, we demonstrated the same cell type composition effects on DNA cytosine modification alterations in the pediatric CNS tumors when compared to non-tumor pediatric brain tissue. When proportions from major cell types present in the tumors and non-tumor tissue were incorporated into the models for epigenome wide association studies, the number of differentially hydroxymethylated CpGs and differentially methylated CpGs decreased drastically. In some tumor types, no significantly differentially hydroxymethylated or methylated CpGs were observed when adjusting for cell type proportions. These results suggested that differentially hydroxymethylated or methylated CpGs from our EWAS analyses in our small sample size were almost all due to cell type composition differences.

Despite a reduced scope of CpGs whose modifications were associated with pediatric CNS tumors compared to normal tissue, we utilized a computational approach called CellDMC[37] to identify CpGs that were associated with the tumors at a cell type-specific level. At the cell type level, especially in progenitor-like cell types, we identified thousands of CpGs that were differentially hydroxymethylated and even identified differentially methylated CpGs that we were not able to detect at all from bulk tissue EWAS analyses with adjustment for cell type. The differentially hydroxymethylated CpGs in different tumor types were enriched in separate genomic contexts, suggesting tumor type specific and cell type specific changes in hydroxymethylation contributing to the underlying tumor biology.

Associations between hydroxymethylation and gene expression especially at the cell type level have not been published in any tumor contexts to our knowledge. From our integrative multi-omic approach, the results from this chapter revealed the relationship between changes in gene expression with differential hydroxymethylation in neuron-like cell types and in oligodendrocyte precursor cells across almost all pediatric

CNS tumor types included in our study. Oligodendrocyte precursors cells had higher gene expression levels for genes that had cell type specific differential hydroxymethylated CpGs. Neuronal-like cell types had decreased gene expression levels for genes that had cell type specific differentially hydroxymethylated CpGs. Two genes widely suggested to play a role in tumor progression, *HDAC4* and *IGF1R,* were couple of examples of genes with differential hydroxymethylated CpGs and differential gene expression at the cell type specific level.

Chapter 4 demonstrates epigenetic heterogeneity in pediatric CNS tumors and the significant cell type-specific aberrations that exist in hydroxymethylation compared to non-tumor pediatric brains. In addition, chapter 4 brings forth more clarity in the potential roles of 5-hmC in regulating cell type-specific gene expression.

5.2. Perspectives and future directions

Research in the epigenetics field in the past few decades has established the role of DNA methylation in many biological processes. However, studies of locus-specific states and alterations of DNA hydroxymethylation have emerged much more recently. In this thesis, I demonstrate the cell type specificity and the critical roles of DNA hydroxymethylation in 1) epithelial-to-mesenchymal transition, 2) tumoral hydroxymethylation alterations, with major distinctions from DNA methylation.

As epithelial-to-mesenchymal transition is a dynamic process, many studies have focused more on molecular features that are more likely to be transient such as gene expression or chromatin accessibility. However, as previous studies have shown that DNA methylation and hydroxymethylation play key roles in normal developmental processes, we aimed to understand cytosine modifications in single cells that are undergoing EMT. In Chapter 2, we identified that intermediate/hybrid cell states particularly have high levels of 5-hmC and are particularly distinct from the differentiated cell states. As, intermediate EMT states are associated with higher levels of invasion and metastasis in tumors[363], opportunities to exploit hydroxymethylation or TET enzymes to halt EMT progression exist. *In vitro* and *in vivo* models to test for such hypothesis are needed to experimentally validate whether interruption of EMT may be possible through altering hydroxymethylation.

In addition, as our results illustrate the distinct cytosine modification profiles that exist in the various EMT cell states, cell state/type deconvolution methods with both hydroxymethylation and methylation may be utilized as biomarkers for assessing potential for tumor progression and metastasis in patient samples. Current methods in estimating EMT levels are generally restricted to gene expression or protein levels[579]. DNA cytosine modification-based methods to complement previously developed methods can improve accuracy to estimate tumor progression and metastasis. To develop a DNA cytosine modification-based deconvolution method for EMT, primary tumor samples of various tumor types that have been associated with EMT should be used. The primary tumors can be dissociated and FACS-sorted to isolate multiple cell states in the EMT processes in real tumors. Differentially methylated and hydroxymethylated CpGs associated with each isolated EMT cell states can be determined to develop libraries for deconvolution. The deconvolution method then can be used to first, compare FACS-based quantification of EMT cell state proportions, second, be compared to existing methods for assessing EMT levels, and lastly, test for associations between EMT cell state proportions and tumor progression and metastasis.

In the past five decades, the mortality rates for many cancer types have drastically been reduced. While survival rates within pediatric CNS tumor types vary, collectively, these tumors have still not have had the drastic reduction in mortality rates seen in some other cancer types like hematological malignancies. Moreover, pediatric CNS tumor patients face the highest cumulative chronic conditions after surviving their initial tumors. The burden is largely due to the limited type of treatment options that are currently available for pediatric CNS tumors. Many patients undergo radiation therapy as one of their main treatment options. Radiation has been associated with later in life health effects such as decline in intellectual ability, strokes, seizures, short term memory decline, and neuromuscular dysfunction[337–339,342,499]. To improve survival rates and quality of life after tumors, novel treatment options and strategies need to be developed for pediatric CNS tumors. To contribute to those efforts, we furthered the current knowledge of transcriptome and epigenome of pediatric CNS tumors in chapters 3 and 4. Our dataset contributes >40% increase in pediatric CNS tumor sample size and >74% increase in number of pediatric CNS tumor nuclei available for analysis. Moreover, pediatric CNS tumor datasets that have measured single nuclei RNA-seq with DNA cytosine modifications on the same tumors do not yet exist to our knowledge. Furthermore, we demonstrate shared transcriptomic and epigenomic alterations to suggest a tumor type-agnostic approach in identifying potential therapeutic targets as these tumors are very rare in the general population. Results from chapter 3 indicates that epigenomic alterations are essential features of pediatric CNS tumors, and additional studies are needed to expand the understanding of other epigenomic features, like histone modifications and chromatin structures. Multi-omic approaches, like ones used for this body of work, for multiple epigenetic systems, will improve our understanding of the critical epigenomic aberrations that underlies in these tumors. As some single cell level epigenomic technologies like single cell ATAC-seq are commercially and more readily available, chromatin accessibility at the single cell level will complement what we have found with our single cell gene expression profiles and cell type dependent DNA cytosine modification alterations.

In addition to furthering our understanding of the underlying biology of pediatric CNS tumors, epigenomic profiling may serve to identify potential targets for treatment. Limited clinical trials are investigating epigenetic modifier drugs in pediatric CNS tumors. We identified some epigenetic modifiers like *HDAC4* that have differentially hydroxymethylated CpGs and differential gene expression across all our pediatric CNS tumor types. Experimental evidence from *in vitro* and preclinical models targeting epigenetic modifiers may lead to the development of novel approaches for treatment and may determine if survival and quality of life post-tumor could be improved. Few HDAC inhibitors already have been FDA approved for other cancer types[580]. As these drugs have already been tested for safety, it may be clinically beneficial to repurpose these drugs for pediatric CNS tumors.

In conjunction with potential neoadjuvant therapy targeting epigenetic modifiers, future experiments testing for prospective adjuvant therapy is warranted. One source of recurrence and therapy resistance in pediatric CNS tumors are the neural stem cells. In chapter 3, our single cell gene expression experiment revealed sets of potential neural stem cell type-specific targetable pathways like Aurora B kinase pathway and retinoic acid pathway. From these sets of pathways, *in vitro* and *in vivo* models should be used to validate each of the significant pathways. These results allow us to develop adjuvant therapies for specific cell types that remain after surgical resections. Pediatric CNS tumor mice models can be used to test if given adjuvant therapies for targeting validated neural stem cell specific pathways will prevent recurrence or metastasis after primary surgical resection.

Additional molecular characteristics on pediatric CNS tumors may be delineated by featuring our current dataset. With our Illumina Human Methylation EPIC arrays on bulk tumor tissue, we determined some copy number variations in our pediatric CNS tumors, particularly gain of chromosome 1q in the ependymomas. Chromosome 1q gain in ependymoma have been well documented[581–583]. While not too many copy number alterations were identified, few other copy number alterations that were detected varied among tumor types. Cell type level copy number alterations by incorporating the single nuclei RNA-seq data may identify additional copy number alterations that may have been obscured due to cell type composition of the tumors. It may also identify cell type associated copy number alterations that may drive transcriptomic alterations and cytosine modification alterations associated with the pediatric CNS tumor types. Furthermore, although there were not too many non-tumor nuclei in most of our pediatric CNS tumor samples, these sample matched non-tumor nuclei may be utilized to identify germline variants in these patients. While it may be difficult to call variants due to limited depth of sequencing per nuclei, it still may be useful to identify the more prevalent germline variants. This additional layer of germline variants may facilitate identifying more pertinent molecular alterations associated with pediatric CNS tumors.

Adult brain and CNS tumors and even some pediatric CNS tumors have been shown to be composed of heterogeneous cell types[285,297,303–305,444,501,526,527,529,530,584–587]. Supporting previous literature, our results from chapter 3 identify and enumerate the heterogenous cell types that exist in pediatric CNS tumors. With a multi-omic approach, our results from chapters 3 and 4 highlight the influence of heterogeneous cell type effects on identification of gene expression and cytosine modifications. While we could not deconvolute the more granular cell types in our data due to the current lack of reference-based deconvolution methods for brain tissue or brain tumors, we were able to utilize our single nuclei RNA-seq data to understand the cell type populations that exist in these tumors. Moving forward, bulk tissue deconvolution methods specifically to be used in normal or tumor brain tissue will be very beneficial to remove cell type specific effects and to study cell-specific programs and alterations in tumors. Our results strongly suggest that cell type specific 5-hmC marks should absolutely be incorporated in deconvolution approaches that utilize DNA cytosine modifications.

In the DNA methylation-based classification method for CNS tumors, traditional bisulfite treatment is used for methylation arrays. As mentioned in previous chapters, traditional bisulfite treatment cannot distinguish between 5-mC and 5-hmC. A previous investigation in the lab demonstrated that when only 5-mC specific signals from oxidative bisulfite treated DNA are used for this method, the classifications changed from 5-mC + 5-hmC signals from bisulfite treated DNA[243]. In addition, our results from chapter 4 suggest a greater role of 5-hmC than 5-mC in the pediatric CNS tumors. Therefore, there exists a potential for substantial improvement in accuracy and specificity of tumor type classification when 1) cell type heterogeneity is taken into context and 2) 5-hmC is incorporated in the methylation-based classification methods.

Because pediatric CNS tumors are so rare in the population, it is difficult to accrue large enough sample size for high powered statistical analyses in a single center. The samples used in chapters 3 and 4 were collected over a period of more than 20 years. While we have collected a sample size that are relatively larger than some of the other published single cell genomics studies, it is still not particularly large. Moreover, the cohort in these chapters is also restricted to a rural and ethnically homogenous population in northern New England. To capture a broader and more generalizable pediatric CNS tumor patient population, collaborative efforts from multiple institutions are needed. Collaborative studies with demographically and geographically diverse groups like International Childhood Cancer Cohort Consortium, Children's Oncology Group, or Pediatric Brain Tumor Consortium would improve generalizability and statistical power of our studies of hydroxymethylation roles in pediatric CNS tumors.

In this body of work, I showed the importance of considering 5-hmC in addition to 5-mC. While hydroxymethylation marks have high potential of serving as effective biomarkers in other published studies[588–592], there are few limitations that need to be improved upon before it can be mainstreamed in the clinical space. One of the biggest challenges with 5-hmC is that it is unable to detected after the tissue has been fixed in to FFPE blocks due to formalin interaction with the hydroxymethyl mark. To be broadly used in clinical settings, fresh frozen tissue needs to be preserved or DNA must be extracted prior to fixation as currently most surgically resected or biopsied tissues are currently stored in FFPE.

5.3. Concluding remarks

The studies included in this thesis explore the cell type-specific molecular heterogeneity that exists in various cancer cell states and cell types. It also expands the field's current understanding of DNA cytosine modifications in different cancer contexts. While most studies have really focused on 5-mC up to this point, studies within this thesis suggests that 5-hmC must be investigated separately from 5-mC. In addition, works from this thesis suggests that incorporating both 5-mC and 5-hmC will dramatically improve computational methods for cell type deconvolution and tumor type classification. Lastly, the combined works of this thesis further establishes the importance of accounting for cell type composition in transcriptomic and epigenomic investigations that use bulk tissue as it can obscure critical molecular underpinnings of diseases and hinder progress in understanding disease biology and developing novel therapies.

References

1. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann Hum Biol* 40, 463–471 (2013).

2. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *Plos Biol* 14, e1002533 (2016).

3. Liu, Z. & Zhang, Z. Mapping cell types across human tissues. *Science* 376, 695–696 (2022).

4. Consortium*, T. S. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* 376, eabl4896 (2022).

5. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* 376, eabl4290 (2022).

6. Conde, C. D. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Sci New York N Y* 376, eabl5197–eabl5197 (2022).

7. Suo, C. *et al.* Mapping the developing human immune system across organs. *Sci New York N Y* 376, science.abo0510 (2022).

8. Coskun, A. F., Eser, U. & Islam, S. Cellular identity at the single-cell level. *Mol Biosyst* 12, 2965–2979 (2016).

9. Zeng, H. What is a cell type and how to define it? *Cell* 185, 2739–2755 (2022).

10. Morris, S. A., Klein, A. & Treutlein, B. The evolving concept of cell identity in the single cell era. *Development* 146, dev169748 (2019).

11. Mincarelli, L., Lister, A., Lipscombe, J. & Macaulay, I. C. Defining Cell Identity with Single-Cell Omics. *Proteomics* 18, 1700312 (2018).

12. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* 18, 83 (2017).

13. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* 25, 1491–1498 (2015).

14. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* 34, 1145–1160 (2016).

15. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853-1866.e17 (2016).

16. Schraivogel, D. *et al.* Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat Methods* 17, 629–635 (2020).

17. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol* 17, 72 (2016).

18. Chappell, L., Russell, A. J. C. & Voet, T. Single-Cell (Multi)omics Technologies. *Annu Rev Genom Hum G* 19, 1–27 (2016).

19. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 10, 1093–1095 (2013).

20. Grün, D., Kester, L. & Oudenaarden, A. van. Validation of noise models for single-cell transcriptomics. *Nat Methods* 11, 637–640 (2014).

21. Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 16, 716–726 (2015).

22. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat Rev Genet* 1–23 (2022) doi:10.1038/s41576-022-00509-1.

23. Shema, E., Bernstein, B. E. & Buenrostro, J. D. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat Genet* 51, 19–25 (2019).

24. Arneson, D., Yang, X. & Wang, K. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. *Commun Biology* 3, 422 (2020).

25. Rowland, B. *et al.* THUNDER: A reference-free deconvolution method to infer cell type proportions from bulk Hi-C data. *Plos Genet* 18, e1010102 (2022).

26. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol* 21, 221 (2020).

27. Li, H. *et al.* DeconPeaker, a Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of Mixture Samples. *Frontiers Genetics* 11, 392 (2020).

28. Zeng, W. *et al.* DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nat Commun* 10, 4613 (2019).

29. Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & Preter, K. D. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun* 11, 5650 (2020).

30. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol* 22, 102 (2021).

31. Salas, L. A. *et al.* An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol* 19, 64 (2018).

32. Salas, L. A. *et al.* Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun* 13, 761 (2022).

33. Koestler, D. C. *et al.* Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *Bmc Bioinformatics* 17, 120 (2016).

34. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *Bmc Bioinformatics* 13, 86 (2012).

35. Scott, C. A. *et al.* Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol* 21, 156 (2020).

36. Rahmani, E. *et al.* Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat Commun* 10, 3417 (2019).

37. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods* 15, 1059–1066 (2018).

38. Wu, C. -t. & Morris, J. R. Genes, Genetics, and Epigenetics: A Correspondence. *Science* 293, 1103–1105 (2001).

39. Holliday, R. Epigenetics: An overview. *Dev. Genet.* 15, 453–457 (1994).

40. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* 17, 487–500 (2016).

41. Bird, A. DNA methylation patterns and epigenetic memory. *Gene Dev* 16, 6–21 (2002).

42. Hemberger, M., Dean, W. & Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat Rev Mol Cell Bio* 10, 526–537 (2009).

43. Fisher, A. G. & Brockdorff, N. Epigenetic memory and parliamentary privilege combine to evoke discussions on inheritance. *Development* 139, 3891–3896 (2012).

44. Kim, M. & Costello, J. DNA methylation: an epigenetic mark of cellular memory. *Exp Mol Medicine* 49, e322–e322 (2017).

45. Koh, K. P. & Rao, A. DNA methylation and methylcytosine oxidation in cell fate decisions. *Curr Opin Cell Biol* 25, 152–161 (2013).

46. Bogdanović, O. & Lister, R. DNA methylation and the preservation of cell identity. *Curr Opin Genet Dev* 46, 9–14 (2017).

47. Petryk, N., Bultmann, S., Bartke, T. & Defossez, P.-A. Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. *Nucleic Acids Res* 49, gkaa1154 (2020).

48. Mattei, A. L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective. *Trends Genet* 38, 676–707 (2022).

49. Smith, J. D. & Markham, R. Polynucleotides from Deoxyribonucleic Acids. *Nature* 170, 120–121 (1952).

50. Sinsheimer, R. L. The action of pancreatic desoxyribonuclease: I. Isolation of mono- and dinucleotides. *J. Biol. Chem.* 208, 445–459 (1953).

51. Trunbull, J. F. & Adams, R. L. P. DNA methylase: purification from ascites cells and the effect of various DNA substrates on its activity. *Nucleic Acids Res* 3, 677–695 (1975).

52. Jones, P. A. & Taylor, S. M. Hemimethylated duplex DNAs prepared from 5-azacytidine-treated cells. *Nucleic Acids Res* 9, 2933–2947 (1981).

53. Bestor, T., Laudano, A., Mattaliano, R. & Ingram, V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* 203, 971–983 (1988).

54. Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A. & Cedar, H. Clonal inheritance of the pattern of DNA methylation in mouse cells. *Proc National Acad Sci* 79, 61–65 (1982).

55. Yoder, J. A., Soman, N. S., Verdine, G. L. & Bestor, T. H. DNA (cytosine-5)-methyltransferases in mouse cells and tissues. studies with a mechanism-based probe 1 1 Edited by K. Yamamoto. *J Mol Biol* 270, 385–395 (1997).

56. Goll, M. G. & Bestor, T. H. EUKARYOTIC CYTOSINE METHYLTRANSFERASES. *Biochemistry-us* 74, 481–514 (2005).

57. Xie, S. *et al.* Cloning, expression and chromosome locations of the human DNMT3 gene family. *Gene* 236, 87–95 (1999).

58. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc National Acad Sci* 97, 5237–5242 (2000).

59. Edwards, J. R., Yarychkivska, O., Boulard, M. & Bestor, T. H. DNA methylation and DNA methyltransferases. *Epigenet Chromatin* 10, 23 (2017).

60. Arand, J. *et al.* In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases. *Plos Genet* 8, e1002750 (2012).

61. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* 19, 219–220 (1998).

62. Gowher, H. & Jeltsch, A. Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpA sites1 1Edited by J. Karn. *J Mol Biol* 309, 1201–1208 (2001).

63. Jeong, S. *et al.* Selective Anchoring of DNA Methyltransferases 3A and 3B to Nucleosomes Containing Methylated DNA. *Mol Cell Biol* 29, 5366–5376 (2009).

64. Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. *Mol Cell Biol* 23, 5594–5605 (2003).

65. Jones, P. A. & Liang, G. Rethinking how DNA methylation patterns are maintained. *Nat Rev Genet* 10, 805–811 (2009).

66. Walton, E. L., Francastel, C. & Velasco, G. Maintenance of DNA methylation: Dnmt3b joins the dance. *Epigenetics* 6, 1373–1377 (2011).

67. Farthing, C. R. *et al.* Global Mapping of DNA Methylation in Mouse Promoters Reveals Epigenetic Reprogramming of Pluripotency Genes. *Plos Genet* 4, e1000116 (2008).

68. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 39, 457–466 (2007).

69. Mohn, F. *et al.* Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. *Mol Cell* 30, 755–766 (2008).

70. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766–770 (2008).

71. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* 454, 49–55 (2008).

72. Bernstein, B. E., Meissner, A. & Lander, E. S. The Mammalian Epigenome. *Cell* 128, 669–681 (2007).

73. Mohn, F. & Schübeler, D. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet* 25, 129–136 (2009).

74. Gifford, C. A. *et al.* Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. *Cell* 153, 1149–1163 (2013).

75. Farlik, M. *et al.* DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell* 19, 808–822 (2016).

76. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471, 68–73 (2011).

77. Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* 467, 338–342 (2010).

78. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* 27, 361–368 (2009).

79. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466, 253–257 (2010).

80. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* 513, 115–119 (2014).

81. Consortium, D. *et al.* A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biol* 19, 150 (2018).

82. Illingworth, R. *et al.* A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci. *Plos Biol* 6, e22 (2008).

83. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38, 1378–1385 (2006).

84. Christensen, B. C. *et al.* Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *Plos Genet* 5, e1000602 (2009).

85. Ghosh, S. *et al.* Tissue specific DNA methylation of CpG islands in normal human adult somatic tissues distinguishes neural from non-neural tissues. *Epigenetics* 5, 527–538 (2010).

86. Wan, J. *et al.* Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *Bmc Genomics* 16, 49 (2015).

87. Yagi, S. *et al.* DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. *Genome Res* 18, 1969–1978 (2008).

88. Avraham, A. *et al.* Tissue Specific DNA Methylation in Normal Human Breast Epithelium and in Breast Cancer. *Plos One* 9, e91805 (2014).

89. Zhou, J. *et al.* Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation. *Bmc Genomics* 18, 724 (2017).

90. Luo, Y., Lu, X. & Xie, H. Dynamic Alu Methylation during Normal Development, Aging, and Tumorigenesis. *Biomed Res Int* 2014, 784706 (2014).

91. Babenko, V. N., Chadaeva, I. V. & Orlov, Y. L. Genomic landscape of CpG rich elements in human. *Bmc Evol Biol* 17, 19 (2017).

92. Bestor, T. H. DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philosophical Transactions Royal Soc Lond B Biological Sci* 326, 179–187 (1990).

93. Zhou, W., Liang, G., Molloy, P. L. & Jones, P. A. DNA methylation enables transposable element-driven genome expansion. *Proc National Acad Sci* 117, 19359–19366 (2020).

94. Jansz, N. DNA methylation dynamics at transposable elements in mammals. *Essays Biochem* 63, 677–689 (2019).

95. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 8, 272–285 (2007).

96. Consortium, I. H. G. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).

97. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol* 19, 199 (2018).

98. Konkel, M. K. & Batzer, M. A. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* 20, 211–221 (2010).

99. Ayarpadikannan, S. & Kim, H.-S. The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. *Genom Informatics* 12, 98–104 (2014).

100. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mobile Dna-uk* 7, 9 (2016).

101. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* 20, 116–117 (1998).

102. Chernyavskaya, Y. *et al.* Loss of DNA methylation in zebrafish embryos activates retrotransposons to trigger antiviral signaling. *Development* 144, 2925–2939 (2017).

103. Zhou, Y., Cambareri, E. & Kinsey, J. DNA methylation inhibits expression and transposition of the Neurospora Tad retrotransposon. *Mol Genet Genomics* 265, 748–754 (2001).

104. Deniz, Ö., Frost, J. M. & Branco, M. R. Regulation of transposable elements by DNA modifications. *Nat Rev Genet* 20, 417–431 (2019).

105. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J Mol Biol* 196, 261–282 (1987).

106. Ioshikhes, I. P. & Zhang, M. Q. Large-scale human promoter mapping using CpG islands. *Nat Genet* 26, 61–63 (2000).

107. Edwards, J. R. *et al.* Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res* 20, 972–980 (2010).

108. Erfurth, F. E. *et al.* MLL protects CpG clusters from methylation within the Hoxa9 gene, maintaining transcript expression. *Proc National Acad Sci* 105, 7517–7522 (2008).

109. Clouaire, T. *et al.* Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Gene Dev* 26, 1714–1728 (2012).

110. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Bio* 16, 519–532 (2015).

111. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 14, 204–220 (2013).

112. Cano-Rodriguez, D. *et al.* Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner. *Nat Commun* 7, 12284 (2016).

113. Taberlay, P. C. *et al.* Polycomb-Repressed Genes Have Permissive Enhancers that Initiate Reprogramming. *Cell* 147, 1283–1294 (2011).

114. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 13, 484–492 (2012).

115. Brinkman, A. B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res* 22, 1128–1138 (2012).

116. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452, 215–219 (2008).

117. Zhang, X. *et al.* Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell* 126, 1189–1201 (2006).

118. Chotalia, M. *et al.* Transcription is required for establishment of germline methylation marks at imprinted genes. *Gene Dev* 23, 105–117 (2009).

119. Tufarelli, C. *et al.* Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat Genet* 34, 157–165 (2003).

120. Amante, S. M. *et al.* Transcription of intragenic CpG islands influences spatiotemporal host gene pre-mRNA processing. *Nucleic Acids Res* 48, gkaa556- (2020).

121. Williamson, C. M. *et al.* Uncoupling Antisense-Mediated Silencing and DNA Methylation in the Imprinted Gnas Cluster. *Plos Genet* 7, e1001347 (2011).

122. Dahlet, T. *et al.* Genome-wide analysis in the mouse embryo reveals the importance of DNA methylation for transcription integrity. *Nat Commun* 11, 3153 (2020).

123. Jeziorska, D. M. *et al.* DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc National Acad Sci* 114, E7526–E7535 (2017).

124. Baubec, T. *et al.* Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* 520, 243–247 (2015).

125. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* 543, 72–77 (2017).

126. Cain, J. A., Montibus, B. & Oakey, R. J. Intragenic CpG Islands and Their Impact on Gene Regulation. *Frontiers Cell Dev Biology* 10, 832348 (2022).

127. Nan, X., Meehan, R. R. & Bird, A. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res* 21, 4886–4892 (1993).

128. Wade, P. A. *et al.* Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat Genet* 23, 62–66 (1999).

129. Boeke, J., Ammerpohl, O., Kegel, S., Moehren, U. & Renkawitz, R. The Minimal Repression Domain of MBD2b Overlaps with the Methyl-CpG-binding Domain and Binds Directly to Sin3A*. *J Biol Chem* 275, 34963–34967 (2000).

130. Ohki, I. *et al.* Solution Structure of the Methyl-CpG Binding Domain of Human MBD1 in Complex with Methylated DNA. *Cell* 105, 487–497 (2001).

131. Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet* 19, 269–277 (2003).

132. Ginder, G. D. & Williams, D. C. Readers of DNA methylation, the MBD family as potential therapeutic targets. *Pharmacol Therapeut* 184, 98–111 (2018).

133. Boyes, J. & Bird, A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* 64, 1123–1134 (1991).

134. Meehan, R., Lewis, J. D. & Bird, A. P. Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. *Nucleic Acids Res* 20, 5085–5092 (1992).

135. N, H. K. *et al.* Brahma links the SWI/SNF chromatin-remodeling complex with MeCP2-dependent transcriptional silencing. *Nat Genet* 37, 254–264 (2005).

136. Kokura, K. *et al.* The Ski Protein Family Is Required for MeCP2-mediated Transcriptional Repression*. *J Biol Chem* 276, 34115–34121 (2001).

137. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* 393, 386–389 (1998).

138. Fuks, F. *et al.* The Methyl-CpG-binding Protein MeCP2 Links DNA Methylation to Histone Methylation*. *J Biol Chem* 278, 4035–4040 (2003).

139. Du, Q., Luu, P.-L., Stirzaker, C. & Clark, S. J. Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics-uk* 7, 1051–1073 (2015).

140. WYATT, G. R. & COHEN, S. S. A New Pyrimidine Base from Bacteriophage Nucleic Acids. *Nature* 170, 1072–1073 (1952).

141. Wyatt, G. R. & Cohen, S. S. The bases of the nucleic acids of some bacterial and animal viruses: the occurrence of 5-hydroxymethylcytosine. *Biochem J* 55, 774–782 (1953).

142. Penn, N. W., Suwalski, R., O'Riley, C., Bojanowski, K. & Yura, R. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochem J* 126, 781–790 (1972).

143. Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* 324, 930–935 (2009).

144. Kriaucionis, S. & Heintz, N. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science* 324, 929–930 (2009).

145. He, S. *et al.* Passive DNA demethylation preferentially up-regulates pluripotency-related genes and facilitates the generation of induced pluripotent stem cells. *J Biol Chem* 292, 18542–18555 (2017).

146. Maiti, A. & Drohat, A. C. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES*. *J Biol Chem* 286, 35334–35338 (2011).

147. Prakash, A., Carroll, B. L., Sweasy, J. B., Wallace, S. S. & Doublié, S. Genome and cancer single nucleotide polymorphisms of the human NEIL1 DNA glycosylase: Activity, structure, and the effect of editing. *Dna Repair* 14, 17–26 (2014).

148. Schomacher, L. *et al.* Neil DNA glycosylases promote substrate turnover by Tdg during DNA demethylation. *Nat Struct Mol Biol* 23, 116–124 (2016).

149. Lister, R. *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* 341, 1237905 (2013).

150. Münzel, M. *et al.* Quantification of the Sixth DNA Base Hydroxymethylcytosine in the Brain. *Angew. Chem. Int. Ed.* 49, 5375–5377 (2010).

151. Hahn, M. A., Szabó, P. E. & Pfeifer, G. P. 5-Hydroxymethylcytosine: A stable or transient DNA modification? *Genomics* 104, 314–323 (2014).

152. Münzel, M., Globisch, D. & Carell, T. 5-Hydroxymethylcytosine, the Sixth Base of the Genome. *Angew. Chem. Int. Ed.* 50, 6460–6468 (2011).

153. Diep, D. & Zhang, K. Genome-wide mapping of the sixth base. *Genome Biol* 12, 116 (2011).

154. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* 6, 1049–1055 (2014).

155. Globisch, D. *et al.* Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. *Plos One* 5, e15367 (2010).

156. Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* 151, 1417–30 (2012).

157. Song, C.-X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* 30, 1107–1116 (2012).

158. Nestor, C. E. *et al.* Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res* 22, 467–477 (2012).

159. He, B. *et al.* Tissue-specific 5-hydroxymethylcytosine landscape of the human genome. *Nat Commun* 12, 4249 (2021).

160. Jin, S.-G., Wu, X., Li, A. X. & Pfeifer, G. P. Genomic mapping of 5-hydroxymethylcytosine in the human brain. *Nucleic Acids Res* 39, 5015–5024 (2011).

161. Szulwach, K. E. *et al.* 5-hmC–mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat Neurosci* 14, 1607–1616 (2011).

162. Yildirim, O. *et al.* Mbd3/NURD Complex Regulates Expression of 5-Hydroxymethylcytosine Marked Genes in Embryonic Stem Cells. *Cell* 147, 1498–1510 (2011).

163. Wang, T. *et al.* Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. *Hum Mol Genet* 21, 5500–5510 (2012).

164. Wen, L. *et al.* Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol* 15, R49 (2014).

165. Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* 473, 394–397 (2011).

166. Yu, M. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* 149, 1368–1380 (2012).

167. Hahn, M. A. *et al.* Dynamics of 5-Hydroxymethylcytosine and Chromatin Marks in Mammalian Neurogenesis. *Cell Reports* 3, 291–300 (2013).

168. Stroud, H., Feng, S., Kinney, S. M., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol* 12, R54 (2011).

169. Xu, Y. *et al.* Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Mol Cell* 42, 451–464 (2011).

170. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* 473, 389–393 (2011).

171. Wu, H. *et al.* Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Gene Dev* 25, 679–684 (2011).

172. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* 473, 343–348 (2011).

173. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 473, 398–402 (2011).

174. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* 466, 1129–1133 (2010).

175. Shi, D.-Q., Ali, I., Tang, J. & Yang, W.-C. New Insights into 5hmC DNA Modification: Generation, Distribution and Function. *Frontiers Genetics* 8, 100 (2017).

176. Gu, T.-P. *et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* 477, 606–610 (2011).

177. Huang, Y. *et al.* Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proc National Acad Sci* 111, 1361–1366 (2014).

178. Jin, S.-G. *et al.* 5-Hydroxymethylcytosine Is Strongly Depleted in Human Cancers but Its Levels Do Not Correlate with IDH1 Mutations. *Cancer Res* 71, 7360–7365 (2011).

179. Liu, Y. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat Biotechnol* 37, 424–429 (2019).

180. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc National Acad Sci* 89, 1827–1831 (1992).

181. Huang, Y. *et al.* The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. *Plos One* 5, e8888 (2010).

182. Hayatsu, H. & Shiragami, M. Reaction of bisulfite with the 5-hydroxymethyl group in pyrimidines and in phage DNAs. *Biochemistry-us* 18, 632–637 (1979).

183. Booth, M. J. *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* 8, 1841–1851 (2013).

184. Houseman, E. A., Johnson, K. C. & Christensen, B. C. OxyBS: estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. *Bioinformatics* 32, 2505–2507 (2016).

185. Xu, Z., Taylor, J. A., Leung, Y.-K., Ho, S.-M. & Niu, L. oxBS-MLE: an efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. *Bioinformatics* 32, 3667–3669 (2016).

186. Heiss, J. A. *et al.* Battle of epigenetic proportions: comparing Illumina's EPIC methylation microarrays and TruSeq targeted bisulfite sequencing. *Epigenetics* 15, 174–182 (2020).

187. Shu, C., Zhang, X., Aouizerat, B. E. & Xu, K. Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells. *Epigenet Chromatin* 13, 51 (2020).

188. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* 12, 31–46 (2022).

189. Hitchins, M. P. *et al.* Dominantly Inherited Constitutional Epigenetic Silencing of MLH1 in a Cancer-Affected Family Is Linked to a Single Nucleotide Variant within the 5′UTR. *Cancer Cell* 20, 200–213 (2011).

190. Muse, M. E. *et al.* Enrichment of CpG island shore region hypermethylation in epigenetic breast field cancerization. *Epigenetics* 15, 1093–1106 (2020).

191. You, J. S. & Jones, P. A. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell* 22, 9–20 (2012).

192. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339 (2013).

193. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* 578, 82–93 (2020).

194. Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Csh Perspect Biol* 8, a019505 (2016).

195. Yamazaki, J. *et al.* Effects of TET2 mutations on DNA methylation in chronic myelomonocytic leukemia. *Epigenetics* 7, 201–207 (2012).

196. Yamazaki, J. *et al.* TET2 Mutations Affect Non-CpG Island DNA Methylation at Enhancers and Transcription Factor–Binding Sites in Chronic Myelomonocytic Leukemia. *Cancer Res* 75, 2833–2843 (2015).

197. Tulstrup, M. *et al.* TET2 mutations are associated with hypermethylation at key regulatory enhancers in normal and malignant hematopoiesis. *Nat Commun* 12, 6061 (2021).

198. López-Moyado, I. F. *et al.* Paradoxical association of TET loss of function with genome-wide DNA hypomethylation. *Proc National Acad Sci* 116, 16933–16942 (2019).

199. Asmar, F. *et al.* Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma. *Haematologica* 98, 1912–1920 (2013).

200. Rasmussen, K. D. & Helin, K. Role of TET enzymes in DNA methylation, development, and cancer. *Gene Dev* 30, 733–750 (2016).

201. Figueroa, M. E. *et al.* Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell* 18, 553–567 (2010).

202. Bledea, R. *et al.* Functional and topographic effects on DNA methylation in IDH1/2 mutant cancers. *Sci Rep-uk* 9, 16830 (2019).

203. Unruh, D. *et al.* Methylation and transcription patterns are distinct in IDH mutant gliomas compared to other IDH mutant cancers. *Sci Rep-uk* 9, 8946 (2019).

204. Glowacka, W. K. *et al.* 5-Hydroxymethylcytosine preferentially targets genes upregulated in isocitrate dehydrogenase 1 mutant high-grade glioma. *Acta Neuropathol* 135, 617–634 (2018).

205. Ehrlich, M. & Lacey, M. Epigenetic Alterations in Oncogenesis. *Adv Exp Med Biol* 754, 31–56 (2012).

206. Chen, R. Z., Pettersson, U., Beard, C., Jackson-Grusby, L. & Jaenisch, R. DNA hypomethylation leads to elevated mutation rates. *Nature* 395, 89–93 (1998).

207. Gaudet, F. *et al.* Induction of Tumors in Mice by Genomic Hypomethylation. *Science* 300, 489–492 (2003).

208. Narayan, A. *et al.* Hypomethylation of pericentromeric DNA in breast adenocarcinomas. *Int. J. Cancer* 77, 833–838 (1998).

209. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics-uk* 1, 239–259 (2009).

210. Sheaffer, K. L., Elliott, E. N. & Kaestner, K. H. DNA Hypomethylation Contributes to Genomic Instability and Intestinal Cancer Initiation. *Cancer Prev Res* 9, 534–546 (2016).

211. Zeggar, H. R. *et al.* Tumor DNA hypomethylation of LINE-1 is associated with low tumor grade of breast cancer in Tunisian patients. *Oncol Lett* 20, 1999–2006 (2020).

212. Cunningham, J. M. *et al.* Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res* 58, 3455–60 (1998).

213. Myöhänen, S. K., Baylin, S. B. & Herman, J. G. Hypermethylation can selectively silence individual p16ink4A alleles in neoplasia. *Cancer Res* 58, 591–3 (1998).

214. Ohtani-Fujita, N. *et al.* CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene. *Oncogene* 8, 1063–7 (1993).

215. Greger, V., Passarge, E., Höpping, W., Messmer, E. & Horsthemke, B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum Genet* 83, 155–158 (1989).

216. Esteller, M. *et al.* Promoter Hypermethylation and BRCA1 Inactivation in Sporadic Breast and Ovarian Tumors. *Jnci J National Cancer Inst* 92, 564–569 (2000).

217. Herman, J. G. *et al.* Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc National Acad Sci* 95, 6870–6875 (1998).

218. Zheng, Y. *et al.* A pan-cancer analysis of CpG Island gene regulation reveals extensive plasticity within Polycomb target genes. *Nat Commun* 12, 2485 (2021).

219. Sproul, D. & Meehan, R. R. Genomic insights into cancer-associated aberrant CpG island hypermethylation. *Brief Funct Genomics* 12, 174–190 (2013).

220. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. *Nat Genet* 44, 40–46 (2012).

221. Moarii, M., Boeva, V., Vert, J.-P. & Reyal, F. Changes in correlation between promoter methylation and gene expression in cancer. *Bmc Genomics* 16, 873 (2015).

222. Ng, J. M.-K. & Yu, J. Promoter Hypermethylation of Tumour Suppressor Genes as Potential Biomarkers in Colorectal Cancer. *Int J Mol Sci* 16, 2472–2496 (2015).

223. Liyanage, C. *et al.* Promoter Hypermethylation of Tumor-Suppressor Genes p16INK4a, RASSF1A, TIMP3, and PCQAP/MED15 in Salivary DNA as a Quadruple Biomarker Panel for Early Detection of Oral and Oropharyngeal Cancers. *Biomol* 9, 148 (2019).

224. Pfeifer, G. P. p53 mutational spectra and the role of methylated CpG sequences. *Mutat Res Fundam Mol Mech Mutagen* 450, 155–166 (2000).

225. You, Y.-H., Li, C. & Pfeifer, G. P. Involvement of 5-methylcytosine in sunlight-induced mutagenesis. *J Mol Biol* 293, 493–503 (1999).

226. Rideout, W. M., Coetzee, G. A., Olumi, A. F. & Jones, P. A. 5-Methylcytosine as an Endogenous Mutagen in the Human LDL Receptor and p53 Genes. *Science* 249, 1288–1290 (1990).

227. Lehman, T. A., Greenblatt, M., Bennett, W. P. & Harris, C. C. Mutational Spectrum of the P53 Tumor Suppressor Gene: Clues to Cancer Etiology and Molecular Pathogenesis. *Drug Metab Rev* 26, 221–235 (1994).

228. Gao, F. *et al.* Integrated analyses of DNA methylation and hydroxymethylation reveal tumor suppressive roles of ECM1, ATF5, and EOMESin human hepatocellular carcinoma. *Genome Biol* 15, 533 (2014).

229. Gambichler, T., Sand, M. & Skrygan, M. Loss of 5-hydroxymethylcytosine and ten-eleven translocation 2 protein expression in malignant melanoma. *Melanoma Res* 23, 218–220 (2013).

230. Park, J.-L. *et al.* Decrease of 5hmC in gastric cancers is associated with TET1 silencing due to with DNA methylation and bivalent histone marks at TET1 CpG island 3′-shore. *Oncotarget* 6, 37647–37662 (2015).

231. Lian, C. G. *et al.* Loss of 5-Hydroxymethylcytosine Is an Epigenetic Hallmark of Melanoma. *Cell* 150, 1135–1146 (2012).

232. Larson, A. R. *et al.* Loss of 5-hydroxymethylcytosine correlates with increasing morphologic dysplasia in melanocytic tumors. *Modern Pathol* 27, 936–944 (2014).

233. Chen, K. *et al.* Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Res* 26, 103–118 (2016).

234. Qi, J. *et al.* Regional gain and global loss of 5-hydroxymethylcytosine coexist in genitourinary cancers and regulate different oncogenic pathways. *Clin Epigenetics* 14, 117 (2022).

235. Tanager, K. S. *et al.* Loss of 5-hydroxymethylcytosine expression is nearuniversal in B-cell lymphomas with variable mutations in epigenetic regulators. *Haematologica* 107, 966–969 (2021).

236. Munari, E. *et al.* Global 5-Hydroxymethylcytosine Levels Are Profoundly Reduced in Multiple Genitourinary Malignancies. *Plos One* 11, e0146302 (2016).

237. Shi, X. *et al.* Loss of 5-Hydroxymethylcytosine Is an Independent Unfavorable Prognostic Factor for Esophageal Squamous Cell Carcinoma. *Plos One* 11, e0153100 (2016).

238. LIAO, Y. *et al.* Low level of 5-Hydroxymethylcytosine predicts poor prognosis in non-small cell lung cancer. *Oncol Lett* 11, 3753–3760 (2016).

239. Orr, B. A., Haffner, M. C., Nelson, W. G., Yegnasubramanian, S. & Eberhart, C. G. Decreased 5-Hydroxymethylcytosine Is Associated with Neural Progenitor Phenotype in Normal Brain and Shorter Survival in Malignant Glioma. *Plos One* 7, e41036 (2012).

240. Misawa, K. *et al.* 5-Hydroxymethylcytosine and ten-eleven translocation dioxygenases in head and neck carcinoma. *J Cancer* 10, 5306–5314 (2019).

241. Liu, C. *et al.* Decrease of 5-Hydroxymethylcytosine Is Associated with Progression of Hepatocellular Carcinoma through Downregulation of TET1. *Plos One* 8, e62828 (2013).

242. Zhao, F. *et al.* Loss of 5-Hydroxymethylcytosine as an Epigenetic Signature That Correlates With Poor Outcomes in Patients With Medulloblastoma. *Frontiers Oncol* 11, 603686 (2021).

243. Azizgolshani, N. *et al.* DNA 5-hydroxymethylcytosine in pediatric central nervous system tumors may impact tumor classification and is a positive prognostic marker. *Clin Epigenetics* 13, 176 (2021).

244. Lemonnier, F. *et al.* Loss of 5-hydroxymethylcytosine is a frequent event in peripheral T-cell lymphomas. *Haematologica* 103, e115–e118 (2018).

245. Putiri, E. L. *et al.* Distinct and overlapping control of 5-methylcytosine and 5-hydroxymethylcytosine by the TET proteins in human cancer cells. *Genome Biol* 15, R81 (2014).

246. Pronier, E. *et al.* Inhibition of TET2-mediated conversion of 5-methylcytosine to 5-hydroxymethylcytosine disturbs erythroid and granulomonocytic differentiation of human hematopoietic progenitors. *Blood* 118, 2551–2555 (2011).

247. Ye, D., Ma, S., Xiong, Y. & Guan, K.-L. R-2-Hydroxyglutarate as the Key Effector of IDH Mutations Promoting Oncogenesis. *Cancer Cell* 23, 274–276 (2013).

248. Natsumeda, M. *et al.* Accumulation of 2-hydroxyglutarate in gliomas correlates with survival: a study by 3.0-tesla magnetic resonance spectroscopy. *Acta Neuropathologica Commun* 2, 158 (2014).

249. Han, S. *et al.* IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. *Brit J Cancer* 122, 1580–1589 (2020).

250. Losman, J.-A. *et al.* (R)-2-Hydroxyglutarate Is Sufficient to Promote Leukemogenesis and Its Effects Are Reversible. *Science* 339, 1621–1625 (2013).

251. Lemonnier, F. *et al.* The IDH2 R172K mutation associated with angioimmunoblastic T-cell lymphoma produces 2HG in T cells and impacts lymphoid development. *Proc National Acad Sci* 113, 15084–15089 (2016).

252. Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *Science* 194, 23–28 (1976).

253. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 46, 225–233 (2014).

254. Cooper, C. S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet* 47, 367–372 (2015).

255. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New Engl J Medicine* 366, 883–892 (2012).

256. Marusyk, A., Janiszewska, M. & Polyak, K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell* 37, 471–484 (2020).

257. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* 357, (2017).

258. Merlo, L. M. F. *et al.* A Comprehensive Survey of Clonal Diversity Measures in Barrett's Esophagus as Biomarkers of Progression to Esophageal Adenocarcinoma. *Cancer Prev Res* 3, 1388–1397 (2010).

259. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet* 38, 468–473 (2006).

260. Okosun, J. *et al.* Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat Genet* 46, 176–181 (2014).

261. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* 173, 879-893.e13 (2018).

262. Brady, S. W. *et al.* Combating subclonal evolution of resistant cancer phenotypes. *Nat Commun* 8, 1231 (2017).

263. Juric, D. *et al.* Convergent loss of PTEN leads to clinical resistance to a PI(3)Kα inhibitor. *Nature* 518, 240–244 (2015).

264. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 15, 81–94 (2018).

265. Hua, X. *et al.* Genetic and epigenetic intratumor heterogeneity impacts prognosis of lung adenocarcinoma. *Nat Commun* 11, 2459 (2020).

266. Morris, L. G. T. *et al.* Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget* 7, 10051–10063 (2016).

267. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New Engl J Medicine* 376, 2109–2121 (2017).

268. Stacker, S. A. *et al.* Lymphangiogenesis and lymphatic vessel remodelling in cancer. *Nat Rev Cancer* 14, 159–172 (2014).

269. Carmeliet, P. & Jain, R. K. Angiogenesis in cancer and other diseases. *Nature* 407, 249–257 (2000).

270. Korenchan, D. E. & Flavell, R. R. Spatiotemporal pH Heterogeneity as a Promoter of Cancer Progression and Therapeutic Resistance. *Cancers* 11, 1026 (2019).

271. Yuan, Y. Spatial Heterogeneity in the Tumor Microenvironment. *Csh Perspect Med* 6, a026583 (2016).

272. Hinohara, K. & Polyak, K. Intratumoral Heterogeneity: More Than Just Mutations. *Trends Cell Biol* 29, 569–579 (2019).

273. Kreso, A. *et al.* Variable Clonal Repopulation Dynamics Influence Chemotherapy Response in Colorectal Cancer. *Science* 339, 543–548 (2013).

274. Sharma, S. V. *et al.* A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell* 141, 69–80 (2010).

275. Johnson, B. E. *et al.* Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science* 343, 189–193 (2014).

276. Almendro, V. *et al.* Inference of Tumor Evolution during Chemotherapy by Computational Modeling and In Situ Analysis of Genetic and Phenotypic Cellular Diversity. *Cell Reports* 6, 514–527 (2014).

277. Hinohara, K. *et al.* KDM5 Histone Demethylase Activity Links Cellular Transcriptomic Heterogeneity to Therapeutic Resistance. *Cancer Cell* 34, 939-953.e9 (2018).

278. Hata, A. N. *et al.* Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat Med* 22, 262–269 (2016).

279. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat Med* 10, 789–799 (2004).

280. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* 144, 646–674 (2011).

281. Shen, H. & Laird, P. W. Interplay between the Cancer Genome and Epigenome. *Cell* 153, 38–55 (2013).

282. Vitale, I., Shema, E., Loi, S. & Galluzzi, L. Intratumoral heterogeneity in cancer progression and response to immunotherapy. *Nat Med* 27, 212–224 (2021).

283. Schmitt, M. W., Loeb, L. A. & Salk, J. J. The influence of subclonal resistance mutations on targeted cancer therapy. *Nat Rev Clin Oncol* 13, 335–347 (2016).

284. Raynaud, F., Mina, M., Tavernari, D. & Ciriello, G. Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. *Plos Genet* 14, e1007669 (2018).

285. Schramm, A. *et al.* Mutational dynamics between primary and relapse neuroblastomas. *Nat Genet* 47, 872–877 (2015).

286. Janiszewska, M. *et al.* In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. *Nat Genet* 47, 1212–1219 (2015).

287. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 22, 105–113 (2016).

288. Hu, Z., Li, Z., Ma, Z. & Curtis, C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nat Genet* 52, 701–708 (2020).

289. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* 6, 924–935 (2006).

290. Biswas, A. & De, S. Drivers of dynamic intratumor heterogeneity and phenotypic plasticity. *Am J Physiol-cell Ph* 320, C750–C760 (2021).

291. Mazor, T., Pankov, A., Song, J. S. & Costello, J. F. Intratumoral Heterogeneity of the Epigenome. *Cancer Cell* 29, 440–451 (2016).

292. Rastetter, M. *et al.* Frequent intra-tumoural heterogeneity of promoter hypermethylation in malignant melanoma. *Histol Histopathol* 22, 1005–15 (2007).

293. Varley, K. E., Mutch, D. G., Edmonston, T. B., Goodfellow, P. J. & Mitra, R. D. Intra-tumor heterogeneity of MLH1 promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Res* 37, 4603–4612 (2009).

294. Sigalotti, L. *et al.* Intratumor Heterogeneity of Cancer/Testis Antigens Expression in Human Cutaneous Melanoma Is Methylation-Regulated and Functionally Reverted by 5-Aza-2′-deoxycytidine. *Cancer Res* 64, 9167–9171 (2004).

295. Guo, M. *et al.* Accumulation of Promoter Methylation Suggests Epigenetic Progression in Squamous Cell Carcinoma of the Esophagus. *Clin Cancer Res* 12, 4515–4522 (2006).

296. Torres, C. M. *et al.* The linker histone H1.0 generates epigenetic and functional intratumor heterogeneity. *Science* 353, (2016).

297. Johnson, K. C. *et al.* Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response. *Nat Genet* 53, 1456–1468 (2021).

298. Almendro, V., Marusyk, A. & Polyak, K. Cellular Heterogeneity and Molecular Evolution in Cancer. *Annu Rev Pathology Mech Dis* 8, 277–302 (2013).

299. Flavahan, W. A. Epigenetic plasticity, selection, and tumorigenesis. *Biochem Soc T* 48, 1609–1621 (2020).

300. Oakes, C. C. *et al.* Evolution of DNA Methylation Is Linked to Genetic Aberrations in Chronic Lymphocytic Leukemia. *Cancer Discov* 4, 348–361 (2014).

301. Mazor, T. *et al.* DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. *Cancer Cell* 28, 307–317 (2015).

302. Lopes, M. B. & Vinga, S. Tracking intratumoral heterogeneity in glioblastoma via regularized classification of single-cell RNA-Seq data. *Bmc Bioinformatics* 21, 59 (2020).

303. Qazi, M. A., Bakhshinyan, D. & Singh, S. K. Deciphering brain tumor heterogeneity, one cell at a time. *Nat Med* 25, 1474–1476 (2019).

304. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401 (2014).

305. Gojo, J. *et al.* Single-Cell RNA-Seq Reveals Cellular Hierarchies and Impaired Developmental Trajectories in Pediatric Ependymoma. *Cancer Cell* 38, 44-59.e9 (2020).

306. Regner, M. J. *et al.* A multi-omic single-cell landscape of human gynecologic malignancies. *Mol Cell* 81, 4924-4941.e10 (2021).

307. Hu, J. *et al.* Single-Cell Transcriptome Analysis Reveals Intratumoral Heterogeneity in ccRCC, which Results in Different Clinical Outcomes. *Mol Ther* 28, 1658–1672 (2020).

308. Contreras-Trujillo, H. *et al.* Deciphering intratumoral heterogeneity using integrated clonal tracking and single-cell transcriptome analyses. *Nat Commun* 12, 6522 (2021).

309. Zhou, S. *et al.* Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks. *Mol Ther - Nucleic Acids* 23, 682–690 (2021).

310. Tokura, M. *et al.* Single-Cell Transcriptome Profiling Reveals Intratumoral Heterogeneity and Molecular Features of Ductal Carcinoma In Situ. *Cancer Res* 82, 3236–3248 (2022).

311. Liang, J., Chen, W., Ye, J., Ni, C. & Zhai, W. Single-cell transcriptomics analysis reveals intratumoral heterogeneity and identifies a gene signature associated with prognosis of hepatocellular carcinoma. *Bioscience Rep* 42, BSR20212560 (2022).

312. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat Commun* 12, 2540 (2021).

313. Zhou, Y. *et al.* Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nat Commun* 11, 6322 (2020).

314. Zhang, Y. & Weinberg, R. A. Epithelial-to-mesenchymal transition in cancer: complexity and opportunities. *Front Med-prc* 12, 361–373 (2018).

315. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* 166, 21–45 (2016).

316. Aggarwal, V., Montoya, C. A., Donnenberg, V. S. & Sant, S. Interplay between tumor microenvironment and partial EMT as the driver of tumor progression. *Iscience* 24, 102113 (2021).

317. Junk, D. J., Cipriano, R., Bryson, B. L., Gilmore, H. L. & Jackson, M. W. Tumor Microenvironmental Signaling Elicits Epithelial-Mesenchymal Plasticity through Cooperation with Transforming Genetic Events. *Neoplasia* 15, 1100–1109 (2013).

318. Solinas, G., Marchesi, F., Garlanda, C., Mantovani, A. & Allavena, P. Inflammation-mediated promotion of invasion and metastasis. *Cancer Metast Rev* 29, 243–248 (2010).

319. Brabletz, T. *et al.* Variable β-catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment. *Proc National Acad Sci* 98, 10356–10361 (2001).

320. Miettinen, P. J., Ebner, R., Lopez, A. R. & Derynck, R. TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J Cell Biology* 127, 2021–2036 (1994).

321. Jing, Y., Han, Z., Zhang, S., Liu, Y. & Wei, L. Epithelial-Mesenchymal Transition in tumor microenvironment. *Cell Biosci* 1, 29 (2011).

322. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nat Rev Mol Cell Bio* 20, 69–84 (2019).

323. Sánchez-Tilló, E. *et al.* ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene* 29, 3490–3500 (2010).

324. Cano, A. *et al.* The transcription factor Snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. *Nat Cell Biol* 2, 76–83 (2000).

325. Batlle, E. *et al.* The transcription factor Snail is a repressor of E-cadherin gene expression in epithelial tumour cells. *Nat Cell Biol* 2, 84–89 (2000).

326. Spaderna, S. *et al.* The Transcriptional Repressor ZEB1 Promotes Metastasis and Loss of Cell Polarity in Cancer. *Cancer Res* 68, 537–544 (2008).

327. Miyoshi, A. *et al.* Snail and SIP1 increase cancer invasion by upregulating MMP family in hepatocellular carcinoma cells. *Brit J Cancer* 90, 1265–1273 (2004).

328. Huang, R. Y.-J., Guilford, P. & Thiery, J. P. Early events in cell adhesion and polarity during epithelial-mesenchymal transition. *J Cell Sci* 125, 4417–4422 (2012).

329. Huang, Y., Hong, W. & Wei, X. The molecular mechanisms and therapeutic strategies of EMT in tumor progression and metastasis. *J Hematol Oncol* 15, 129 (2022).

330. Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* 556, 463–468 (2018).

331. Tam, W. L. & Weinberg, R. A. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med* 19, 1438–1449 (2013).

332. Maruyama, R. *et al.* Epigenetic Regulation of Cell Type–Specific Expression Patterns in the Human Mammary Epithelium. *Plos Genet* 7, e1001369 (2011).

333. Chaffer, C. L. *et al.* Poised Chromatin at the ZEB1 Promoter Enables Breast Cancer Cell Plasticity and Enhances Tumorigenicity. *Cell* 154, 61–74 (2013).

334. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *Ca Cancer J Clin* 72, 7–33 (2022).

335. Ostrom, Q. T., Cioffi, G., Waite, K., Kruchko, C. & Barnholtz-Sloan, J. S. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014–2018. *Neuro-oncology* 23, iii1–iii105 (2021).

336. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-oncology* 23, 1231–1251 (2021).

337. Bhakta, N. *et al.* The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE). *Lancet* 390, 2569–2582 (2017).

338. Palmer, S. L. *et al.* Patterns of Intellectual Development Among Survivors of Pediatric Medulloblastoma: A Longitudinal Analysis. *J Clin Oncol* 19, 2302–2308 (2001).

339. Merchant, T. E. *et al.* Critical Combinations of Radiation Dose and Volume Predict Intelligence Quotient and Academic Achievement Scores After Craniospinal Irradiation in Children With Medulloblastoma. *Int J Radiat Oncol Biology Phys* 90, 554–561 (2014).

340. Brinkman, T. M. *et al.* Long-Term Neurocognitive Functioning and Social Attainment in Adult Survivors of Pediatric CNS Tumors: Results From the St Jude Lifetime Cohort Study. *J Clin Oncol* 34, 1358–1367 (2016).

341. Armstrong, G. T. *et al.* Survival and long-term health and cognitive outcomes after low-grade glioma. *Neuro-oncology* 13, 223–234 (2011).

342. Ris, M. D. *et al.* Intellectual and academic outcome following two chemotherapy regimens and radiotherapy for average-risk medulloblastoma: COG A9961. *Pediatr Blood Cancer* 60, 1350–1357 (2013).

343. Brière, M., Scott, J. G., McNall-Knapp, R. Y. & Adams, R. L. Cognitive outcome in pediatric brain tumor survivors: Delayed attention deficit at long-term follow-up. *Pediatr Blood Cancer* 50, 337–340 (2008).

344. Bowers, D. C. *et al.* Late-Occurring Stroke Among Long-Term Survivors of Childhood Leukemia and Brain Tumors: A Report From the Childhood Cancer Survivor Study. *J Clin Oncol* 24, 5277–5282 (2006).

345. Board, P. P. T. E. PDQ Late Effects of Treatment for Childhood Cancer. *National Cancer Institute*.

346. Smith, M. A., Altekruse, S. F., Adamson, P. C., Reaman, G. H. & Seibel, N. L. Declining childhood and adolescent cancer mortality. *Cancer* 120, 2497–2506 (2014).

347. Cho, Y.-J. *et al.* Integrative Genomic Analysis of Medulloblastoma Identifies a Molecular Subgroup That Drives Poor Clinical Outcome. *J Clin Oncol* 29, 1424–1430 (2010).

348. Northcott, P. A. *et al.* Medulloblastoma Comprises Four Distinct Molecular Variants. *J Clin Oncol* 29, 1408–1414 (2010).

349. Thompson, M. C. *et al.* Genomics Identifies Medulloblastoma Subgroups That Are Enriched for Specific Genetic Alterations. *J Clin Oncol* 24, 1924–1931 (2006).

350. Kool, M. *et al.* Integrated Genomics Identifies Five Medulloblastoma Subtypes with Distinct Genetic Profiles, Pathway Signatures and Clinicopathological Features. *Plos One* 3, e3088 (2008).

351. Taylor, M. D. *et al.* Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol* 123, 465–472 (2012).

352. Juraschka, K. & Taylor, M. D. Medulloblastoma in the age of molecular subgroups: a review: JNSPG 75th Anniversary Invited Review Article. *J Neurosurg Pediatrics* 24, 353–363 (2019).

353. Pajtler, K. W. *et al.* Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell* 27, 728–743 (2015).

354. Witt, H. *et al.* DNA methylation-based classification of ependymomas in adulthood: implications for diagnosis and treatment. *Neuro-oncology* 20, 1616–1624 (2018).

355. Patel, R. R., Ramkissoon, S. H., Ross, J. & Weintraub, L. Tumor mutational burden and driver mutations: Characterizing the genomic landscape of pediatric brain tumors. *Pediatr Blood Cancer* 67, e28338 (2020).

356. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* 9, 34 (2017).

357. Gröbner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* 555, 321–327 (2018).

358. Jolly, M. K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Frontiers Oncol* 5, 155 (2015).

359. Chaffer, C. L. & Weinberg, R. A. A Perspective on Cancer Cell Metastasis. *Science* 331, 1559–1564 (2011).

360. Dillekås, H., Rogers, M. S. & Straume, O. Are 90% of deaths from cancer caused by metastases? *Cancer Med-us* 8, 5574–5576 (2019).

361. Nieto, M. A. Epithelial Plasticity: A Common Theme in Embryonic and Cancer Cells. *Science* 342, 1234850 (2013).

362. Jolly, M. K. *et al.* Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol Therapeut* 194, 161–184 (2018).

363. Brown, M. S. *et al.* Phenotypic heterogeneity driven by plasticity of the intermediate EMT state governs disease progression and metastasis in breast cancer. *Sci Adv* 8, eabj8002 (2022).

364. Grosse-Wilde, A. *et al.* Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. *Plos One* 10, e0126522 (2015).

365. Bierie, B. *et al.* Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc National Acad Sci* 114, E2337–E2346 (2017).

366. Ognjenovic, N. B. *et al.* Limiting Self-Renewal of the Basal Compartment by PKA Activation Induces Differentiation and Alters the Evolution of Mammary Tumors. *Dev Cell* 55, 544-557.e6 (2020).

367. Goetz, H., Melendez-Alvarez, J. R., Chen, L. & Tian, X.-J. A plausible accelerating function of intermediate states in cancer metastasis. *Plos Comput Biol* 16, e1007682 (2020).

368. Hong, T. *et al.* An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. *Plos Comput Biol* 11, e1004569 (2015).

369. Huang, R. Y.-J. *et al.* An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). *Cell Death Dis* 4, e915–e915 (2013).

370. Janitz, K. & Janitz, M. Handbook of Epigenetics. *Sect Iii Epigenetic Technology* 173–181 (2011) doi:10.1016/b978-0-12-375709-8.00012-5.

371. Cortellino, S. *et al.* Thymine DNA Glycosylase Is Essential for Active DNA Demethylation by Linked Deamination-Base Excision Repair. *Cell* 146, 67–79 (2011).

372. Tost, J. DNA Methylation: An Introduction to the Biology and the Disease-Associated Changes of a Promising Biomarker. *Mol Biotechnol* 44, 71–81 (2010).

373. Cui, X.-L. *et al.* A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nat Commun* 11, 6161 (2020).

374. Pistore, C. *et al.* DNA methylation variations are required for epithelial-to-mesenchymal transition induced by cancer-associated fibroblasts in prostate cancer cells. *Oncogene* 36, 5551–5566 (2017).

375. Tahara, T. *et al.* DNA Methylation Status of Epithelial-Mesenchymal Transition (EMT) - Related Genes Is Associated with Severe Clinical Phenotypes in Ulcerative Colitis (UC). *Plos One* 9, e107947 (2014).

376. Carmona, F. J. *et al.* A Comprehensive DNA Methylation Profile of Epithelial-to-Mesenchymal Transition. *Cancer Res* 74, 5608–5619 (2014).

377. Dumont, N. *et al.* Sustained induction of epithelial to mesenchymal transition activates DNA methylation of genes silenced in basal-like breast cancers. *Proc National Acad Sci* 105, 14867–14872 (2008).

378. Rajić, J. *et al.* DNA methylation of miR-200 clusters promotes epithelial to mesenchymal transition in human conjunctival epithelial cells. *Exp Eye Res* 197, 108047 (2020).

379. Choi, S. K. *et al.* Epigenetic landscape change analysis during human EMT sheds light on a key EMT mediator TRIM29. *Oncotarget* 8, 98322–98335 (2017).

380. Kumar, S., Chinnusamy, V. & Mohapatra, T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Frontiers Genetics* 9, 640 (2018).

381. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat Rev Mol Cell Bio* 14, 341–356 (2013).

382. Ngo, T. T. M. *et al.* Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat Commun* 7, 10813 (2016).

383. Tekpli, X. *et al.* Changes of 5-hydroxymethylcytosine distribution during myeloid and lymphoid differentiation of CD34+ cells. *Epigenet Chromatin* 9, 21 (2016).

384. Costa, Y. *et al.* NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature* 495, 370–374 (2013).

385. Xiong, J. *et al.* Cooperative Action between SALL4A and TET Proteins in Stepwise Oxidation of 5-Methylcytosine. *Mol Cell* 64, 913–925 (2016).

386. Shenoy, N. *et al.* Ascorbic acid–induced TET activation mitigates adverse hydroxymethylcystosine loss in renal cell carcinoma. *J Clin Invest* 129, 1612–1625 (2019).

387. Yang, H. *et al.* Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* 32, 663–669 (2013).

388. Tsai, K.-W. *et al.* Reduction of global 5-hydroxymethylcytosine is a poor prognostic factor in breast cancer patients, especially for an ER/PR-negative subtype. *Breast Cancer Res Tr* 153, 219–234 (2015).

389. Skvortsova, K. *et al.* DNA Hypermethylation Encroachment at CpG Island Borders in Cancer Is Predisposed by H3K4 Monomethylation Patterns. *Cancer Cell* 35, 297-314.e8 (2019).

390. Arechederra, M. *et al.* Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nat Commun* 9, 3164 (2018).

391. Gonzalez-Zulueta, M. *et al.* Methylation of the 5’ CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. *Cancer Res* 55, 4531–5 (1995).

392. Herman, J. G. *et al.* Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc National Acad Sci* 91, 9700–9704 (1994).

393. Jin, C. *et al.* TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. *Nucleic Acids Res* 42, 6956–6971 (2014).

394. Wilkins, O. M. *et al.* Genome-wide characterization of cytosine-specific 5-hydroxymethylation in normal breast tissue. *Epigenetics* 15, 398–418 (2020).

395. Fleischer, T. *et al.* Genome-wide DNA methylation profiles in progression to in situand invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol* 15, 435 (2014).

396. Johnson, K. C. *et al.* DNA methylation in ductal carcinoma in situ related with future development of invasive breast cancer. *Clin Epigenetics* 7, 75 (2015).

397. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 20, 207–220 (2019).

398. Cieślik, M. *et al.* Epigenetic coordination of signaling pathways during the epithelial-mesenchymal transition. *Epigenet Chromatin* 6, 28 (2013).

399. Bakiri, L. *et al.* Fra-1/AP-1 induces EMT in mammary epithelial cells by modulating Zeb1/2 and TGFβ expression. *Cell Death Differ* 22, 336–350 (2015).

400. Hardy, K. *et al.* Identification of chromatin accessibility domains in human breast cancer stem cells. *Nucleus* 7, 50–67 (2016).

401. Arase, M. *et al.* Dynamics of chromatin accessibility during TGF-β-induced EMT of Ras-transformed mammary gland epithelial cells. *Sci Rep-uk* 7, 1166 (2017).

402. Denny, S. K. *et al.* Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. *Cell* 166, 328–342 (2016).

403. Xin, L. *et al.* SND1 acts upstream of SLUG to regulate the epithelial–mesenchymal transition (EMT) in SKOV3 cells. *Faseb J* 33, 3795–3806 (2019).

404. Frey, W. D. *et al.* BPTF Maintains Chromatin Accessibility and the Self-Renewal Capacity of Mammary Gland Stem Cells. *Stem Cell Rep* 9, 23–31 (2017).

405. Stowers, R. S. *et al.* Matrix stiffness induces a tumorigenic phenotype in mammary epithelium through changes in chromatin accessibility. *Nat Biomed Eng* 3, 1009–1019 (2019).

406. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369 (2014).

407. Hansen, K. *IlluminaHumanMethylationEPICanno.ilm10b4.hg19: Annotation for Illumina's EPIC methylation arrays*. (R package version 0.6.0, 2017).

408. Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res* 44, e20–e20 (2016).

409. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 45, e22–e22 (2017).

410. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43, e47–e47 (2015).

411. Storey, J., Bass, A., Dabney, A., Robinson, D. & Warnes, G. *qvalue: Q-value estimation for false discovery rate control.* (2019).

412. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28, 495–501 (2010).

413. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biology* 109, 21.29.1-21.29.9 (2015).

414. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J* 17, 10–12 (2011).

415. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–359 (2012).

416. Institute, B. Picard Toolkit. https://broadinstitute.github.io/picard/ (2019).

417. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137 (2008).

418. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140 (2010).

419. Maintainer, B. & B, T. *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s)*. (R package 3.4.6, 2019).

420. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31, 2382–2383 (2015).

421. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* 12, 477–479 (2015).

422. Schep, A. *motifmatchr: Fast Motif Matching in R*. (R package version 1.20.0, 2020).

423. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 48, D87–D92 (2019).

424. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 32, 1555–1556 (2016).

425. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).

426. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics* 12, 323 (2011).

427. Yang, J. *et al.* Twist, a Master Regulator of Morphogenesis, Plays an Essential Role in Tumor Metastasis. *Cell* 117, 927–939 (2004).

428. Clayton, N. S. & Ridley, A. J. Targeting Rho GTPase Signaling Networks in Cancer. *Frontiers Cell Dev Biology* 8, 222 (2020).

429. Chen, A. F. *et al.* GRHL2-Dependent Enhancer Switching Maintains a Pluripotent Stem Cell Transcriptional Subnetwork after Exit from Naive Pluripotency. *Cell Stem Cell* 23, 226-238.e4 (2018).

430. Chung, V. Y. *et al.* The role of GRHL2 and epigenetic remodeling in epithelial–mesenchymal plasticity in ovarian cancer cells. *Commun Biology* 2, 272 (2019).

431. Ungefroren, H., Witte, D. & Lehnert, H. The role of small GTPases of the Rho/Rac family in TGF-β-induced EMT and cell motility in cancer. *Dev. Dyn.* 247, 451–461 (2018).

432. Ellenbroek, S. I. J. & Collard, J. G. Rho GTPases: functions and association with cancer. *Clin Exp Metastas* 24, 657–672 (2007).

433. Gong, F. *et al.* Epigenetic silencing of TET2 and TET3 induces an EMT-like process in melanoma. *Oncotarget* 8, 315–328 (2016).

434. Friedman, D. L. *et al.* Subsequent Neoplasms in 5-Year Survivors of Childhood Cancer: The Childhood Cancer Survivor Study. *Jnci J National Cancer Inst* 102, 1083–1095 (2010).

435. Tsui, K. *et al.* Subsequent neoplasms in survivors of childhood central nervous system tumors: risk after modern multimodal therapy. *Neuro-oncology* 17, 448–456 (2015).

436. Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* 547, 311–317 (2017).

437. Hovestadt, V. *et al.* Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* 572, 74–79 (2019).

438. Northcott, P. A. *et al.* Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* 488, 49–56 (2012).

439. Lin, C. Y. *et al.* Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature* 530, 57–62 (2016).

440. Parker, M. *et al.* C11orf95–RELA fusions drive oncogenic NF-κB signalling in ependymoma. *Nature* 506, 451–455 (2014).

441. Doz, F. *et al.* Efficacy and safety of larotrectinib in TRK fusion-positive primary central nervous system tumors. *Neuro-oncology* 24, noab274- (2021).

442. FDA approves larotrectinib for solid tumors with NTRK gene fusions | FDA. https://www.fda.gov/drugs/fda-approves-larotrectinib-solid-tumors-ntrk-gene-fusions.

443. Filbin, M. G. *et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* 360, 331–335 (2018).

444. Gillen, A. E. *et al.* Single-Cell RNA Sequencing of Childhood Ependymoma Reveals Neoplastic Cell Subpopulations That Impact Molecular Classification and Etiology. *Cell Reports* 32, 108023 (2020).

445. Auwera, G. V. der & O'Connor, B. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra.* (O'Reilly Media, 2020).

446. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–1303 (2010).

447. Hovestadt, V. & Zapatka, M. *conumee: Enhanced copy-number variation analysis using Illumina DNA methylation arrays*. (R package version 1.9.0).

448. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* 16, 619–626 (2019).

449. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 33, 495–502 (2015).

450. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* 184, 3573-3587.e29 (2021).

451. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888-1902.e21 (2019).

452. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36, 411–420 (2018).

453. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol* 20, 273 (2019).

454. Weber, L. M. *et al.* Genetic demultiplexing of pooled single-cell RNA-sequencing samples in cancer facilitates effective experimental design. *Gigascience* 10, giab062- (2021).

455. Tommaso, P. D. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 35, 316–319 (2017).

456. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *Biorxiv* 2021.05.05.442755 (2021) doi:10.1101/2021.05.05.442755.

457. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008 (2021).

458. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* 37, 4569–4571 (2021).

459. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Arxiv* (2018) doi:10.48550/arxiv.1802.03426.

460. Frost, H. R. Variance-adjusted Mahalanobis (VAM): a fast and accurate method for cell-specific gene set scoring. *Nucleic Acids Res* 48, gkaa582- (2020).

461. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740 (2011).

462. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* 1, 417–425 (2015).

463. Fan, X. *et al.* Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Res* 28, 730–745 (2018).

464. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 555, 524–528 (2018).

465. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* 370, (2020).

466. Manno, G. L. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* 167, 566-580.e19 (2016).

467. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res* 37, D674–D679 (2009).

468. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* 539, 309–313 (2016).

469. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32, 381–386 (2014).

470. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 14, 979–982 (2017).

471. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502 (2019).

472. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* 14, 309–315 (2017).

473. Carlson, M. *org.Hs.eg.db: Genome wide annotation for Human*. (R package version 3.8.2.).

474. Pagès, H., Carlson, M., Falcon, S. & Li, N. *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. (R package version 1.60.0).

475. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16, 278 (2015).

476. Carter, M. *et al.* Genetic abnormalities detected in ependymomas by comparative genomic hybridisation. *Brit J Cancer* 86, 929–939 (2002).

477. Rajeshwari, M. *et al.* Evaluation of chromosome 1q gain in intracranial ependymomas. *J Neuro-oncol* 127, 271–278 (2016).

478. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc National Acad Sci* 102, 15545–15550 (2005).

479. Zhao, W., Li, Y. & Zhang, X. Stemness-related markers in cancer. *Cancer Transl Medicine* 3, 87 (2017).

480. Kuhn, A. *et al.* Cell population-specific expression analysis of human cerebellum. *Bmc Genomics* 13, 610 (2012).

481. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* 12, 298–306 (2012).

482. Egeblad, M., Nakasone, E. S. & Werb, Z. Tumors as Organs: Complex Tissues that Interface with the Entire Organism. *Dev Cell* 18, 884–901 (2010).

483. Kaur, N. *et al.* Wnt3a mediated activation of Wnt/β-catenin signaling promotes tumor progression in glioblastoma. *Mol Cell Neurosci* 54, 44–57 (2013).

484. Johnson, R. A. *et al.* Cross-species genomics matches driver mutations and cell compartments to model ependymoma. *Nature* 466, 632–636 (2010).

485. Taylor, M. D. *et al.* Radial glia cells are candidate stem cells of ependymoma. *Cancer Cell* 8, 323–335 (2005).

486. Reitman, Z. J. *et al.* Mitogenic and progenitor gene programmes in single pilocytic astrocytoma cells. *Nat Commun* 10, 3731 (2019).

487. Abou-Antoun, T. J., Hale, J. S., Lathia, J. D. & Dombrowski, S. M. Brain Cancer Stem Cells in Adults and Children: Cell Biology and Therapeutic Implications. *Neurotherapeutics* 14, 372–384 (2017).

488. Castelo-Branco, P. & Tabori, U. Promises and challenges of exhausting pediatric neural cancer stem cells. *Pediatr Res* 71, 523–528 (2012).

489. Kong, D.-S. Cancer Stem Cells in Brain Tumors and Their Lineage Hierarchy. *Int J Stem Cells* 5, 12–15 (2012).

490. Bao, S. *et al.* Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* 444, 756–760 (2006).

491. Hemmati, H. D. *et al.* Cancerous stem cells can arise from pediatric brain tumors. *Proc National Acad Sci* 100, 15178–15183 (2003).

492. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* 432, 396–401 (2004).

493. Galli, R. *et al.* Isolation and Characterization of Tumorigenic, Stem-like Neural Precursors from Human Glioblastoma. *Cancer Res* 64, 7011–7021 (2004).

494. Lou, C. H. *et al.* Posttranscriptional Control of the Stem Cell and Neurogenic Programs by the Nonsense-Mediated RNA Decay Pathway. *Cell Reports* 6, 748–764 (2014).

495. Jaffrey, S. R. & Wilkinson, M. F. Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. *Nat Rev Neurosci* 19, 715–728 (2018).

496. Jolly, L. A., Homan, C. C., Jacob, R., Barry, S. & Gecz, J. The UPF3B gene, implicated in intellectual disability, autism, ADHD and childhood onset schizophrenia regulates neural progenitor cell behaviour and neuronal outgrowth. *Hum Mol Genet* 22, 4673–4687 (2013).

497. Robinson, K. E. *et al.* A quantitative meta-analysis of neurocognitive sequelae in survivors of pediatric brain tumors. *Pediatr Blood Cancer* 55, 525–531 (2010).

498. Ellenberg, L. *et al.* Neurocognitive Status in Long-Term Survivors of Childhood CNS Malignancies: A Report From the Childhood Cancer Survivor Study. *Neuropsychology* 23, 705–717 (2009).

499. Pinto, M. D., Conklin, H. M., Li, C. & Merchant, T. E. Learning and Memory Following Conformal Radiation Therapy for Pediatric Craniopharyngioma and Low-Grade Glioma. *Int J Radiat Oncol Biology Phys* 84, e363–e369 (2012).

500. Jessa, S. *et al.* Stalled developmental programs at the root of pediatric brain tumors. *Nat Genet* 51, 1702–1713 (2019).

501. Zhang, L. *et al.* Single-Cell Transcriptomics in Medulloblastoma Reveals Tumor-Initiating Progenitors and Oncogenic Cascades during Tumorigenesis and Relapse. *Cancer Cell* 36, 302-318.e7 (2019).

502. Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* 572, 67–73 (2019).

503. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* 555, 469–474 (2018).

504. Gold, M., Hurwitz, J. & Anders, M. The enzymatic methylation of RNA and DNA, II. on the species specificity. *Proc National Acad Sci* 50, 164–169 (1963).

505. Billen, D. & Hewitt, R. Influence of Starvation for Methionine and Other Amino Acids on Subsequent Bacterial Deoxyribonucleic Acid Replication. *J Bacteriol* 92, 609–617 (1966).

506. Billen, D. Methylation of the bacterial chromosome: an event at the “replication point”? *J Mol Biol* 31, 477–486 (1968).

507. Lark, C. Studies on the in vivo methylation of DNA in Escherichia coli 15T−. *J Mol Biol* 31, 389–399 (1968).

508. Srinivasan, P. R. & Borek, E. Enzymatic Alteration. *Science* 145, 548–553 (1964).

509. Ambrosi, C., Manzo, M. & Baubec, T. Dynamics and Context-Dependent Roles of DNA Methylation. *J Mol Biol* 429, 1459–1475 (2017).

510. Rose, N. R. & Klose, R. J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochimica Et Biophysica Acta Bba - Gene Regul Mech* 1839, 1362–1372 (2014).

511. Zilberman, D., Coleman-Derr, D., Ballinger, T. & Henikoff, S. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature* 456, 125–129 (2008).

512. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 43, 768–775 (2011).

513. Ito, S. *et al.* Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* 333, 1300–1303 (2011).

514. He, Y.-F. *et al.* Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science* 333, 1303–1307 (2011).

515. Thomson, J. P. & Meehan, R. R. The application of genome-wide 5-hydroxymethylcytosine studies in cancer research. *Epigenomics-uk* 9, 77–91 (2017).

516. Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C. & Greenberg, M. E. Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proc National Acad Sci* 112, 6800–6806 (2015).

517. Thomson, J. P. *et al.* Comparative analysis of affinity-based 5-hydroxymethylation enrichment techniques. *Nucleic Acids Res* 41, e206–e206 (2013).

518. Spada, F. *et al.* Active turnover of genomic methylcytosine in pluripotent cells. *Nat Chem Biol* 16, 1411–1419 (2020).

519. Stoyanova, E., Riad, M., Rao, A. & Heintz, N. 5-Hydroxymethylcytosine-mediated active demethylation is required for mammalian neuronal differentiation and function. *Elife* 10, e66973 (2021).

520. Kudo, Y. *et al.* Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. *Cancer Sci* 103, 670–676 (2012).

521. Ficz, G. & Gribben, J. G. Loss of 5-hydroxymethylcytosine in cancer: Cause or consequence? *Genomics* 104, 352–357 (2014).

522. Johnson, K. C. *et al.* 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nat Commun* 7, 13177 (2016).

523. Lu, C. *et al.* IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature* 483, 474–478 (2012).

524. Rampal, R. *et al.* DNA Hydroxymethylation Profiling Reveals that WT1 Mutations Result in Loss of TET2 Function in Acute Myeloid Leukemia. *Cell Reports* 9, 1841–1855 (2014).

525. Duncan, C. G. *et al.* A heterozygous IDH1R132H/WT mutation induces genome-wide alterations in DNA methylation. *Genome Res* 22, 2339–2355 (2012).

526. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc National Acad Sci* 110, 4009–4014 (2013).

527. Hoffman, M. *et al.* Intratumoral genetic and functional heterogeneity in pediatric glioblastoma. *Cancer Res* 79, canres.3441.2018 (2019).

528. Kim, E. L. *et al.* Intratumoral Heterogeneity and Longitudinal Changes in Gene Expression Predict Differential Drug Sensitivity in Newly Diagnosed and Recurrent Glioblastoma. *Cancers* 12, 520 (2020).

529. Qazi, M. A. *et al.* Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. *Ann Oncol* 28, 1448–1456 (2017).

530. Gularyan, S. K. *et al.* Investigation of Inter- and Intratumoral Heterogeneity of Glioblastoma Using TOF-SIMS*. *Mol Cell Proteomics* 19, 960–970 (2020).

531. Larsson, I. *et al.* Modeling glioblastoma heterogeneity as a dynamic network of cell states. *Mol Syst Biol* 17, e10105 (2021).

532. Berens, M. E. *et al.* Multiscale, multimodal analysis of tumor heterogeneity in IDH1 mutant vs wild-type diffuse gliomas. *Plos One* 14, e0219724 (2019).

533. Lam, K. H. B., Valkanas, K., Djuric, U. & Diamandis, P. Unifying models of glioblastoma's intra-tumoral heterogeneity. *Neuro-oncology Adv* 2, vdaa096- (2020).

534. Sproul, D. *et al.* Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol* 13, R84 (2012).

535. Zhang, B. *et al.* Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res* 23, 1522–1540 (2013).

536. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* 9, 5068 (2018).

537. Kim, S. *et al.* Enlarged leukocyte referent libraries can explain additional variance in blood-based epigenome-wide association studies. *Epigenomics-uk* 8, 1185–1192 (2016).

538. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 15, R31 (2014).

539. You, C. *et al.* A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. *Nat Commun* 11, 4779 (2020).

540. Reinius, L. E. *et al.* Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *Plos One* 7, e41361 (2012).

541. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 13, 229–232 (2016).

542. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 11, 817–820 (2014).

543. Teschendorff, A. E. & Zheng, S. C. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics-uk* 9, 757–768 (2017).

544. Rahmani, E. *et al.* BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. *Genome Biol* 19, 141 (2018).

545. Rahmani, E. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat Methods* 13, 443–445 (2016).

546. Waite, L. L. *et al.* Estimation of Cell-Type Composition Including T and B Cell Subtypes for Whole Blood Methylation Microarray Data. *Frontiers Genetics* 7, 23 (2016).

547. Zhang, Z. *et al.* HiTIMED: hierarchical tumor immune microenvironment epigenetic deconvolution for accurate cell type resolution in the tumor microenvironment using tumor-type-specific DNA methylation data. *J Transl Med* 20, 516 (2022).

548. Andrews, S. FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

549. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

550. Zhou, W., Triche, T. J., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res* 46, gky691- (2018).

551. Qin, Y., Feng, H., Chen, M., Wu, H. & Zheng, X. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. *Genes Dis* 5, 43–45 (2018).

552. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).

553. O'Sullivan, D. E., Johnson, K. C., Skinner, L., Koestler, D. C. & Christensen, B. C. Epigenetic and genetic burden measures are associated with tumor characteristics in invasive breast carcinoma. *Epigenetics* 11, 344–353 (2016).

554. Gruhn, B. *et al.* The expression of histone deacetylase 4 is associated with prednisone poor-response in childhood acute lymphoblastic leukemia. *Leukemia Res* 37, 1200–1207 (2013).

555. Kang, Z.-H. *et al.* Histone Deacetylase HDAC4 Promotes Gastric Cancer SGC-7901 Cells Progression via p21 Repression. *Plos One* 9, e98894 (2014).

556. Kaowinn, S., Kaewpiboon, C., Koh, S. S., Krämer, O. H. & Chung, Y.-H. STAT1-HDAC4 signaling induces epithelial-mesenchymal transition and sphere formation of cancer cells overexpressing the oncogene, CUG2. *Oncol Rep* 40, 2619–2627 (2018).

557. Mottet, D. *et al.* HDAC4 represses p21WAF1/Cip1 expression in human cancer cells through a Sp1-dependent, p53-independent mechanism. *Oncogene* 28, 243–256 (2009).

558. Cheng, W. *et al.* HDAC4, a prognostic and chromosomal instability marker, refines the predictive value of MGMT promoter methylation. *J Neuro-oncol* 122, 303–312 (2015).

559. Cheng, C. *et al.* HDAC4 promotes nasopharyngeal carcinoma progression and serves as a therapeutic target. *Cell Death Dis* 12, 137 (2021).

560. Cai, J.-Y. *et al.* Histone deacetylase HDAC4 promotes the proliferation and invasion of glioma cells. *Int J Oncol* 53, 2758–2768 (2018).

561. Wilson, A. J. *et al.* HDAC4 Promotes Growth of Colon Cancer Cells via Repression of p21. *Mol Biol Cell* 19, 4062–4075 (2008).

562. Zeng, L.-S. *et al.* Overexpressed HDAC4 is associated with poor survival and promotes tumor progression in esophageal carcinoma. *Aging Albany Ny* 8, 1236–1248 (2016).

563. Creighton, C. J. *et al.* Insulin-Like Growth Factor-I Activates Gene Transcription Programs Strongly Associated With Poor Breast Cancer Prognosis. *J Clin Oncol* 26, 4078–4085 (2008).

564. Farabaugh, S. M., Boone, D. N. & Lee, A. V. Role of IGF1R in Breast Cancer Subtypes, Stemness, and Lineage Differentiation. *Front Endocrinol* 6, 59 (2015).

565. Maris, C. *et al.* IGF-IR: a new prognostic biomarker for human glioblastoma. *Brit J Cancer* 113, 729–737 (2015).

566. Doepfner, K. T., Spertini, O. & Arcaro, A. Autocrine insulin-like growth factor-I signaling promotes growth and survival of human acute myeloid leukemia cells via the phosphoinositide 3-kinase/Akt pathway. *Leukemia* 21, 1921–1930 (2007).

567. Chng, W. J., Gualberto, A. & Fonseca, R. IGF-1R is overexpressed in poor-prognostic subtypes of multiple myeloma. *Leukemia* 20, 174–176 (2006).

568. Svalina, M. N. *et al.* IGF1R as a Key Target in High Risk, Metastatic Medulloblastoma. *Sci Rep-uk* 6, 27012 (2016).

569. Tirrò, E. *et al.* Prognostic and Therapeutic Roles of the Insulin Growth Factor System in Glioblastoma. *Frontiers Oncol* 10, 612385 (2021).

570. Vewinger, N. *et al.* IGF1R Is a Potential New Therapeutic Target for HGNET-BCOR Brain Tumor Patients. *Int J Mol Sci* 20, 3027 (2019).

571. Zhang, Y. *et al.* Pan-Cancer Analysis of IGF-1 and IGF-1R as Potential Prognostic Biomarkers and Immunotherapy Targets. *Frontiers Oncol* 11, 755341 (2021).

572. Wang, P., Mak, V. CY. & Cheung, L. WT. Drugging IGF-1R in cancer: new insights and emerging opportunities. *Genes Dis* (2022) doi:10.1016/j.gendis.2022.03.002.

573. Hua, H., Kong, Q., Yin, J., Zhang, J. & Jiang, Y. Insulin-like growth factor receptor signaling in tumorigenesis and drug resistance: a challenge for cancer therapy. *J Hematol Oncol* 13, 64 (2020).

574. Savary, C. *et al.* Depicting the genetic architecture of pediatric cancers through an integrative gene network approach. *Sci Rep-uk* 10, 1224 (2020).

575. Huether, R. *et al.* The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. *Nat Commun* 5, 3630 (2014).

576. Lawlor, E. R. & Thiele, C. J. Epigenetic Changes in Pediatric Solid Tumors: Promising New Targets. *Clin Cancer Res* 18, 2768–2779 (2012).

577. Ecker, J., Witt, O. & Milde, T. Targeting of histone deacetylases in brain tumors. *Cns Oncol* 2, 359–376 (2013).

578. Bielen, A. *et al.* Enhanced Efficacy of IGF1R Inhibition in Pediatric Glioblastoma by Combinatorial Targeting of PDGFRα/β. *Mol Cancer Ther* 10, 1407–1418 (2011).

579. Brown, M. S., Muller, K. E. & Pattabiraman, D. R. Quantifying the Epithelial-to-Mesenchymal Transition (EMT) from Bench to Bedside. *Cancers* 14, 1138 (2022).

580. Yoon, S. & Eom, G. H. HDAC and HDAC Inhibitor: From Cancer to Cardiovascular Diseases. *Chonnam Medical J* 52, 1–11 (2015).

581. Reardon, D. A. *et al.* Chromosome arm 6q loss is the most common recurrent autosomal alteration detected in primary pediatric ependymoma. *Genes Chromosom. Cancer* 24, 230–237 (1999).

582. Ward, S. *et al.* Gain of 1q and loss of 22 are the most common changes detected by comparative genomic hybridisation in paediatric ependymoma. *Genes Chromosom. Cancer* 32, 59–66 (2001).

583. Rand, V. *et al.* Investigation of chromosome 1q reveals differential expression of members of the S100 family in clinical subgroups of intracranial paediatric ependymoma. *Brit J Cancer* 99, 1136–1143 (2008).

584. Parker, J. J. *et al.* Intratumoral heterogeneity of endogenous tumor cell invasive behavior in human glioblastoma. *Sci Rep-uk* 8, 18002 (2018).

585. Bernstock, J. D. *et al.* Molecular and cellular intratumoral heterogeneity in primary glioblastoma: clinical and translational implications. *J Neurosurg* 133, 655–663 (2020).

586. Bergmann, N. *et al.* The Intratumoral Heterogeneity Reflects the Intertumoral Subtypes of Glioblastoma Multiforme: A Regional Immunohistochemistry Analysis. *Frontiers Oncol* 10, 494 (2020).

587. Liu, I. *et al.* The landscape of tumor cell states and spatial organization in H3-K27M mutant diffuse midline glioma across age and location. *Nat Genet* 1–14 (2022) doi:10.1038/s41588-022-01236-3.

588. Zhang, J. *et al.* 5-Hydroxymethylome in Circulating Cell-free DNA as A Potential Biomarker for Non-small-cell Lung Cancer. *Genom Proteom Bioinform* 16, 187–199 (2018).

589. Xu, L. *et al.* Deoxyribonucleic Acid 5-Hydroxymethylation in Cell-Free Deoxyribonucleic Acid, a Novel Cancer Biomarker in the Era of Precision Medicine. *Frontiers Cell Dev Biology* 9, 744990 (2021).

590. Sjöström, M. *et al.* 5-hydroxymethylcytosine as a liquid biopsy biomarker in mCRPC. *J Clin Oncol* 39, 148–148 (2021).

591. Hu, X. *et al.* Integrated 5-hydroxymethylcytosine and fragmentation signatures as enhanced biomarkers in lung cancer. *Clin Epigenetics* 14, 15 (2022).

592. Li, W. *et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res* 27, 1243–1257 (2017).