Title

Multi-omic analyses and organoid models for identification of therapeutic vulnerabilities and developmental origins in childhood cancer

Author
Sanders, Lauren M

Publication Date
2020

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

MULTI-OMIC ANALYSES AND ORGANOID MODELS FOR IDENTIFICATION OF THERAPEUTIC VULNERABILITIES AND DEVELOPMENTAL ORIGINS IN CHILDHOOD CANCER

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Lauren M. Sanders

June 2020

The Dissertation of Lauren M. Sanders is approved:

____
Professor David Haussler, chair

____
Professor Olena M. Vaske

____
Professor Sameer Agnihotri

____
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Multi-omic analyses and organoid models for identification of therapeutic vulnerabilities and developmental origins in childhood cancer

by

Lauren M. Sanders

Pediatric cancers are different from adult cancers in that they often have few targetable DNA mutations, and in most cases are thought to be developmental in origin rather than environmental. While overall survival rates for pediatric cancer have increased in the past few decades, there remain several difficult-to-treat pediatric cancer types, including deadly pediatric diffuse midline and brainstem gliomas. However, advances in genomics data generation and analysis methods have made it possible to start identifying the dominant signaling pathways, developmental origins, and ultimately therapeutic vulnerabilities of these tumors.

Here I present my work in the UCSC Treehouse Childhood Cancer Initiative using comparative gene expression analysis methods to identify a rare cancer subtype with treatment and outcome implications (Chapter 2). This work demonstrates the utility of gene expression data for molecular subtyping in the clinic, especially in rare or difficult-to-diagnose pediatric cancers. I also present my work identifying a developmental window of opportunity for the histone H3 K27M mutation event, which characterizes a majority of brainstem gliomas and is associated with especially poor overall survival (Chapter 3). The cell-of-origin and developmental timing for

gliomagenesis in this tumor type has been a subject of much research in the past decade, and my work helps to address this clinically relevant issue.

Finally, in Chapter 4 I present an analysis of multiple types of genomic data across thousands of primary tumors and diverse cancer laboratory models (cell lines, organoids, and patient-derived mouse xenografts). I report that cancer organoids have the advantage of being relatively accurate cellular representations of primary tumors, and time- and cost-effective. Therefore, in Chapter 5 I present the development of a novel human embryonic stem cell based cerebral organoid model of the histone H3 K27M mutation in early human embryonic brain development. This model can be used to answer outstanding questions in the field to better help treat these devastating tumors.

To all the families who have been affected
by childhood cancer.

Acknowledgements

I gratefully acknowledge the support of my advisors Drs. Olena Vaske and David Haussler, and my mentors Drs. Sameer Agnihotri and Sofie Salama.

I thank the members of the Treehouse Childhood Cancer Initiative and the Haussler-Salama lab for being outstanding fellow researchers and friends.

I thank the undergraduate students who have shared my passion and worked with me on each of my projects.

I am grateful for encouragement from my parents, my brother and sister, and for the unconditional support of my husband.

Chapter 1. Introduction

1.1 Pediatric Cancer Background

Although survival rates for pediatric cancer have dramatically improved over the past five decades, cancer is still the leading cause of disease-related death in childhood. With recent advances in sequencing technology, pediatric pan-cancer genomic studies at several clinical centers have well described the childhood cancer mutational and epigenetic landscape[1–3]. It is now clear that the mutational burden is significantly lower in pediatric cancer overall as compared to adult cancer[2]. But epigenetic alterations and gene expression dysregulation are prevalent in pediatric cancer[4]. These results, combined with experimental modeling studies, have indicated that many pediatric cancers arise from progenitor cells during important developmental windows[5,6]. The oncogenesis mechanisms are varied, including chromosomal fusion events, histone mutations, epigenetic modifier mutations, or non-genetic epigenetic dysregulation events that are still poorly understood[6]. The common theme is that each of these oncogenetic events stalls or otherwise interferes with normal development, resulting in maintenance of a progenitor cell state which is particularly susceptible to tumorigenesis. It is thought that in many tumor types, specific developmental times and cell types are most vulnerable to these alterations. The resulting tumors often harbor gene expression signatures which are strongly reminiscent of the developmental cell of origin.

While the 5-year survival rate for pediatric cancer is now at 80%, the remaining 20% of patients are very difficult to treat effectively[7]. Due to the developmental origins of these diseases, many pediatric cancers display overexpression of oncogenes and oncogenic pathways without a targetable oncogenic mutation. Although many pediatric cancers harbor an epigenetic alteration that indicates their developmental origin, epigenetic alterations are in general very difficult to target therapeutically[8]. Epigenetically-directed treatment regimens can be especially dangerous in pediatric patients, because normal development requires delicately maintained epigenetic processes and children are particularly susceptible to stunted growth, tissue development, and neurocognitive function[6].

Faced with this evidence, over the past few years several clinical centers have begun to incorporate high-throughput RNA sequencing (RNA-seq) into treatment decision-making[9–12]. Analysis of cancer gene expression via RNA-seq has been shown in many cases to increase the identification of actionable genetic abnormalities, especially for children with relapsed or refractory cancer who have failed standard-of-care and run out of treatment options. Although there are currently few targeted molecular therapies designed specifically for pediatric cancers, the adult cancer field has generated a wide library of precision medicine targeted agents. Because childhood cancers often display oncogene overexpression or activated oncogenic signaling, it is possible in many cases to repurpose a targeted therapeutic which has displayed efficacy in adult cancers.

In addition to its usefulness in detecting aberrant oncogenic signaling, cancer gene expression often carries the signature of the tumor cell and tissue of origin. Identification of the cell and tissue of origin of many pediatric cancer types is an ongoing research effort. Once we better understand the developmental timing of each cancer type, we can identify therapeutic vulnerabilities for improved treatment. Single cell RNA-seq has proven particularly useful in this regard, as this technique can characterize individual cell type populations in each tumor[13–16]. Separately, identification of the tissue of origin or tissues with similar gene expression profiles can be of immediate clinical relevance for molecular subtyping and classification of pediatric cancers, which may be especially difficult to classify through histopathological analysis alone[17].

1.2 Gene Expression in Pediatric Cancer

Previous studies have shown that transcriptomic analysis of patient cancer samples can aid in comprehensive molecular characterization of poorly understood cancer types. The most influential transcriptomic studies have come from The Cancer Genome Atlas, which included over 11,000 patient tumor samples and yielded 27 papers on 33 of the most common adult cancer types[18]. These studies influenced the inclusion of genomics in cancer treatment decisions, and helped establish the paradigm that comparative transcriptomic analysis of patient samples can yield clinically relevant findings.

This paradigm has held up through its application in pediatric cancer, as several studies have characterized the transcriptomic landscape of pediatric cancers such as leukemia[1], neuroblastoma[19], and glioma[3]. Many of these studies were made possible through the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) project to generate genomic data on pediatric cancers. However, the most difficult-to-treat cancers are rare, and an individual institution may only see a few cases per year, making it particularly challenging to perform large-scale comparative transcriptomic analyses. Therefore, the UC Santa Cruz Treehouse Childhood Cancer Initiative (Treehouse) assembles cancer RNA-seq datasets from many sources to enable data sharing and analysis[20]. These datasets together make up the Treehouse cancer compendium of over 12,000 uniformly processed tumor RNA-seq samples.

Accordingly, my thesis research leverages Treehouse transcriptomic data and other publicly available datasets to identify therapeutic vulnerabilities and developmental origins of the most devastating pediatric brain cancers. While survival rates of other pediatric cancers have risen, the overall survival rate for pediatric brainstem gliomas (diffuse intrinsic pontine glioma, DIPG) remains at less than a year for initial diagnosis[21]. Over four decades of clinical trials in DIPG have resulted in several failed therapies, partly because sequencing analysis of biopsy samples only began in the early 2000's[22–24]. A majority of these tumors are characterized by a lysine-to-methionine mutation in the K27 residue of histone H3, resulting in loss of the

H3K27 trimethyl transcriptional repressive mark and causing widespread oncogenic dysregulation of developmental genes[25].

The rare and deadly nature of these brain tumors has resulted in a limited amount of genomic information, a situation which has only improved somewhat in the last decade. Since much of our knowledge about these rare brain tumors has derived from experimental models, part of my thesis research focuses on evaluating the biologically representative nature of tumor models and identifying their limits. In particular, for cancers with epigenetic and developmental origins, models derived from biopsy material may not be useful for studying the early events leading to tumorigenesis. Ultimately, this led me to pioneer the development of a new experimental model of H3K27M in early brain development.

Finally, I would like to present my thesis research as a message for future researchers. At the time of this writing, a search for the keyword "cancer" on the National Center for Biotechnology Information Gene Expression Omnibus database returns 768,734 public human datasets. Only a fraction of these high-dimensional gene expression data are currently being repurposed to answer additional research questions. The majority of the data in these studies is often only used for the original authors' purpose, which too often is to study only one or two biological gene sets, leaving a great deal of data uninvestigated. The *status quo* in biology is to pioneer a new experimental study instead of mining the thousands of publicly available datasets and leveraging the power of previously published expression data. In some cases, a novel experimental study is justified, but the huge amount of existing data should warrant at least a preliminary analysis. Accordingly, a secondary motivation of my research is to demonstrate the possibility and the necessity of repurposing existing expression datasets from a variety of sample types and experimental models to develop and test novel biological hypotheses.

Chapter 2: Treehouse comparative RNA-seq analysis for molecular subtyping of pediatric cancer

2.1 Chapter Introduction

To aid in identification of personalized treatments for individual childhood cancer patients, the Treehouse Childhood Cancer Initiative has developed a pipeline for comparative analysis of RNA expression (Treehouse CARE). Treehouse deploys this analysis in partnership with multiple clinical sites including Stanford, University of California at San Francisco (UCSF), and Children's Hospital of Orange County. The Treehouse CARE analysis provides two clinically-relevant outputs: first, identification of oncogenes with outlier expression in a single tumor sample as compared to a background cohort, and second, identification of other tumors within the Treehouse cancer compendium with similar gene expression profiles to the tumor of interest. The first goal can aid in identification of targeted therapeutics, and the second goal can help in molecular subtyping or classification.

Treehouse uses the UCSC TumorMap tool[26] to identify and visualize molecular similarity between a focus sample and the top 6 most similar tumors in the Treehouse cancer compendium. TumorMap was developed as a tool for finding similarities among tumors from different tissues, based on TCGA findings that clinically and prognostically relevant cancer subtypes can originate from different tissues but share underlying signaling[27]. TumorMap similarity analysis is particularly relevant in the context of pediatric cancers, for which the tissue-of-origin is not always easy to determine. In many cases, identification of similar tumors among a background cancer cohort can help assign a molecular subtype based on clinically relevant cancer signaling.

The following publication details a case of ovarian cancer in a 10-year-old girl treated at Stanford, in which the Treehouse CARE pipeline and TumorMap analysis helped refine the subtype diagnosis with implications for treatment and outcome. Although I served as the case analyst for several Treehouse cases, the original analysis for this case was performed by Du Linh Lam, a previous Treehouse research analyst. I was responsible for re-analyzing the case against a more recent version of the Treehouse compendium, completing all writing, and generating all figures for this publication except Figure 1 in which the H&E images were provided by Inge Behroozfard and Dr. Florette Hazard at UCSF.

COLD SPRING HARBOR
Molecular Case Studies

RESEARCH REPORT

Comparative RNA-seq analysis aids in diagnosis of a rare pediatric tumor

Lauren M. Sanders,[1] Arun Rangaswami,[2] Isabel Bjork,[1] Du Linh Lam,[1] Holly C. Beale,[3] Ellen Towle Kephart,[1] Ann Durbin,[1] Katrina Learned,[1] Rob Currie,[1] A. Geoffrey Lyle,[3] Jacob Pfeil,[1] Avanthi Tayi Shah,[4] Alex G. Lee,[4] Stanley G. Leung,[4] Inge H. Behroozfard,[4] Marcus R. Breese,[4] Jennifer Peralez,[2] Florette K. Hazard,[2] Norman Lacayo,[2] Sheri L. Spunt,[2] David Haussler,[1,5] Sofie R. Salama,[1,5] E. Alejandro Sweet-Cordero,[4] and Olena M. Vaske[3]

[1]Department of Biomolecular Engineering, UC Santa Cruz Genomics Institute, Santa Cruz, California 95064, USA; [2]Stanford University School of Medicine and Stanford Cancer Institute, Stanford, California 94305, USA; [3]Department of Molecular, Cell and Developmental Biology, UC Santa Cruz Genomics Institute, Santa Cruz, California 95064, USA; [4]Department of Pediatrics, Division of Hematology and Oncology, University of California San Francisco, San Francisco, California 94143, USA; [5]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA

Abstract Gliomatosis peritonei is a rare pathologic finding that is associated with ovarian teratomas and malignant mixed germ cell tumors. The occurrence of gliomatosis as a mature glial implant can impart an improved prognosis to patients with immature ovarian teratoma, making prompt and accurate diagnosis important. We describe a case of recurrent immature teratoma in a 10-yr-old female patient, in which comparative analysis of the RNA sequencing gene expression data from the patient's tumor was used effectively to aid in the diagnosis of gliomatosis peritonei.

[Supplemental material is available for this article.]

INTRODUCTION

Immature ovarian teratomas are malignant tumors of germ cell origin (Gheorghisan-Galateanu et al. 2013). Teratomas are the most common germ cell tumor, but, in rare cases, immature teratoma can occur with gliomatosis peritonei, which is characterized by mature glial tissue in the peritoneum (Liang et al. 2015). The presence of mature glial tissue implants can indicate a favorable prognosis in patients with immature ovarian teratoma (Marwah et al. 2016). However, all lesions must be sampled to confirm mature histological status, and full excision is important. Additionally, recurrence potential is high, requiring careful follow-up and monitoring.

Gliomatosis peritonei can be difficult to identify through histopathological analysis alone. Recent studies have shown that molecular analysis can aid cancer type classification (Cancer Genome Atlas Research Network 2015). In many cases, DNA sequencing and variant identification can help subtype cancers by grade and outcome. However, the paucity of recurrent DNA mutations in rare pediatric cancers can make variant-based disease classification difficult. RNA sequencing (RNA-seq) of tumor gene expression can provide additional classification information, through a comparative analysis of RNA-seq from the pediatric tumor and with a compendium of RNA-seq data from known cancer types (Newton et al. 2018). In this study, we describe the use of comparative RNA-seq analysis in a case of relapsed

Corresponding author: olena@ucsc.edu

Ontology term: neoplasm of the nervous system

Published by Cold Spring Harbor Laboratory Press

doi:10.1101/mcs.a004317

pediatric immature teratoma cooccurring with gliomatosis peritonei with no informative DNA mutations.

RESULTS

Clinical Presentation and Family History

A 10-yr-old female patient was diagnosed with immature teratoma, relapsed to the pericardium and diaphragm. Treatment history included laparotomy with resection of ovarian and fallopian tube mass, video-assisted thoracoscopic resection of lung nodule, and resection of diaphragmatic and pericardial lesions. The pericardial lesion was submitted for RNA sequencing. Foundation Medicine DNA testing of the diaphragmatic lesion identified only one variant: MLL3 p.C310S. This mutation was not informative for diagnosis or subtyping. Whole-genome sequencing (WGS) detected three somatic coding variants in genes *FCGR1A*, *ANKRD36C*, and *HLA-DRB1* at hg38 Chr 1:g.[149790230C>T], Chr 2:g.[95855406C>G], and Chr 6:g.[32584172C>G] (Table 1). Interestingly, *ANKRD36* has been previously characterized by TCGA as a significantly mutated gene in adult glioblastoma (Brennan et al. 2013).

Histologic sections of the pericardium (Fig. 1A) showed fibroadipose tissue and mature glial cells, with lymphovascular invasion. Histologic sections of the right diaphragm (Fig. 1B) showed extensive involvement by teratoma, with no malignant elements present.

Genomic Analyses

The patient was enrolled in the "Clinical Implementation of Genomic Analysis in Pediatric Malignancies" study at Stanford University, and through this trial, her tumor RNA-seq data set was analyzed. This analysis uses an *N*-of-1 analysis approach, which compares an individual pediatric tumor to a cancer compendium of uniformly processed RNA-seq data from 11,456 other tumors (https://treehousegenomics.ucsc.edu/public-data/). This approach aids in the molecular classification of the pediatric tumor through the identification of most similar tumors in the Treehouse cancer compendium.

We calculated pairwise Spearman correlation scores between the patient's tumor and all tumors in the Treehouse cancer compendium. For 232 samples, the pairwise correlation scores with the focus sample exceeded the 95th percentile correlation score in the cancer compendium (0.875); 228 of these (98%) were glioma or glioblastoma multiforme samples, and the remaining four were various brain tumors (Supplemental Fig. S1). The top 6 most correlated tumors are shown in Table 2. In addition, glioma samples had a significantly higher correlation to the patient's tumor than to other tumor types in the cancer compendium (Supplemental Fig. S2). A neural network classification approach (Abadi et al. 2016) also classified the patient's tumor as most similar to glioma (Supplemental Fig. S3). Overall this indicates a strong gene expression similarity between the patient's pericardial lesion and high-grade adult glioma tumors.

Table 1. Variants detected in the pericardial lesion by whole-genome sequencing

Gene	Chr	HGVS DNA reference	HGVS protein reference	Variant type	Predicted effect (substitution, deletion, etc.)	dbSNP/ dbVar ID	Genotype (heterozygous/ homozygous)
FCGR1A	1	GRCh38	NP_000557.1	Missense	Substitution	rs637882	Heterozygous
ANKRD36C	2	GRCh38	NP_001297083.1	Missense	Substitution	rs77972623	Heterozygous
HLA-DRB1	6	GRCh38	NP_001230894.1	Missense	Substitution	rs16822805	Heterozygous

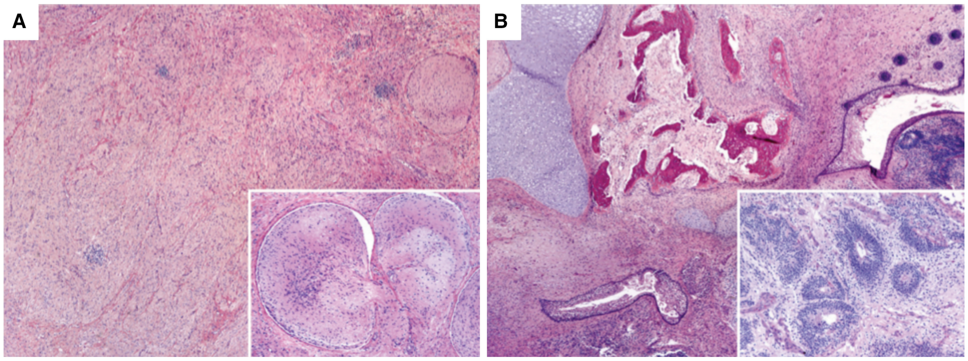


Figure 1. Teratoma involving the pericardium and diaphragm. (*A*) (H&E stain, 40×) Pericardial involvement by mature glial implant composed of mature neurons, neuropil, and schwannian stroma. (*Inset*) (H&E stain, 100×) Nodules of implants within vascular spaces. (*B*) (H&E stain, 40×) Diaphragm involvement by a mixture of mature and immature germ cell components. (*Inset*) (H&E stain, 100×) Immature neuroepithelium forming rosettes set within neuropil.

The TumorMap algorithm (see Methods) was used to visualize the top six most correlated tumors in the context of all tumors in the Treehouse cancer compendium (Newton et al. 2017). TumorMap visualization generates a two-dimensional (2D) "map" of the similarity between tumor RNA-seq samples based on pairwise Spearman correlation. The six most correlated samples to the focus tumor are indicated using red pins (Fig. 2). All six most correlated samples fall in a modular cluster that includes both adult and pediatric glioma (yellow) and glioblastoma brain tumors (green). This indicates that the patient's pericardial lesion is most transcriptionally similar to high-grade glial brain tumors.

Diagnosis of Gliomatosis Peritonei

As a result of the combined histologic and genomic analysis, the patient was subsequently diagnosed with gliomatosis peritonei. This diagnosis is consistent both with the presence of glial tissue in the pericardial lesion, and with the Treehouse genomic finding that the patient's tumor is most similar to high-grade glioma tumors. Molecular similarities between high-grade glioma and gliomatosis peritonei include high expression of the stem cell marker *SOX2* and low expression of transcription factors *OCT4* and *NANOG* (Nogales et al. 2014; Liang et al. 2015). Consistent with these characteristics, Figure 3A shows that the patient's

Table 2. The top six most correlated RNA-seq samples to the patient's RNA-seq sample belong to older patients with glioma or glioblastoma

Sample ID	Diagnosis (grade)	Histology	Patient age (yr)	Spearman correlation
TCGA-DU-7012-01	Glioma (3)	Astrocytoma	74	0.93
THR14_0312_S01	Glioma (3)	Astrocytoma	18	0.92
TCGA-CS-4941-01	Glioma (3)	Astrocytoma	67	0.91
TCGA-HT-7680-01	Glioma (2)	Astrocytoma	32	0.91
TCGA-DU-8158-01	Glioma (3)	Astrocytoma	57	0.91
TCGA-28-5215-01	Glioblastoma (4)	Astrocytoma	62	0.91

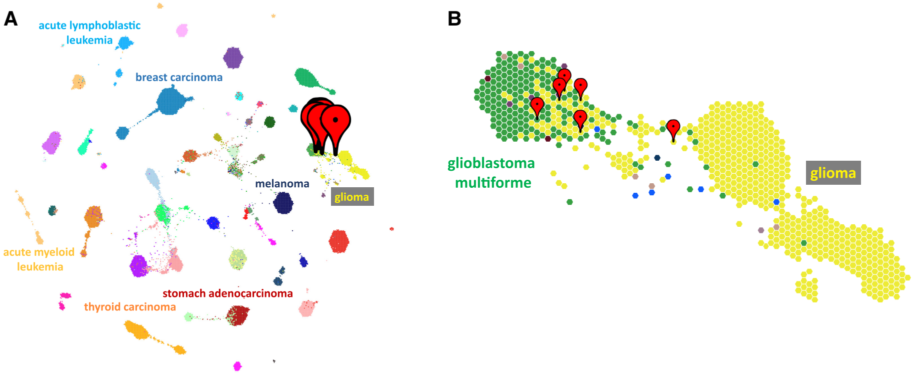


Figure 2. TumorMap clustering visualization of the Treehouse cancer compendium. (*A*) Treehouse cancer compendium v8 shown visualized with the TumorMap tool. Each colored dot represents an individual patient's tumor RNA-seq data. Tumors are grouped based on RNA-seq similarity and selected tumor types are labeled. The top six most similar tumors to the patient's pericardial lesion are indicated with red pins. (*B*) Zoomed-in image of the location of the six most correlated tumors on the TumorMap. Five out of six of the most correlated tumors fall in the leftmost brain tumor cluster, which contains the majority of glioblastoma samples and high-grade glioma samples.

tumor expresses *SOX2* at very high levels, comparable to glioma and glioblastoma tumors in the Treehouse cancer compendium. Figure 3B,C shows that the patient's tumor also expresses very low levels of *OCT4* and *NANOG*, similar to gliomas.

The patient underwent resection of the pericardial gliomatosis implant as well as the diaphragmatic immature teratoma implant. Two years post-resection, she was healthy and was discharged from oncology.

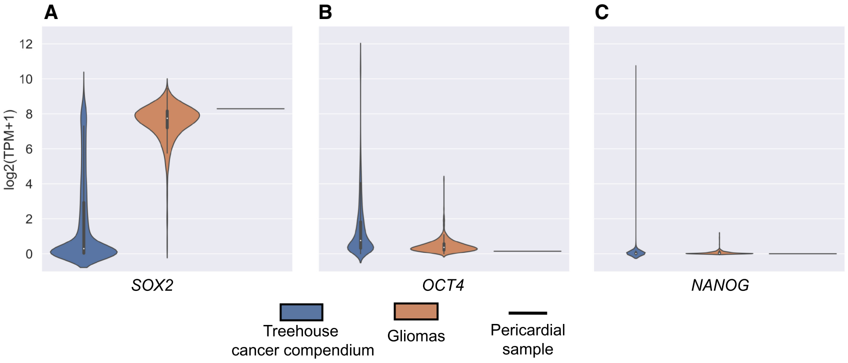


Figure 3. Expression levels of *SOX2*, *OCT4*, and *NANOG*. (*A*) *SOX2* expression. The majority of the tumors in the Treehouse cancer compendium express *SOX2* at very low levels. However, the glioma tumors exhibit exceptionally high *SOX2* expression. The patient's pericardial tumor expresses *SOX2* at a level comparable with the glioma tumor group. (*B*,*C*) *OCT4*, *NANOG* expression. The majority of Treehouse cancer compendium tumors express *OCT4* and *NANOG* at very low levels, including most gliomas. The patient's pericardial tumor also expresses *OCT4* and *NANOG* at very low levels, comparable with the glioma tumor group.

DISCUSSION

We describe here the utility of comparative RNA-seq analysis using the TumorMap method for molecular classification and diagnosis of a rare pediatric tumor. The TumorMap algorithm has been used previously to describe the global similarity between tumor types and to discover novel subtypes (Ceccarelli et al. 2016; Farshidfar et al. 2017), but this is the first published use of TumorMap for *N*-of-1 tumor classification.

We demonstrate the utility of the publicly available Treehouse cancer compendium, an extensive database of thousands of tumor RNA-seq samples. *N*-of-1 comparison of a pediatric pericardial tumor aided in molecular classification by identifying other tumors with similar gene expression profiles, all of which were glioma or glioblastoma brain tumors. The diagnoses of the most similar tumors were clinically meaningful and consistent with the subsequent diagnosis of gliomatosis peritonei in this pediatric patient.

The methods described here are widely applicable for enabling precision molecular classification or diagnosis in cases of rare or difficult-to-diagnose cancer. Beyond the application described here, the Treehouse cancer compendium and TumorMap clustering analysis can also be used to accurately identify molecular subtypes of cancer, a useful application for cancer types with the subtype-dependent outcome and treatment differences (Newton et al. 2017). Additionally, in some cases it can be impossible to determine cancer tissue of origin using clinical or radiologic data, making it difficult to design a treatment regimen (Park et al. 2018). These methods could aid diagnosis and treatment strategies for both childhood and adult cancers with unknown tissue of origin, by clustering a tumor tissue RNA-seq sample in the TumorMap and identifying other tumors with most similar molecular features. Overall, the comparative RNA sequencing analysis presented here is a powerful tool for precision molecular subtype classification and diagnosis of cancer.

METHODS

Tissue Source and Processing

A sample of the pericardial lesion was flash frozen, embedded into OCT, sectioned to a depth of 5 µm, and stained with H&E. The sample was evaluated for tumor content by a certified pathologist. The tumor was macro-dissected from the OCT block to a depth of up to 5 mm, disrupted with a mortar and pestle under liquid nitrogen, and homogenized with a QIAshredder (QIAGEN, 79654). Nucleic acids were extracted using the AllPrep DNA/RNA kit (QIAGEN, 80204). The RNA integrity was quantified using the RNA 6000 Pico kit (Agilent, 5067-1513) on the Bioanalyzer (Agilent).

Whole-Genome Sequencing

WGS was performed on the pericardial lesion. Average WGS depth was 60.67× (tumor), 29.10× (germline). The read length was 2 × 150 bp (paired-end). The read depths for reported somatic variants are as follows: Chr 1:g.[149790230C>T] Tumor ref,alt: 45, 6; Germline ref,alt: 31,0. Chr 2:g.[95855406C>G] Tumor ref,alt: 59, 9; Germline ref,alt: 35, 1. Chr 6:g.[32584172C>G] Tumor ref,alt: 11, 5; Germline ref, alt: 19, 0 (Table 3).

RNA Sequencing

Libraries were prepared using the TruSeq Stranded mRNA kit (Illumina, RS-122-2101) with an input of 400 ng in accordance with manufacturer's instructions. All manufacturer controls were used in preparation. Libraries were quantified using the High Sensitivity DNA kit (Agilent, 5067-4626) on the BioAnalyzer (Agilent). Sequencing was performed on the

Table 3. Sequencing coverage table for somatic variants detected in the pericardial lesion

Gene	Chr	Pos	Ref allele	Alt allele	Avg depth (tumor)	Ref (tumor)	Alt (tumor)	Avg depth (germline)	Ref (germline)	Alt (germline)
FCGR1A	1	149790230	C	T	60.67	45	6	29.10	31	0
ANKRD36C	2	95855406	C	G	60.67	59	9	29.10	35	1
HLA-DRB1	6	32584172	C	G	60.67	11	5	29.10	19	0

(Avg depth) average read depth, (Ref) number of reads for reference allele, (Alt) number of reads for alternative allele.

Illumina HiSeq 4000 with PE75 chemistry at the Stanford Functional Genomics Facility. The total sequence depth for this sample was 97,983,221 reads.

Comparative RNA-seq Analysis

The RNA-seq data from the patient's pericardial lesion was processed at UC Santa Cruz and gene expression quantification was performed using the TOIL RNA-seq pipeline (Vivian et al. 2017). Genome alignment was performed with genome assembly hg38. RSEM quantification TPM measurements were used as input to normalization and compendium building. The Treehouse comparative RNA-seq analysis is designed to compare an *N*-of-1 sample against a larger background cohort of RNA-seq data from many cancer samples. Pairwise Spearman correlation scores are calculated between the gene expression vector from the focus sample and all other samples in the background cohort. The top six most correlated samples are used to identify tumor types with gene expression are most similar to the focus sample.

TumorMap

TumorMap (tumormap.ucsc.edu) is a hexagonal 2D representation of similarity between samples based on gene expression (Newton et al. 2017). The spatial representation of samples in the TumorMap is based on vector similarity using a multidimensional scaling tool called OpenOrd. Sample information, such as patient age, tumor grade, or clinical outcome, can be displayed as attributes on the map. TumorMap also has built-in correlation analysis tools for discovering correlations between attributes.

Neural Network Classification

As validation, we trained a fully connected neural network to classify disease type. The network was comprised of an input layer, a batch normalization layer, two hidden layers of size 32 with a dropout of 0.5 and relu activation, and a one-hot output layer, one per disease type with sigmoid activation. We trained the network on 80% of the data using binary cross entropy as a loss function and tested it using the remaining 20% of the data. The train and test were stratified by disease to ensure an equal representation of each disease type. The network was implemented using the Keras library in a Jupyter notebook. All code is available here: https://nbviewer.jupyter.org/github/rcurrie/pancan-gtex/tree/cf249e64f7ac5da95c264f946f2ae5fd69410f63/

ADDITIONAL INFORMATION

Data Deposition and Access

All processed RNA sequencing data is publicly available in the Treehouse cancer compendium: https://treehousegenomics.soe.ucsc.edu/public-data/. The interpreted variants were submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) under accession numbers SCV000994650–SCV000994652.

Ethics Statement

The patient was enrolled on a Stanford protocol "Clinical implementation of genomic analysis in pediatric malignancies" (IRB#34383). The UCSC Treehouse protocol was approved by the institutional review board at the University of California Santa Cruz (No. HS2648).

Acknowledgments

We thank the patient and her family for their support. We thank the California Initiative to Advance Precision Medicine, St Baldrick's Foundation, Alex's Lemonade Stand Foundation, Unravel Pediatric Cancer, Team G Childhood Cancer Foundation, Live for Others Foundation, and Lucile Packard Children's Hospital Stanford.

Author Contributions

L.M.S., O.M.V., I.B., S.R.S., S.L.S., E.A.S.-C., A.R., and A.T.S. were responsible for conception and design. D.L.L., H.C.B., E.T.K., A.D., K.L., R.C., A.G.L., J.P., A.G.L., S.G.L., I.H.B., M.R.B., F.K.H., D.H., and S.R.S. collected and assembled the data. L.M.S., O.M.V., D.L.L., H.C.B., E.T.K., A.D., K.L., R.C., A.G.L., J.P., and S.R.S. analyzed and interpreted the data. A.R., A.T.S., E.A.S.-C., N.L., and S.L.S. provided study material or patients. I.B., J.P., and A.D. provided administrative support. All authors wrote the manuscript.

Competing Interest Statement

The authors have declared no competing interest.

Received May 10, 2019; accepted in revised form August 15, 2019.

Funding

This work was supported by the California Initiative to Advance Precision Medicine, St Baldrick's Foundation, Alex's Lemonade Stand Foundation, Unravel Pediatric Cancer, Team G Childhood Cancer Foundation, Live for Others Foundation, and Lucile Packard Children's Hospital Stanford.

REFERENCES

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al. 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467.

Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, et al. 2013. The somatic genomic landscape of glioblastoma. *Cell* **155:** 462–477. doi:10.1016/j.cell.2013.09.034

Cancer Genome Atlas Research Network. 2015. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med* **372:** 2481–2498. doi:10.1056/NEJMoa1402121

Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, et al. 2016. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164:** 550–563. doi:10.1016/j.cell.2015.12.028

Farshidfar F, Zheng S, Gingras M-C, Newton Y, Shih J, Robertson AG, Hinoue T, Hoadley KA, Gibb EA, Roszik J, et al. 2017. Integrative genomic analysis of cholangiocarcinoma identifies distinct IDH-mutant molecular profiles. *Cell Rep* **18:** 2780–2794. doi:10.1016/j.celrep.2017.02.033

Gheorghisan-Galateanu A, Terzea DC, Carsote M, Poiana C. 2013. Immature ovarian teratoma with unusual gliomatosis. *J Ovarian Res* **6:** 28. doi:10.1186/1757-2215-6-28

Liang L, Zhang Y, Malpica A, Ramalingam P, Euscher ED, Fuller GN, Liu J. 2015. Gliomatosis peritonei: a clinicopathologic and immunohistochemical study of 21 cases. *Mod Pathol* **28:** 1613–1620. doi:10.1038/modpathol.2015.116

Marwah N, Batra A, Gupta S, Singhal SR, Sen R. 2016. Gliomatosis peritonei arising in setting of immature teratoma of ovary: a case report and review of literature. *J Obstet Gynecol India* **66:** 192–195. doi:10.1007/s13224-015-0708-7

Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, Weinstein AS, Baertsch R, Salama SR, Ellrott K, et al. 2017. TumorMap: exploring the molecular similarities of cancer samples in an interactive portal. *Cancer Res* **77:** e111–e114. doi:10.1158/0008-5472.CAN-17-0580

Newton Y, Rassekh SR, Deyell RJ, Shen Y, Jones MR, Dunham C, Yip S, Leelakumari S, Zhu J, McColl D, et al. 2018. Comparative RNA-sequencing analysis benefits a pediatric patient with relapsed cancer. *JCO Precis Oncol* **2:** 1–16. doi:10.1200/PO.17.00198

Nogales FF, Dulcey I, Preda O. 2014. Germ cell tumors of the ovary: an update. *Arch Pathol Lab Med* **138:** 351–362. doi:10.5858/arpa.2012-0547-RA

Park CK, Malinowski DP, Cho NH. 2018. Diagnostic algorithm for determining primary tumor sites using peritoneal fluid. *PLoS One* **13:** e0199715. doi:10.1371/journal.pone.0199715

Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, Pfeil J, Narkizian J, Deran AD, Musselman-Brown A, et al. 2017. Toil enables reproducible, open source, big biomedical data analyses. *Nat Biotechnol* **35:** 314–316. doi:10.1038/nbt.3772

COLD SPRING HARBOR
Molecular Case Studies

Comparative RNA-seq analysis aids in diagnosis of a rare pediatric tumor

Lauren M. Sanders, Arun Rangaswami, Isabel Bjork, et al.

Cold Spring Harb Mol Case Stud 2019, **5:** a004317
Access the most recent version at doi:10.1101/mcs.a004317

Chapter 3: Comparative analysis and data re-use to investigate developmental origins of histone mutant pediatric glioma

3.1 Chapter Introduction

There is an open question in the pediatric brain cancer field about the developmental timing and cell of origin for histone H3 K27M mutant gliomas. Several studies have addressed this question, and have contributed valuable insights, but the issue remains unclear. This is in part due to limitations of experimental mouse and cell models to recapitulate accurately the oncogenic histone mutation event in early neural development. It is important to understand the origins of these tumors since the H3 K27M mutation results in global transcriptional dysregulation of thousands of developmental genes. Identifying the correct developmental cell of origin is the key to understanding the underlying signaling driving the mature tumor, and leveraging this knowledge into therapeutic opportunities.

As a doctoral student in the labs of Drs. David Haussler and Olena Vaske, I was uniquely positioned to help answer this question, because the Haussler lab specializes in early cerebral development, while the Vaske lab / Treehouse has assembled a large cohort of pediatric glioma RNA-seq data.

In this study, I performed a unique comparative analysis, leveraging the publicly available high-grade pediatric glioma (pHGG) RNA-seq cohort in the

Treehouse cancer compendium and a previously published normal cerebral organoid development RNA-seq dataset from the Haussler lab[28]. I also included a publicly available single-cell RNA-seq glioma dataset from the Broad Institute[14]. Following the tradition of previous landmark papers which used single cell RNA-seq to deconvolute signatures in bulk data[16,29], this study links H3K27M-specific expression to specific types of early neural cells, identifying a developmental window for H3K27M-driven tumorigenesis. In addition to its novel biological contributions, this study is a demonstration of the power of data reanalysis as the main results come from comparative analysis of three independent datasets.

I led this study and was responsible for the analysis and writing. The contributions of others are as follows: the original analysis of the Treehouse pHGG cohort was performed by Allison Cheney (co-first author) and she developed the epithelial mesenchymal transition hypothesis via literature review and her analysis. Allison contributed Figure 2A (top panel) and wrote the Introduction, Lucas Seninge contributed Figure 2E, and Anouk van den Bout contributed Figure 4C and E.

This manuscript has been submitted to *Giga Science*, April 2020.

Identification of a differentiation stall in epithelial mesenchymal transition in histone H3 mutant diffuse midline glioma

Lauren M. Sanders[1,4*#], Allison Cheney[2#], Lucas Seninge[1,4], Anouk van den Bout[2,4], Marissa Chen[2,4], Holly C. Beale[2,4], Ellen Towle Kephart[4], Jacob Pfeil[1,4], Katrina Learned[4], A. Geoffrey Lyle[2,4], Isabel Bjork[4], David Haussler[1,3,4], Sofie R. Salama[1,3,4+], Olena M. Vaske[2,4+]

[1]Department of Biomolecular Engineering, [2]Department of Molecular, Cell and Developmental Biology, [3]Howard Hughes Medical Institute, [4]University of California Santa Cruz Genomics Institute, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA, USA, 95064

*Corresponding author

#Co-first author

+Co-senior author

Abstract

Background

Diffuse midline gliomas with Histone H3 K27M (H3K27M) mutations occur in early childhood and are marked by an invasive phenotype and global decrease in H3K27me3, an epigenetic mark which regulates differentiation and development. H3K27M mutation timing and effect on early embryonic brain development are not fully characterized.

Results

We analyzed multiple publicly available RNA sequencing datasets to identify differentially expressed genes between H3K27M and nonK27M pediatric gliomas. We found that genes involved in the epithelial-mesenchymal transition (EMT) were significantly overrepresented among differentially expressed genes. Overall, the expression of pre-EMT genes was increased in the H3K27M tumors as compared to nonK27M tumors, while the expression of post-EMT genes was decreased. We hypothesized that H3K27M may contribute to gliomagenesis by stalling an EMT in early brain development, and evaluated this hypothesis by employing another publicly available dataset of single-cell and bulk RNA sequencing data from developing cerebral organoids. This analysis revealed similarities between H3K27M tumors and pre-EMT normal brain cells. Finally, a previously published single-cell RNA sequencing dataset of H3K27M and nonK27M gliomas revealed subgroups of cells at different stages of EMT. In particular, H3.1K27M tumors resemble a later EMT stage compared to H3.3K27M tumors.

Conclusions

Our data analyses indicate that this mutation may be associated with EMT arrest, and that H3K27M cells preferentially exist in a pre-EMT cell phenotype. This study demonstrates how novel biological insights could be derived from combined analysis of previously published datasets, highlighting the importance of making genomic data available to the community in a timely manner.

Background

Pediatric high grade gliomas (pHGGs) are aggressive brain tumors occurring at a median age of 6[1]. Sixty percent of pHGGs harbor a histone H3 K27M mutation, which is associated with an aggressive phenotype and dismal survival rates[2]. H3K27M-mutant pHGG tumors are located along the midline, including in the pons, cerebellum, and brainstem. A diffuse phenotype and delicate location leave them unsuitable for surgery, and their pronounced chemoresistance renders the standard treatments for gliomas ineffective, resulting in a median survival time of only 12 months[3,4]. The prognostic significance of the H3 K27M mutation in these gliomas resulted in a new WHO tumor classification, diffuse midline glioma with H3K27M mutation[5].

The H3K27M mutation results in a global decrease in H3K27me3, an epigenetic repressive mark and posttranslational histone modification[6]. Seventy five percent of gene loci lose or have reduced H3K27me3, although a few loci gain the mark as a result of the H3K27M mutation[2,7]. H3K27me3 is deposited predominantly by EZH2, the catalytic subunit of the PRC2 methyltransferase complex. By regulating H3K27me3, EZH2 maintains cell identity and regulates cellular differentiation[8–11]. Silencing EZH2 in neuroepithelial cells before their differentiation alters the distribution of the progeny cell types[12]. EZH2 also maintains neuroepithelial cell integrity, and midbrain identity[13,14].

Because H3K27me3 is globally lost in H3K27M-mutant glioma, the subsequent deregulation of gene expression is thought to lead to tumorigenesis, although the developmental timing of the mutational event is important[15]. H3K27M expression in neural stem cells has led to tumorigenesis in mice when accompanied by *TP53* knockout and/or *PDGFRA* amplification, but this combination of molecular aberrations failed to result in tumorigenesis when introduced in mature astrocytes[16,17]. However, the precise cell type of origin for H3K27M gliomas is not yet known. Candidate cell types include neuroepithelial cells (also known as neural stem cells), radial glia (also known as neural progenitor cells), and oligodendrocyte precursor cells (OPCs)[16–18].

Many important brain developmental processes are regulated by H3K27me3 deposition and could contribute to gliomagenesis if not well controlled. One of these is the epithelial-mesenchymal transition (EMT), which is essential for gastrulation, migration of neural crest cells, and neural tube formation[19–21]. The EMT is regulated by SNAI1, a transcription factor master regulator[22–24]. By regulating EMT, SNAI1 plays a critical role in many developmental processes, including gastrulation and differentiation of embryonic stem cells[25–27]. SNAI1 induces EMT through direct recruitment of PRC2, resulting in H3K27 trimethylation of key epithelial genes such as concurrently upregulating mesenchymal genes[28,29].

In the brain, processes closely resembling EMT are involved in key developmental steps such as the differentiation of neuroepithelial cells to both neuronal and glial cells[30,31]. These processes, which control cell fate and identity in early neural progenitor cell development, are regulated by EZH2[32]. Interestingly, while EMT mainly results in a differentiation event, in some cases EMT causes increased stem cell properties[33–37]. Recent research potentially reconciles these results by introducing the hybrid epithelial/mesenchymal phenotype: the result of a partial EMT in which both epithelial and mesenchymal genes are expressed[38,39]. This process may allow cancer cells to revert to a more stem cell-like phenotype.

Given the regulation of the EMT by H3K27me3 deposition, and the disruption of this deposition by the H3K27M mutation, we sought to investigate the EMT status in pHGGs with and without the H3K27M mutation. We analyzed RNA sequencing data from 78 pHGGs obtained from three different studies. First, we performed differential expression analysis using RNA sequencing (RNA-seq) derived gene expression from bulk tumor samples, and found that H3K27M gliomas differentially express pre-EMT genes[40]. Secondly, we examined previously published cerebral organoid data and observed similarities between pre-EMT neural stem cells and H3K27M gliomas[41]. Finally, we leveraged a recent single cell RNA sequencing dataset to uncover multiple stages of EMT in H3K27M tumor cells[18]. Overall, our results suggest that the H3K27M mutation may cause an arrest in development of a neural stem cell type at an early stage of EMT, indicating a developmental window of opportunity for H3K27M occurrence.

Our study highlights the importance of genomic data sharing for rare diseases, such as pHGGs. By combining RNA sequencing data from multiple previously published studies, we were able to assemble a cohort of 78 pHGG, large enough for the differential expression analysis of pHGGs with and without the H3K27M mutation.

We used this new cohort of previously published data to derive a novel biological model to describe the molecular pathogenesis of the disease.

Analyses

A. Differential expression analysis of pediatric gliomas with and without H3K27M mutation reveals deregulation of genes involved in epithelial-mesenchymal transition.

We obtained RNA-seq data from 33 H3K27M pediatric high grade gliomas (pHGG) and 45 nonK27M pHGG from the Treehouse Childhood Cancer Initiative public cancer compendium [42] (Supplementary Table 1). These data came from several cohorts including the Pacific Pediatric Neuro-Oncology Consortium (PNOC), Dr. Michelle Monje's studies, and The Cancer Genome Atlas[43–48].

Using the *limma* package in R[49], we conducted differential expression analysis between the H3K27M and nonK27M pHGG cohorts. A total of 1905 genes are differentially expressed between the two tumor types (Supplementary Table 2). Using Gene Set Enrichment Analysis (GSEA) and the Molecular Signatures Database (MSigDB)[50], we found 23 biological signaling pathways with significant enrichment in coding genes overexpressed in the H3K27M cohort (Supplementary Table 2). The top 5 most significantly enriched gene pathways included "Hallmark KRAS Signaling Down" (genes repressed by KRAS activation) and the "Hallmark Epithelial Mesenchymal Transition" (Figure 1A). KRAS pathway enrichment is consistent with a recent study which found RAS signaling to be activated in H3K27M gliomas[51].

Because the epithelial-mesenchymal transition (EMT) is regulated by deposition of H3K27me3, an epigenetic transcriptional repressive mark that is lost in H3K27M cells, we were particularly interested in the differential expression of genes involved in the EMT pathway. The Hallmark EMT pathway gene list is limited to 200 genes[52], so to comprehensively characterize differential EMT activity in H3K27M mutant versus nonK27M tumors, we generated a master list of non-redundant EMT-related genes (n=1226) by merging all MSigDB EMT-related gene sets and by identifying EMT-related genes through manual literature curation (Supplementary Table 2). This list includes genes implicated in both epithelial and mesenchymal cell states, as well as several intermediate EMT cell states and EMT-like processes.

To investigate differential EMT gene expression, we calculated the overlap between the EMT master list and the differentially expressed genes (Supplementary Table 2). We found 123 differentially expressed genes from the EMT master list, indicating potential differential activity of the EMT pathway in H3K27M mutant gliomas (pvalue<2.38^{-28}, hypergeometric test). Of these genes, 73 were more highly expressed in H3K27M tumors, and the remaining 50 were more highly expressed in nonK27M tumors. (Figure 1B). Further investigation revealed that, in general, the EMT-related genes overexpressed in the H3K27M cohort are associated with epithelial-like cell states, and are normally upregulated prior to the EMT. In contrast, many of the EMT genes underexpressed in H3K27M tumors are mesenchymal markers or associated with a post-EMT cell state.

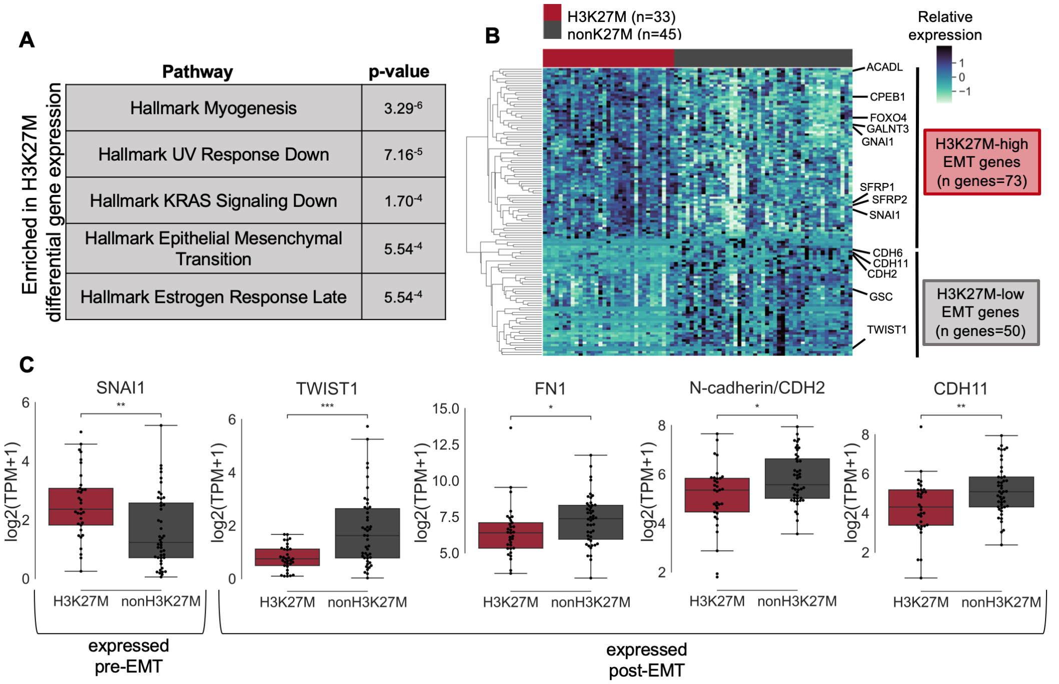
A few examples illustrate this striking trend. *SFRP1* and *SFRP2*, which are more highly expressed in H3K27M tumors, have been shown to inhibit pro-EMT transcription factors and thereby increase expression of E-cadherin in epithelial cells

(*SRFP1* log fold change (LFC)=0.5, *SFRP2* LFC=0.8)[53]. *GALNT3*, which has been characterized as one of the best expression markers for epithelial cells, has higher expression in H3K27M tumors (LFC=0.6)[54]. In contrast, *GSC*/Goosecoid is a key marker of mesenchymal cells, and displays lower expression in H3K27M tumors compared to nonK27M tumors (LFC=-3.1)[55,56].

In particular, we noted that *SNAI1*, a transcription factor and key regulator of the EMT, is significantly overexpressed in H3K27M tumors (LFC=0.6; Figure 1C). High expression of *SNAI1* is a marker of EMT induction in epithelial cells. If the EMT is successful, this is followed by high expression of mesenchymal markers *TWIST1[57]*, fibronectin (*FN1*)[58], N-cadherin (*CDH2*)[59] and cadherin-11 (*CDH11*)[60]. Using a Mann-Whitney nonparametric significance test, we found significantly reduced expression of all of these mesenchymal markers in H3K27M tumors (*TWIST1* LFC=-1.2, *FN1* LFC=-0.2, *CDH2* LFC=-0.2, *CDH11* LFC=-0.3; Figure 1C). *TWIST1*, *CDH2* and *CDH11* are also underexpressed in the H3K27M cohort by the *limma* analysis.

Because *SNAI1* induces EMT by directly recruiting PRC2 methyltransferase activity for H3K27-trimethylation, a process blocked by the H3K27M mutation, we hypothesized that the occurrence of the H3K27M mutation may promote tumorigenesis by stalling EMT during early neuroepithelial differentiation. To further investigate this hypothesis, we performed comparative RNA-sequencing expression outlier analysis developed by the Treehouse Childhood Cancer Initiative, which identifies genes with outlier expression in individual samples as compared to a background cohort of highly correlated and disease-matched samples (pan-disease analysis, see Methods) [40]. We identified genes with outlier expression only in nonK27M pHGG samples (but not H3K27M pHGG samples) as compared to a background glioma cohort, and noted that many mesenchymal and post-EMT pathways were identified as enriched among the outlier genes (Supplementary Figure 1, Supplemental Table 1).

Overall, our multiple analyses of the pHGG RNA-seq cohort suggest that H3K27M pHGG tumors are associated with pre-EMT gene expression, while nonK27M pHGG tumors are characterized by post-EMT and mesenchymal gene expression.



Pathway	p-value
Hallmark Myogenesis	3.29^{-6}
Hallmark UV Response Down	7.16^{-5}
Hallmark KRAS Signaling Down	1.70^{-4}
Hallmark Epithelial Mesenchymal Transition	5.54^{-4}
Hallmark Estrogen Response Late	5.54^{-4}

Figure 1. The EMT pathway is differentially expressed in H3K27M gliomas as compared to nonK27M gliomas. A) Differential expression analysis of a cohort of H3K27M and nonK27M pHGG revealed significant enrichment of Hallmark Epithelial Mesenchymal

Transition in genes overexpressed in H3K27M gliomas. B) Heatmap of differentially expressed EMT genes between H3K27M and nonK27M pHGG. C) SNAI1, master regulator of EMT, is overexpressed in H3K27M glioma, while mesenchymal markers TWIST1, FN1, CDH2 and CDH11 are underexpressed in H3K27M glioma as compared to nonK27M gliomas (Mann-Whitney significance test; * pvalue < 0.05, ** pvalue < 0.01, *** pvalue < 0.001).

B. H3K27M-mediated gliomagenesis is associated with pre-EMT cell types.

Consistent with our differential expression analysis, a review of the literature revealed that H3K27M-associated gliomagenesis has been experimentally recapitulated only in cell types which are poised to undergo an EMT differentiation event (Figure 2A). For example, a combination of H3K27M, *p53* loss, and *PDGFRA* constitutive activation in human neural progenitor cells (NPCs) induced low grade gliomas when injected into the pons of neonatal mice[16]. These gliomas expressed markers of pre-EMT neuroepithelial cells. Another study found that H3K27M and *Trp53* loss was sufficient for gliomagenesis in the NPCs of embryonic mice in the forebrain and hindbrain[17]. Strikingly, when introduced post-natally, H3K27M and *p53* loss in NPCs was not sufficient for gliomagenesis, although post-natal induction of H3K27M, *Trp53* loss and *PDGFRA* amplification in neural stem cells resulted in glioma formation[61,62]. Additionally, no tumorigenesis was observed upon introduction of H3K27M, *p53* loss and *PDGFRA* constitutive activation in mature astrocytes, a post-

EMT cell type[16]. These observations indicate that experimental H3K27M-mediated gliomagenesis occurs in a pre-EMT cell type.

Based on our gene expression analysis and review of the literature, we hypothesized that H3K27M gliomas arise in pre-EMT cell types and retain the EMT signature of the cell type in which the mutation arises. Given this hypothesis, we expect that H3K27M gliomas harbor gene expression signatures of normal pre-EMT cell types that exist during neuronal development. In order to compare the expression of the EMT-related genes of interest between H3K27M tumors and normal developing brain cells, we examined total and single cell RNA-seq data from a human embryonic stem cell-derived cerebral cortex organoid time course experiment (Figure 2B)[41]. These organoid cultures mimic the early weeks of human prenatal cortical development and generate relevant cell types, uniquely allowing us to investigate early time-points in development which are not available in existing human fetal brain datasets. After induction of neural epithelium by week 1, at week 2 radial glia cells and Cajal-Retzius neurons are present in addition to some remaining neuroepithelial cells. By week 5, the organoids contain populations of radial glia, intermediate progenitors and deep-layer neurons.

When we investigated EMT-related gene expression in cerebral organoids during gestational weeks 1-6, we noted the presence of 2 distinct EMT processes (Figure 2A, lower panel). The first process starts as *SNAI1* expression peaks in neural stem cells (week 1), coincident with low expression of mesenchymal markers *TWIST1*, *CDH2, CDH11* and *FN1*. As differentiation from neural epithelial cells to early radial glia occurs, *SNAI1* expression decreases while mesenchymal marker expression increases. In the second process, as radial glia cells prepare to undergo a second EMT into intermediate progenitor cells, *SNAI1* expression increases once again.

To further characterize the EMT states represented in cerebral organoids, we utilized single cell RNA-seq data from the cerebral organoids at gestational weeks 3 and 6[41]. These sample collection times effectively covered all relevant cell type diversity, as gestation week 3 organoids contain substantial populations of neural epithelial cells, early radial glia cells and Cajal-Retzius neurons, while week 6 organoids are composed of late radial glia cells, intermediate progenitors, and immature neurons. We scored the EMT status of each cell using a gene signature representing EMT completion (Figure 2C, Supplementary Table 3)[18,63–66]. Neural epithelial and early (presumably pre-EMT) radial glia cells show significantly lower EMT scores than post-EMT intermediate progenitors, late radial glia and neurons (Mann-Whitney test, pvalue<0.0001). This shows that our assay contains distinct populations of pre- and post-EMT cerebral cells, and is consistent with the levels of *SNAI1*, *CDH2*, *CDH11, FN1* and *TWIST1* in the bulk weeks 1-6 organoid data. This dataset enables us to investigate transcriptional similarities between H3K27M-mutant gliomas and normal pre-EMT cell types during neural development.

We then examined the expression of genes overexpressed in H3K27M gliomas in the single cell organoid RNA-seq dataset, to see which normal cell type is most similar to H3K27M glioma cells. Of the 1180 H3K27M-overexpressed genes, 152 genes passed the single cell RNA-seq expression filter (Supplementary Table 3, see Methods). Hierarchical clustering of the expression profiles of these genes in normal cell types during neural development revealed highest expression in pre-EMT neural epithelium and early radial glia (Figure 2D). We then ranked this gene signature based on each gene's expression in each cell type (see Methods). We found that this signature is ranked most highly in pre-EMT neural epithelium and in early radial glia (pvalue<0.05, Figure 2E).

Overall, these results suggest that the differential EMT gene expression observed in our tumor cohort is related to stages of EMT in the normal developing brain, and that H3K27M tumor cells resemble pre-EMT neural cell types.

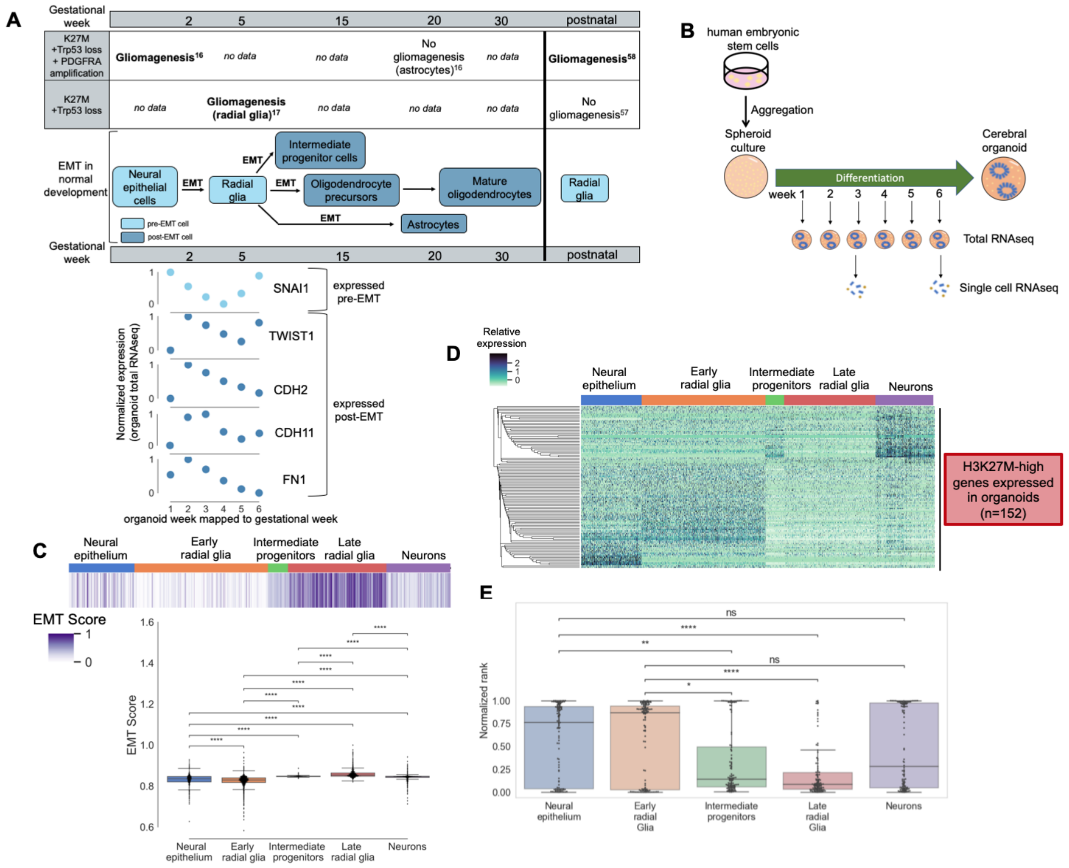


Figure 2. H3K27M-specific EMT transcriptional signature is similar to pre-EMT neural stem cell expression in cerebral organoids. A) In vitro and in vivo experimental H3K27M-associated gliomagenesis occurs exclusively in pre-EMT cell types (upper panel). These

cell types are represented in our cerebral organoid assay, and a time course of these organoid cultures represents 2 EMT events in early brain development (lower panel). B) Experimental workflow for total RNA-seq and single cell RNA-seq from a human embryonic stem cell derived cerebral cortex organoid time course experiment. C) Single cells from cerebral organoids were scored for EMT completion. Pre-EMT neural epithelium and early radial glia were least enriched for the EMT score, while post-EMT intermediate progenitors, late radial glia and neurons were the most enriched. D) A signature of genes differentially expressed in H3K27M gliomas and expressed in cerebral organoids shows highest expression in pre-EMT neural epithelium and early radial glia. E) EMT-related genes highly expressed in H3K27M-mutant gliomas are also highly expressed in neural epithelium and early radial glia. (Mann-Whitney significance test; * pvalue < 0.05, ** pvalue < 0.01, **** pvalue < 0.0001)

C. Single-cell profiling of H3K27M gliomas reveals groups of cells at different stages of EMT.

We utilized recently published single cell RNA-seq data from 6 H3K27M and 2 H3 wild type (H3WT) gliomas to directly investigate the EMT signatures of single cell populations within each tumor type[18]. One of the H3K27M tumors harbors the mutation in the *HIST1H3B* gene (referenced as H3.1K27M), while the remaining 5 H3K27M tumors harbor the mutation in the *H3F3A* gene (referenced as H3.3K27M).

We performed hierarchical clustering of 3057 tumor cells using 629 genes from the EMT master list which passed expression filters (see Methods, Supplementary Table 4)[67]. Ten EMT-related clusters were discovered and named A-J (Figure 3A, Supplementary Table 4). Cluster gene signatures were identified by assigning each cluster the genes with maximum mean expression in that cell cluster across the dataset (Supplementary Table 4).

We assigned cluster function based on manual review of genes in each signature, and observed several populations of cells whose presence in this dataset has already been noted[18]. Cluster C has highest expression of cell cycle markers including *E2F2* and *MCM2-7*, indicating that these are actively cycling cells[68]. Cluster E is composed predominantly of non-malignant immune cells, indicated by comparatively highest expression of immune markers such as *CD68[69]*. Cluster I resembles oligodendrocytic cells, with highest expression of *CD9* and *ZEB2*, and cluster J resembles oligodendrocyte precursor cells with the highest expression of *PDGFRA[70–73]*. The presence of each of these cell types has already been noted in H3K27M gliomas, and these cell type signatures are not informative for assessing EMT state[18].

However, the remaining clusters are defined by gene expression representing various stages of EMT. We again scored the EMT status of each cell with a gene signature representing EMT completion (Figure 3A, see Methods)[18,63–66]. Cluster A scored the lowest overall, while clusters F, G, and H scored the highest overall. Cluster relationships are shown with Uniform Manifold Approximation and Projection (UMAP) in Figure 3B, and expression patterns of selected EMT marker genes are shown in the lower panel of Figure 3B. Of the EMT marker genes identified in the bulk RNA sequencing analysis (Figure 1C), only *FN1*, *CDH2* and *CDH11* were expressed in the glioma single cell RNA-seq data, so we also visualized *VIM* as a post-EMT marker and *OCLN* as a pre-EMT marker.

In keeping with our previous analysis, we noted that clusters F and G, which are composed mainly of H3WT glioma cells, strongly resemble post-EMT cells and most highly express canonical mesenchymal markers including *CDH2*, *CDH11, FN1* and *VIM[74,75]*. This is consistent with our observation that nonK27M gliomas transcriptionally resemble a post-EMT state as compared to H3K27M in the bulk RNA-seq pHGG cohort. Thus, we defined Clusters F and G “post-EMT”.

Interestingly, within the clusters composed predominantly of H3K27M cells, multiple stages of EMT emerged. Cluster A cells exhibit comparatively highest expression of several genes known to be active in epithelial or pre-EMT cell types, including *CADM1*, *EGR1, PTEN, NOTCH1,* and *OCLN[76–80]*. Additionally, cluster A cells are characterized by high expression of genes activated at the early stages of the *SMAD3*-induced EMT pathway, including *SMAD3*, *CTNNB1*, *FOS*, and *FOSB[79,81,82]*. Therefore, we defined Cluster A “pre-EMT”. In contrast, Cluster B has comparatively highest expression of only 6 genes (*ACTG1, BMP2, COPA, PLXNA2, RPS27A and TP53INP1*) and has no clear expression signature of any stage of EMT, so we defined Cluster B “EMT-ambiguous”.

Clusters D and H were defined “EMT-intermediate”, because both clusters display high expression of genes normally expressed while the EMT process is taking place, without a clear bias towards epithelial or mesenchymal gene expression. For example, cluster D has the highest expression of *MMP2*, *VCAN*, and *SMAD2*, which are activated during the EMT process rather than before or after[79,83]. Cluster H cells display both pro-EMT and anti-EMT signaling, as evidenced by expression of genes involved in activating EMT (*TNC*, *MMP14*, and *FGFR3*), and genes implicated in suppressing EMT (*DLG5*, *LRIG1*, and *WWC1*)[84–89]. Cluster H also has the highest expression of several genes previously identified as characterizing an intermediate epithelial/mesenchymal (E/M) state (*COL6A1*, *NR2F1*, *TFPI*, *WNT5A*)[39].

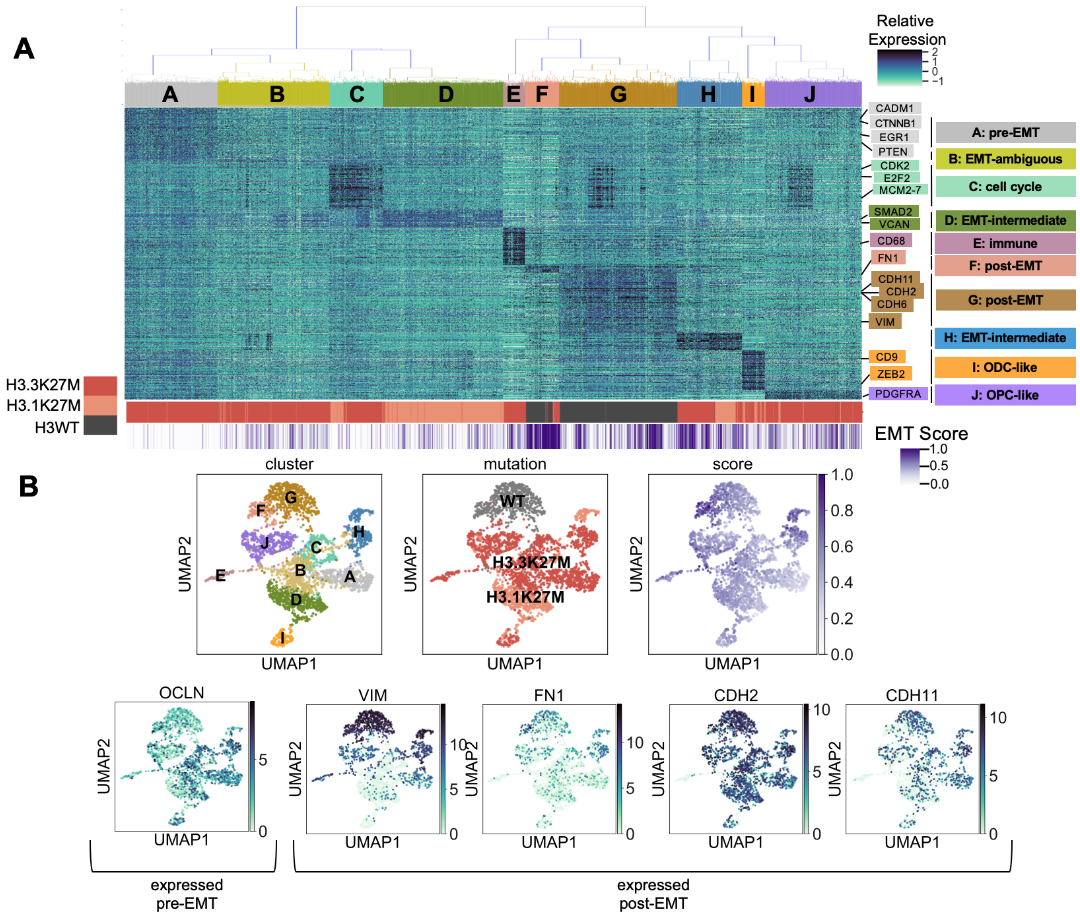


Figure 3. Single cell RNA sequencing of H3K27M and nonK27M gliomas reveals multiple EMT stages within tumors. A) Expression heatmap showing hierarchical clustering of 3,057 cells from 6 H3K27M and 2 nonK27M high-grade gliomas, with a master list of EMT genes. Ten clusters (A-J) were assigned gene signatures based on maximum mean gene expression in each cluster, and clusters were classified based on manual review of each gene signature. Histone H3 mutation status and EMT score are shown at the bottom

of the heatmap (ODC=oligodendrocyte, OPC=oligodendrocyte precursor). B) UMAP dimensionality reduction projection of the same expression data as the heatmap and labeled by cluster, Histone H3 mutation status and EMT score. Expression of selected epithelial and mesenchymal genes shown in bottom panel.

D. Histone H3.1K27M glioma cells may represent a more advanced stage of EMT than H3.3K27M glioma cells.

Further examination revealed that cluster D mainly consists of cells from the H3.1K27M mutant tumor, and cluster H consists of a mixture of H3.1 and H3.3K27M cells. H3.1 and H3.3K27M characterize two functionally different subtypes of H3K27M gliomas; H3.1K27M gliomas are comparatively rare but have a slightly better prognosis[46,90]. The H3.1 histone is diffusely distributed throughout the genome, while the H3.3 histone is preferentially located at active chromatin[91–93]. This leads to distinct patterns of epigenetic reprogramming in each histone variant, where loss of the H3.3K27me3 mark is directly correlated with areas of H3.3 genomic enrichment, but H3.1K27me3 loss is not localized[93]. Because the H3K27M mutation is known to induce dose-dependent inhibition of PRC2 methyltransferase, this suggests that the localized distribution of H3.3 histone may result in higher local inhibition of PRC2 and loss of H3K27me3 at H3.3K27M sites, whereas the widespread distribution of H3.1K27M results in diffuse PRC2 inhibition[61,93]. Because precise control of gene transcription via active chromatin is necessary for a successful EMT, a H3.3K27M mutation would be particularly damaging to proper regulation of the EMT pathway. Indeed, functional analysis of enhancer regions in H3.3K27M-expressing NPCs revealed enrichment of regions positively regulating EMT, indicating that H3.3 active chromatin regions are directly involved in transcriptional control of EMT genes[93]. This suggests that EMT-poised H3.3K27M cells will be unable to properly complete EMT due to lack of transcriptional control.

Accordingly, we observed EMT-intermediate or E/M hybrid expression genes in glioma single-cell clusters D and H, both of which have substantial numbers of H3.1K27M glioma cells. We hypothesized that H3.1K27M cells may be more differentiated and farther along the EMT process than H3.3K27M cells.

In order to investigate this hypothesis further, we subset the single cell glioma RNA-seq data to 2458 cells with H3.1K27M or H3.3K27M mutation and performed Wilcoxon rank-sum test to identify genes overexpressed in each variant group (Supplementary Table 4; Supplementary Figure 2). Consistent with our previous observations, GSEA of Gene Ontology (GO) gene sets (Figure 4B, Supplementary Table 4) revealed enrichment of epithelial gene sets in H3.3K27M compared to H3.1K27M (GO Adhesion pathways, GO Neurogenesis, GO Embryo Development) and mesenchymal gene sets in H3.1K27M compared to H3.3K27M (GO EMT pathway, GO Mesenchymal Cell Differentiation and GO Mesenchyme Development). Additionally, scoring of all cells for EMT completeness shows that H3.1K27M cells score significantly higher overall than H3.3K27M cells, while nonK27M cells score significantly higher than either mutant cell type (Supplementary Figure 3). However, because the H3.1K27M cells come from a single tumor, we performed additional analysis to investigate this observation.

We cultured diffuse intrinsic pontine glioma (DIPG) primary cell lines isolated in a previous study to investigate the expression of EMT markers in H3.3K27M, H3.1K27M and nonK27M glioma cells[94]. Morphologically, we observed that when cultured in serum-free conditions, the H3.1K27M cell lines preferentially grow attached to the flask (4 of 5 cell lines), while the H3.3K27M cells preferentially grow as neurospheres (8 of 9 cell lines) (Figure 4C). Because differentiation of neurospheres is accompanied by attachment and increased expression of N-cadherin, this morphological trend is consistent with our hypothesis that H3.1K27M cells exist in a more differentiated state than H3.3K27M cells[95].

We analyzed RNA-seq data from 3 DIPG cell lines to compare the expression of EMT genes (SU-DIPG-IV is H3.1K27M mutant; SU-DIPG-VI and JHH-DIPG1 are H3.3K27M mutant). We used 4 replicate samples from each SU-DIPG-IV and SU-DIPG-VI and 3 replicate samples from JHH-DIPG1. Each sample was scored using a gene signature of EMT completion (see Methods), and the H3.1K27M samples scored significantly higher than the H3.3K27M samples (Figure 4D, pvalue<0.05).

We then performed RT-PCR to quantify expression of *FN1* and *CDH2*, canonical mesenchymal marker genes which were previously identified in the bulk glioma RNA sequencing analysis (Figure 4E, full-length gel in Supplementary Figure 4). We attempted to quantify E-cadherin/*CDH1* as it is a canonical epithelial marker, but the levels were so low as to be undetectable by RT-PCR in these cell lines (RNA-seq <1.0 log2(TPM+1)). We compared 9 H3.3K27M cell lines (SU-DIPG-VI, XIII, XVII, XIX, 24, 25, 27, 35 and 43) with 5 H3.1K27M cell lines (SU-DIPG-IV, XXI, 33, 36 and 38) and included 5 H3 wild-type lines (SU-DIPG-48, pcGBM2R, KNS42, SJG2 and normal human astrocytes hTERT) and a negative RT-PCR control (NC). Overall, the H3 wild-type and H3.1K27M cell lines appear to more highly express both mesenchymal markers, in keeping with the bulk and single-cell RNA-seq analyses.

Our computational and *in vitro* observations are consistent with a recent study indicating that H3.1K27M tumor cells are overall more differentiated than H3.3K27M tumor cells[93].

Overall, these data suggest that the histone H3K27M mutation is associated with a preferentially early or pre-EMT cell state as compared to nonK27M cells, but that H3.1K27M cells may represent a somewhat later or intermediate-EMT cell state as compared to H3.3K27M cells.

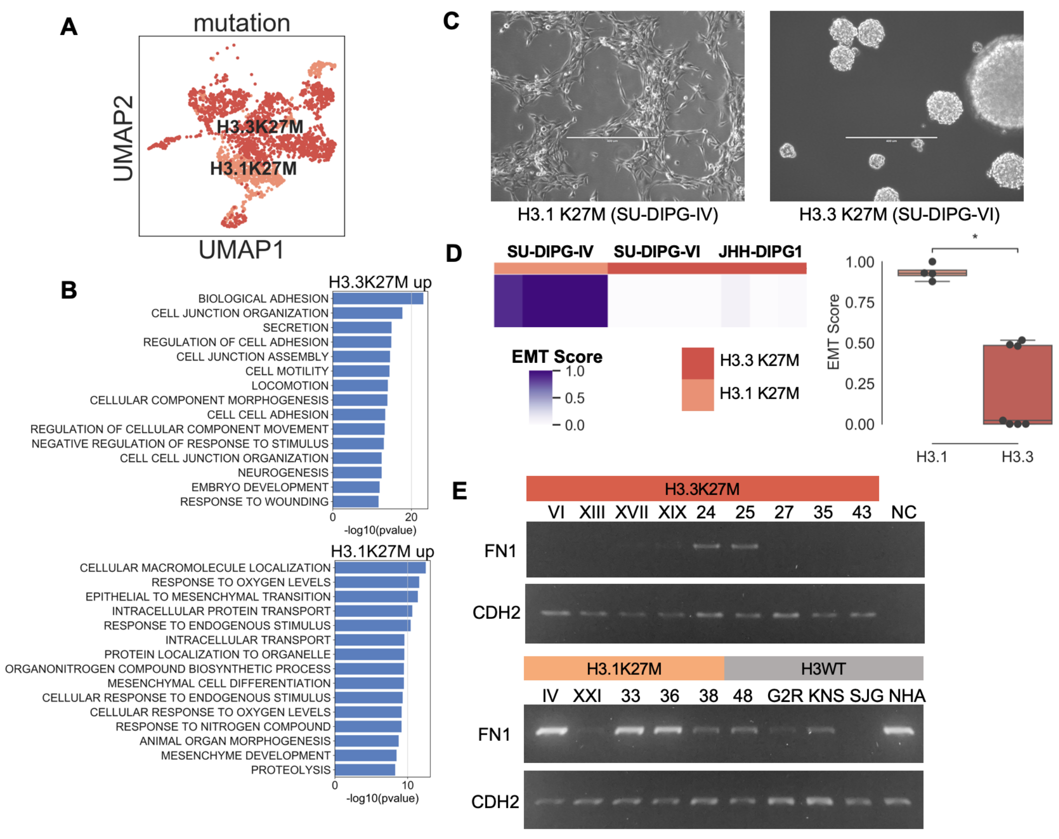


Figure 4. H3.1K27M glioma cells appear more mesenchymal than H3.3K27M glioma cells. A) UMAP dimensionality reduction of 2458 histone mutant glioma single cells. B) Gene set enrichment analysis of genes overexpressed in H3.3K27M versus H3.1K27M

(top graph) or H3.1K27M versus H3.3K27M (lower graph) by Wilcoxon rank-sum test using glioma single cell RNA-seq data. C) Representative images of H3.1K27M and H3.3K27M glioma derived cell cultures. Scale bar 400 um. D) Total RNA sequencing datasets from glioma cell lines were scored for EMT completeness (4 samples from SU-DIPG-IV, 4 samples from SU-DIPG-VI and 3 samples from JHH-DIPG1). Scoring is shown in a heatmap and a boxplot. (Mann-Whitney significance test; * pvalue < 0.05) E) RT-PCR of FN1 and CDH2 expression in glioma primary cell cultures (all numbered lines are SU-DIPG).

Discussion

H3K27M diffuse midline gliomas are aggressive tumors generally occurring in early childhood in the hindbrain or midline. These tumors have poor prognosis and do not respond to standard chemotherapies for adult gliomas[96]. Unlike most adult cancers, pediatric cancers, including pediatric gliomas, are thought to have a developmental origin[15,46,97]. The temporal- and region-specific occurrence of pediatric diffuse midline gliomas reinforces this possible developmental origin. EZH2-deposited H3K27me3 transcriptional marks are known to have crucial roles in cell differentiation and development in the brain and are lost in H3K27M cells[6].

Research has implicated the epithelial mesenchymal transition in pediatric gliomas[98,99], particularly those with a more invasive phenotype. A large portion of diffuse midline glioma tumors highly express genes known to be involved in EMT occurring in adult glioblastomas[100]. EZH2 appears to play an important role in EMT in adult gliomas: EZH2 depletion in adult glioblastomas leads to a reduction in expression of mesenchymal markers, and an increase in epithelial markers[6]. Other studies suggest EZH2 is important for the invasion of gliomas[101–103]. Thus, a molecular aberration affecting the activity of EZH2 might prevent a complete epithelial-mesenchymal transition.

In this study, we observed that various canonical EMT-inducing genes are significantly overexpressed in H3K27M mutant pHGGs, compared to nonK27M pHGGs, while many canonical mesenchymal markers are underexpressed in H3K27M pHGGs as compared to the nonK27M tumors. In particular, we noted higher expression of the pro-EMT transcription factor *SNAI1* in H3K27M-mutant gliomas. Because *SNAI1* relies on PRC2 and H3K27me3 to facilitate EMT through gene expression regulation, this may indicate an arrest in the EMT process. The existence of a hybrid epithelial/mesenchymal phenotype is well-established: the result of a partial EMT is the expression of both epithelial and mesenchymal genes[38]. Studies have shown that a hybrid E/M phenotype may indicate a worse prognosis than mesenchymal-only states in solid tumors[38,39,104].

We hypothesized that if H3K27M mutation prevents full EMT, neural stem cells harboring H3K27M may be forced to retain a proliferative, stem cell phenotype, eventually leading to tumorigenic development. Accordingly, we observed from extensive literature review that experimental induction of H3K27M-associated gliomas has occurred exclusively in pre-EMT cell types, and that two consecutive EMT processes occur early in normal brain development.

Single cell RNA-seq from H3K27M and nonK27M tumors confirmed a more mesenchymal expression signature in the nonK27M cells, and also revealed subsets of H3K27M cells at various stages of EMT. Specifically, we observed an intermediate EMT signature in the H3.1K27M cells as compared to the more epithelial H3.3K27M cells. This was also observed in bulk RNA-seq and *in vitro* analysis. We hypothesize that because the H3.1K27M mutation is not concentrated at active chromatin, it has less repressive power as specific developmental pathways such as EMT are activated over time. If a subset of H3.1K27M cells are able to differentiate through the EMT, this may explain why H3.1K27M gliomas have a slightly better prognosis.

To conclude, we mined 3 publicly available RNA-seq datasets from pediatric gliomas and cerebral organoids to generate a hypothesis for the gliomagenesis of H3K27M gliomas. We propose that the H3K27M mutation is tumorigenic when the mutational hit occurs in a cell poised to undergo the EMT, due to the dependence of normal EMT on the correct timing of the H3K27me3 mark (Figure 5). More work is needed to characterize the observed difference in the EMT status between the H3.1 and H3.3K27M variants. These results hold important implications for better understanding the developmental origin and timing of these aggressive and untreatable cancers. Further, the presence of an epigenetically-driven differentiation stall may imply that a pharmacological methylation agent or a pro-differentiation therapy may aid in future treatment of H3K27M mutant tumors[105].

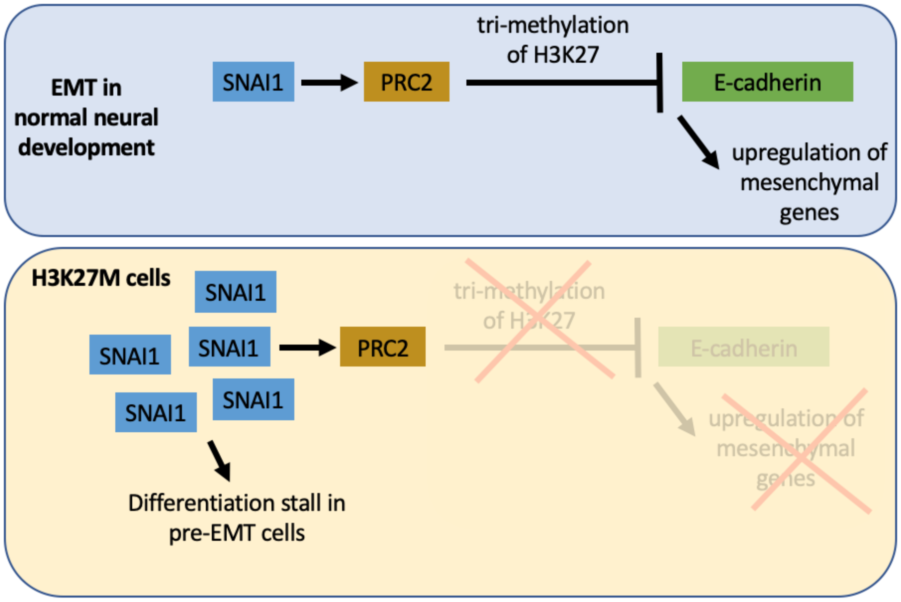

EMT in normal neural development
SNAI1
PRC2
tri-methylation of H3K27
E-cadherin
upregulation of mesenchymal genes
H3K27M cells
SNAI1
SNAI1
SNAI1
SNAI1
SNAI1
PRC2
tri-methylation of H3K27
E-cadherin
upregulation of mesenchymal genes
Differentiation stall in pre-EMT cells

Figure 5. Proposed model for EMT stall in H3K27M cells. We propose that H3K27M cells retain high levels of SNAI1 expression but remain stalled in a pre-EMT state due to inability of PCR2 to tri-methylate H3K27.

Potential Implications

Our study holds implications for other diseases, because H3K27M mutation is not exclusive to diffuse midline gliomas. It can also be found in a fraction of pediatric ependymomas and medulloblastomas[106]. Interestingly, ependymomas located in the posterior fossa typically do not harbor the H3K27M mutation, but exhibit the K27M-associated H3K27 hypomethylation phenotype. Thus, the proposed EMT arrest and differentiation stall as a result of H3K27me3 loss may also apply to these cancers. Beyond the SNAI1-H3K27me3 axis, EMT is also regulated by other epigenetic marks[107]. Given the epigenetically dysfunctional nature of many pediatric cancers[15], EMT arrest could conceivably play a role in the oncogenesis of these tumors as well.

Methods

Glioma bulk RNA sequencing data

Gene expression data from 78 pediatric high grade glioma samples were downloaded from the Treehouse Childhood Cancer Initiative public compendium v8[42]. All samples in the compendium have been uniformly processed using the UC Santa Cruz TOIL RNA-seq pipeline (v3.3.4)[108]. This dataset (n=58581 genes) is in transcripts per million (TPM) and normalized by log2(TPM+1). We divided the dataset into 33 H3K27M mutant samples and 45 nonK27M samples, and performed differential expression analysis of all genes between the two groups using R library *limma* v3.34.9 in R v3.3.4. We performed gene set enrichment analysis (GSEA) of the resulting 1905 differentially expressed genes (pvalue<0.1) with Molecular Signatures Database (MSigDB) v7.0 on the GSEA/MSigDB web site v6.4 (Supplementary Table 2). Since the epithelial-mesenchymal transition (EMT) pathway was in the top 5 most significantly enriched pathways in H3K27M over expressed genes, we created a non-redundant master list of EMT genes (n=1226) by merging 15 EMT related MSigDB pathways and by identifying EMT-related genes through manual literature curation (Supplementary Table 2).

We performed pan-disease outlier analysis on all the pHGG samples using Treehouse CARE (see Availability of source code and requirements section) against the Treehouse Cancer Compendium v10. Pan-disease outlier analysis identifies genes with outlier expression in each sample of interest as compared to a background cohort of tumors identified as most similar[40]. We identified a list of genes with outlier expression in the nonK27M pHGG samples that did not also have outlier expression in the H3K27M pHGG samples, and performed gene set enrichment analysis using Enrichr in the GSEApy package (gseapy-v0.9.17)[109] against BioPlanet_2019 library with p-value cutoff 0.05 (outlier genes and enriched pathways in Supplementary Table 2). We used the EnrichmentMap app in Cytoscape to visualize functionally similar clusters of enriched pathways[110].

Cerebral organoid RNA sequencing data (bulk and single cell)

Gene expression data (TPM) from 6 weekly timepoints of human cerebral organoid growth were downloaded from accession GSE106245[41]. Organoid weeks 0-5 were converted to gestational weeks 1-6 and duplicate gene measurements were averaged. For Figure 2A, expression of each gene was normalized between 0-1. Single cell RNA sequencing data from weeks 2 and 5 (gestational weeks 3 and 6) cerebral organoids were downloaded from accession GSE106245[41]. Expression data were filtered to remove genes with expression in fewer than 10% of cells. Cell types were assigned using a list of marker genes (Supplementary Table 3).

Glioma single cell RNA sequencing data

Smart-seq2 RSEM TPM single cell RNA sequencing data from 3,057 glioma cells were downloaded from accession GSE102130[18]. Data were log2-normalized and filtered to remove genes with expression in fewer than 20% of cells. Hierarchical clustering of all cells was performed using the Python *scipy.cluster.hierarchy* function (scipy v1.4.1) after subsetting to a non-redundant master list of EMT genes (n=1226, Supplementary Table 2). Of these genes, 629 passed the expression filter and were included in the hierarchical clustering. The clustering results were plotted using the *scipy.cluster.hierarchy.dendrogram* function with threshold set to 3.5. Gene signatures for each cluster were assigned by identifying the cluster in which each gene has maximum mean expression, and assigning that gene to that cluster. For UMAP visualizations, Leiden clustering was performed on the single cell data using the *scanpy.tl.leiden* function (scanpy v1.4.5.post1) with resolution set to 0.5 and top 10 principle components used as input.

DIPG Cell Lines

The patient-derived DIPG cell lines (SU-DIPG-IV, SU-DIPG-VI, SU-DIPG-XIII, SU-DIPG-XVII, SU-DIPG-XIX, SU-DIPG-XXI, SU-DIPG-24, SU-DIPG-25, SU-DIPG-27, SU-DIPG-33, SU-DIPG-35, SU-DIPG-36, SU-DIPG-38, SU-DIPG-48) were kindly provided by Dr. Michelle Monje (Stanford University School of Medicine, Stanford CA)[44]. SU-DIPG-IV, SU-DIPG-XXI, SU-DIPG-33, SU-DIPG-36, and SU-DIPG-38 cells harbor a H3.1K27M mutation while SU-DIPG-VI, SU-DIPG-XIII, SU-DIPG-XVII, SU-DIPG-XIX, SU-DIPG-24, SU-DIPG-25, SU-DIPG-27, SU-DIPG-35, SU-DIPG-43 cells harbor a H3.3K27M mutation. SU-DIPG-48 and Glioblastoma cell line SU-pcGBM-2 are H3WT. Glioblastoma H3WT cell lines; KNS-42 (RRID:CVCL_0378), SJ-GBM2 (RRID:CVCL_M141), and one normal astrocyte cell line NHA hTERT were kindly provided by Prof. Sameer Agnihotri (UPMC Children's Hospital of Pittsburgh, Pittsburgh PA). The Universal Mycoplasma Detection Kit (AACC) was used for testing SU-DIPG-XIII, XVII, XIX, and VI latest on January 10, 2020. All cells were cultured in tumor stem medium containing 50X B-27 Supplement Minus Vitamin A (Invitrogen), H-EGF at 20ng/mL (Shenandoah Biotechnology), H-FGF-basic-154 at 20ng/mL (Shenandoah Biotechnology), H-PDGF-AA at 10ng/mL (Shenandoah Biotechnology), H-PDGF-BB at 10ng/mL (Shenandoah Biotechnology), and 0.2% Heparin Solution at 2ug/mL (STEMCELL Technologies). All experiments used cells collected within 5 passages after thawing. The cells were passaged by the treatment of TrypLE (Gibco) and DNAse I (Worthington) rocking at 37°C for 5-15 minutes then HBSS (Corning) was added to deactivate TrypLE. The cells were transferred to new Nunc EasYFlask Cell Culture Flasks (ThermoFisher Scientific) and grown in tumor stem medium as previously described. The bulk RNA sequencing data from lines SU-DIPG-VI, SU-DIPG-IV and JHH-DIPG1 were obtained with permission from Dr. Michelle Monje from dbGap accession phs000900.v1.p1.

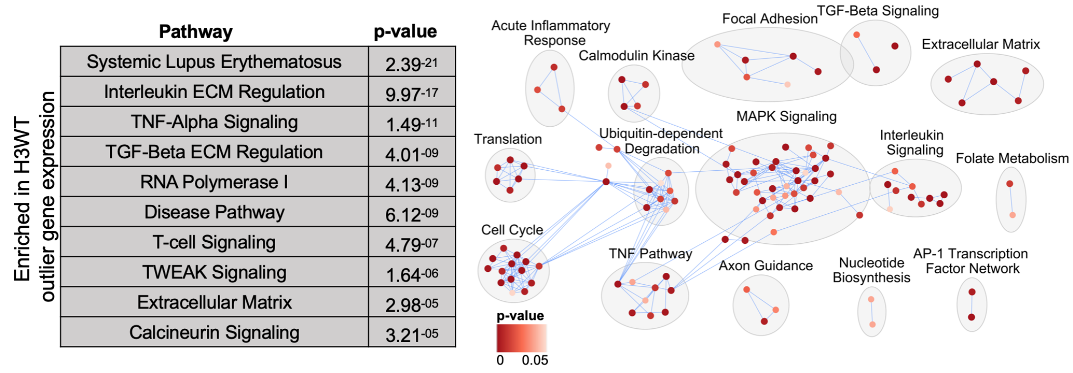
RNA Extraction and RT-PCR

Total RNA was extracted from cell pellets using the Quick-RNA Miniprep Kit (Zymo Research). cDNA was synthesized from 1 ug of total RNA using Oligo(dT)20 primers and the SuperScript III First Strand Synthesis System (Invitrogen). PCR was performed using KAPA HiFi HotStart ReadyMixPCR Kit (KAPA Biosystems), 50 ng of template DNA and the appropriate primers and 27 PCR cycles. *CDH2* primer sequences: forward: ggcttaatggtgattttgctcag reverse: tccataccacaaacatcagcac. *FN1* primer sequences: forward: cttgaaccaacctacggatgac reverse: tcccatcatcataacacgttgc. Primer oligos were purchased from Integrated DNA Technologies.

Data Analysis

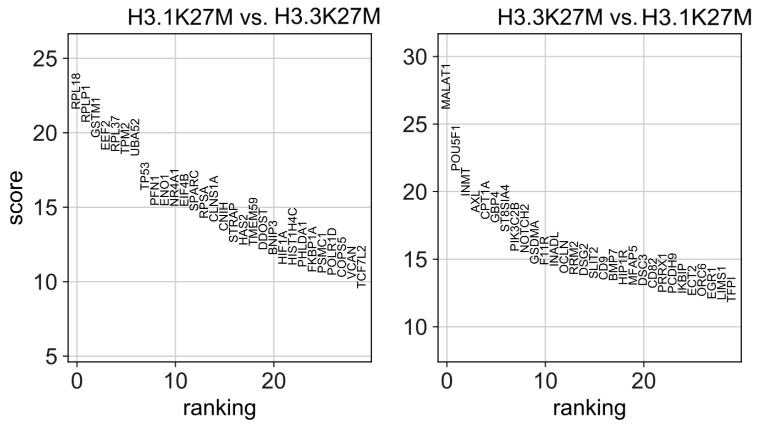
All statistical comparisons are performed with a two-sided Mann-Whitney test, with measurements taken from distinct samples without assumption of normality. Single cell and bulk tumor samples were scored for EMT activity using a manually curated set of mesenchymal genes and a scoring method based on aggregate expression of the gene set as compared to a control gene set (Supplementary Table 3)[18,64,111].

Supplementary Figures

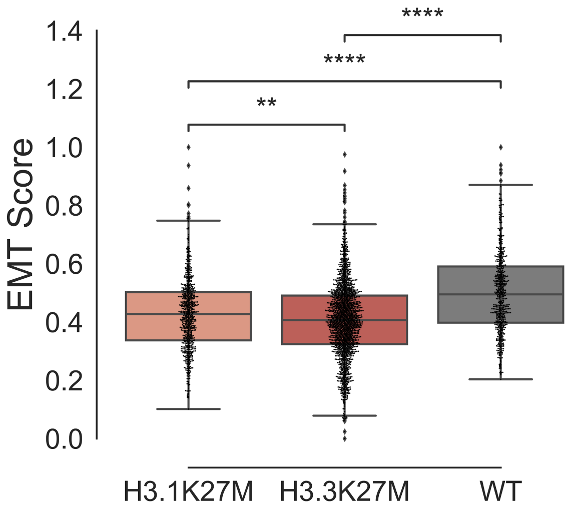


Pathway	p-value
Systemic Lupus Erythematosus	2.39^{-21}
Interleukin ECM Regulation	9.97^{-17}
TNF-Alpha Signaling	1.49^{-11}
TGF-Beta ECM Regulation	4.01^{-09}
RNA Polymerase I	4.13^{-09}
Disease Pathway	6.12^{-09}
T-cell Signaling	4.79^{-07}
TWEAK Signaling	1.64^{-06}
Extracellular Matrix	2.98^{-05}
Calcineurin Signaling	3.21^{-05}

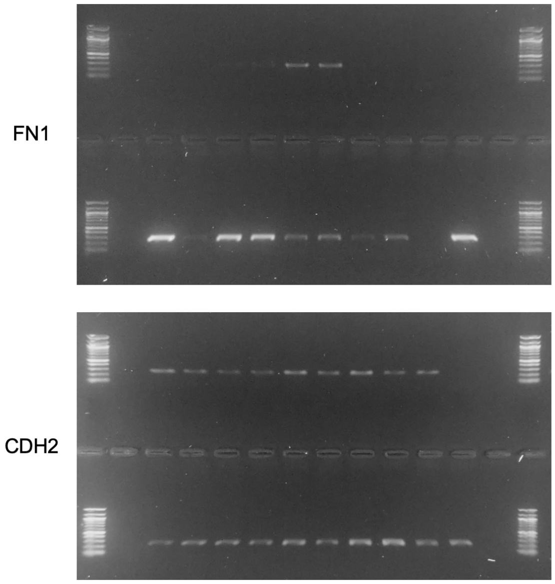
Supplementary Figure 1. Comparative gene expression analysis shows enrichment of post-EMT and mesenchymal pathways in genes with outlier expression in H3WT pHGG tumors. Left: Top 10 pathways from Enrichr gene set enrichment analysis of genes with outlier expression in only nonK27M pHGG samples as compared to a background cohort of similar tumors. Right: Cytoscape EnrichmentMap graph showing clusters of functionally similar pathways enriched for genes with outlier expression in nonK27M pHGG (node p-value < 0.05, edge p-value < 0.38).



Supplementary Figure 2. Top 30 differentially expressed genes by Wilcoxon rank-sum test comparing H3.3K27M mutant glioma cells with H3.1K27M mutant glioma cells.



Supplementary Figure 3. Continuum of EMT completeness scores in glioma single cell RNA-seq data.



Supplementary Figure 4. Full-length RT-PCR gel images for FN1 and CDH2 quantification. Top row of each gel: SU-DIPG-VI, 13, 17, 19, 24, 25, 27, 35, 43, negative RT-PCR control. Bottom row of each gel: SU-DIPG-IV, 21, 33, 36, 38, 48, pcGBM2R, KNS42, SJG2 and normal human astrocytes hTERT.

References

1. Juratli TA, Qin N, Cahill DP, Filbin MG. Molecular pathogenesis and therapeutic implications in pediatric high-grade gliomas. Pharmacol Ther. 2018;182: 70–79.

2. Chan K-M, Fang D, Gan H, Hashizume R, Yu C, Schroeder M, et al. The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. Genes Dev. 2013;27: 985–990.

3. Johung TB, Monje M. Diffuse Intrinsic Pontine Glioma: New Pathophysiological Insights and Emerging Therapeutic Targets. Curr Neuropharmacol. 2017;15: 88–97.

4. Jones C, Baker SJ. Unique genetic and epigenetic mechanisms driving paediatric diffuse high-grade glioma. Nat Rev Cancer. 2014;14. doi:10.1038/nrc3811

5. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131: 803–820.

6. de Vries NA, Hulsman D, Akhtar W, de Jong J, Miles DC, Blom M, et al. Prolonged Ezh2 Depletion in Glioblastoma Causes a Robust Switch in Cell Fate Resulting in Tumor Progression. Cell Rep. 2015;10: 383–397.

7. Mohammad F, Weissmann S, Leblanc B, Pandey DP, Højfeldt JW, Comet I, et al. EZH2 is a potential therapeutic target for H3K27M-mutant pediatric gliomas. Nat Med. 2017;23: 483–492.

8. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature. 2011;469: 343–349.

9. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell. 2008;30: 755–766.

10. Roidl D, Hacker C. Histone methylation during neural development. Cell Tissue Res. 2014;356: 539–552.

11. Sher F, Boddeke E, Olah M, Copray S. Dynamic changes in Ezh2 gene occupancy underlie its involvement in neural stem cell self-renewal and differentiation towards oligodendrocytes. PLoS One. 2012;7: e40399.

12. Sher F, Rössler R, Brouwer N, Balasubramaniyan V, Boddeke E, Copray S. Differentiation of neural stem cells into oligodendrocytes: involvement of the polycomb group protein Ezh2. Stem Cells. 2008;26: 2875–2883.

13. Akizu N, Martínez-Balbás MA. EZH2 orchestrates apicobasal polarity and neuroepithelial cell renewal. Neurogenesis (Austin). 2016;3: e1250034.

14. Zemke M, Draganova K, Klug A, Schöler A, Zurkirchen L, Gay MH-P, et al. Loss of Ezh2 promotes a midbrain-to-forebrain identity switch by direct gene derepression and Wnt-dependent regulation. BMC Biol. 2015;13: 103.

15. Filbin M, Monje M. Developmental origins and emerging therapeutic opportunities for childhood cancer. Nat Med. 2019;25: 367–376.

16. Funato K, Major T, Lewis PW, Allis CD, Tabar V. Use of human embryonic stem cells to model pediatric gliomas with H3.3K27M histone mutation. Science. 2014;346: 1529–1533.

17. Pathania M, De Jay N, Maestro N, Harutyunyan AS, Nitarska J, Pahlavan P, et al. H3.3K27M Cooperates with Trp53 Loss and PDGFRA Gain in Mouse Embryonic Neural Progenitor Cells to Induce Invasive High-Grade Gliomas. Cancer Cell. 2017;32: 684–700.e9.

18. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science. 2018;360: 331–335.

19. Viebahn C. Epithelio-Mesenchymal Transformation during Formation of the Mesoderm in the Mammalian Embryo. Acta Anal. 1995. Available: https://www.karger.com/Article/PDF/147753

20. Duband J-L. Diversity in the molecular and cellular strategies of epithelium-to-mesenchyme transitions: Insights from the neural crest. Cell Adh Migr. 2010;4: 458–482.

21. Kalcheim C. Epithelial-Mesenchymal Transitions during Neural Crest and Somite Development. J Clin Med Res. 2015;5. doi:10.3390/jcm5010001

22. Bolós V, Peinado H, Pérez-Moreno MA, Fraga MF, Esteller M, Cano A. The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: a comparison with Snail and E47 repressors. J Cell Sci. 2003;116: 499–511.

23. Cano A, Pérez-Moreno MA, Rodrigo I, Locascio A, Blanco MJ, del Barrio MG, et al. The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. Nat Cell Biol. 2000;2: 76–83.

24. Lin Y, Dong C, Zhou BP. Epigenetic regulation of EMT: the Snail story. Curr Pharm Des. 2014;20: 1698–1705.

25. Galvagni F, Lentucci C, Neri F, Dettori D, De Clemente C, Orlandini M, et al. Snai1 promotes ESC exit from the pluripotency by direct repression of self-renewal genes. Stem Cells. 2015;33: 742–750.

26. Murray SA, Gridley T. Snail family genes are required for left-right asymmetry determination, but not neural crest formation, in mice. Proc Natl Acad Sci U S A. 2006;103: 10300–10304.

27. Carver EA, Jiang R, Lan Y, Oram KF, Gridley T. The mouse snail gene encodes a key regulator of the epithelial-mesenchymal transition. Mol Cell Biol. 2001;21: 8184–8188.

28. Motta FJN, Valera ET, Lucio-Eterovic AKB, Queiroz RGP, Neder L, Scrideli CA, et al. Differential expression of E-cadherin gene in human neuroepithelial tumors. Genet Mol Res. 2008;7: 295–304.

29. Howng S-L, Wu C-H, Cheng T-S, Sy W-D, Lin P-CK, Wang C, et al. Differential expression of Wnt genes, beta-catenin and E-cadherin in human brain tumors. Cancer Lett. 2002;183: 95–101.

30. Itoh Y, Moriyama Y, Hasegawa T, Endo TA, Toyoda T, Gotoh Y. Scratch regulates neuronal migration onset via an epithelial-mesenchymal transition-like mechanism. Nat Neurosci. 2013;16: 416–425.

31. Ohayon D, Garcès A, Joly W, Soukkarieh C, Takagi T, Sabourin J-C, et al. Onset of Spinal Cord Astrocyte Precursor Emigration from the Ventricular Zone Involves the Zeb1 Transcription Factor. Cell Rep. 2016;17: 1473–1481.

32. Hirabayashi Y, Suzki N, Tsuboi M, Endo TA, Toyoda T, Shinga J, et al. Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. Neuron. 2009;63: 600–613.

33. Li Q, Hutchins AP, Chen Y, Li S, Shan Y, Liao B, et al. A sequential EMT-MET mechanism drives the differentiation of human embryonic stem cells towards hepatocytes. Nat Commun. 2017;8: 15166.

34. Mani SA, Guo W, Liao M-J, Eaton EN, Ayyanan A, Zhou AY, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell. 2008;133: 704–715.

35. Scheel C, Weinberg RA. Cancer stem cells and epithelial-mesenchymal transition: concepts and molecular links. Semin Cancer Biol. 2012;22: 396–403.

36. Ullmann U, In't Veld P, Gilles C, Sermon K, De Rycke M, Van de Velde H, et al. Epithelial-mesenchymal transition process in human embryonic stem cells cultured in feeder-free conditions. Mol Hum Reprod. 2007;13: 21–32.

37. Wang H, Unternaehrer JJ. Epithelial-mesenchymal Transition and Cancer Stem Cells: At the Crossroads of Differentiation and Dedifferentiation. Dev Dyn. 2019;248: 10–20.

38. Christiansen JJ, Rajasekaran AK. Reassessing epithelial to mesenchymal transition as a prerequisite for carcinoma invasion and metastasis. Cancer Res. 2006;66: 8319–8326.

39. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, Ertaylan G, Skupin A, Kuestner RE, et al.

Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. PLoS One. 2015;10: e0126522.

40. Vaske OM, Bjork I, Salama SR, Beale H, Tayi Shah A, Sanders L, et al. Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer. JAMA Netw Open. 2019;2: e1913968.

41. Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, et al. Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. Stem Cell Reports. 2019;12: 245–257.

42. Treehouse Public Data. [cited 21 Apr 2020]. Available: https://treehousegenomics.soe.ucsc.edu/public-data/

43. Mueller S, Jain P, Liang WS, Kilburn L, Kline C, Gupta N, et al. A pilot precision medicine trial for children with diffuse intrinsic pontine glioma - PNOC003: a report from the Pacific Pediatric Neuro-Oncology Consortium. Int J Cancer. 2019. doi:10.1002/ijc.32258

44. Grasso CS, Tang Y, Truffaux N, Berlow NE, Liu L, Debily M-A, et al. Functionally defined therapeutic targets in diffuse intrinsic pontine glioma. Nat Med. 2015;21: 555–559.

45. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 2016;164: 550–563.

46. Mackay A, Burford A, Carvalho D, Izquierdo E, Fazal-Salom J, Taylor KR, et al. Integrated Molecular Meta-Analysis of 1,000 Pediatric High-Grade and Diffuse Intrinsic Pontine Glioma. Cancer Cell. 2017;32: 520–537.e5.

47. Robinson DR, Wu Y-M, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548: 297–303.

48. Sturm D, Orr BA, Toprak UH, Hovestadt V, Jones DTW, Capper D, et al. New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs. Cell. 2016;164: 1060–1072.

49. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.

50. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550.

51. Koncar RF, Dey BR, Stanton A-CJ, Agrawal N, Wassell ML, McCarl LH, et al. Identification of Novel RAS Signaling Therapeutic Vulnerabilities in Diffuse Intrinsic Pontine Gliomas. Cancer Res. 2019;79: 4026–4041.

52. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1: 417–425.

53. Chung M-T, Lai H-C, Sytwu H-K, Yan M-D, Shih Y-L, Chang C-C, et al. SFRP1 and SFRP2 suppress the transformation and invasion abilities of cervical cancer cells through Wnt signal pathway. Gynecol Oncol. 2009;112: 646–653.

54. Maupin KA, Sinha A, Eugster E, Miller J, Ross J, Paulino V, et al. Glycogene expression alterations associated with pancreatic cancer epithelial-mesenchymal transition in complementary model systems. PLoS One. 2010;5: e13002.

55. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proc Natl Acad Sci U S A. 2010;107: 15449–15454.

56. Xue T-C, Ge N-L, Zhang L, Cui J-F, Chen R-X, You Y, et al. Goosecoid promotes the metastasis of hepatocellular carcinoma by modulating the epithelial-mesenchymal transition. PLoS One. 2014;9: e109695.

57. Tran DD, Corsa CAS, Biswas H, Aft RL, Longmore GD. Temporal and spatial cooperation of Snail1 and Twist1 during epithelial-mesenchymal transition predicts for human breast cancer recurrence. Mol Cancer Res. 2011;9: 1644–1657.

58. Stanisavljevic J, Porta-de-la-Riva M, Batlle R, de Herreros AG, Baulida J. The p65 subunit of NF-κB and PARP1 assist Snail1 in activating fibronectin transcription. J Cell Sci. 2011;124: 4161–4171.

59. Javaid S, Zhang J, Anderssen E, Black JC, Wittner BS, Tajima K, et al. Dynamic chromatin modification sustains epithelial-mesenchymal transition following inducible expression of Snail-1. Cell Rep. 2013;5: 1679–1689.

60. Tanaka S, Kobayashi W, Haraguchi M, Ishihata K, Nakamura N, Ozawa M. Snail1 expression in human colon cancer DLD-1 cells confers invasive properties without N-cadherin expression. Biochem Biophys Rep. 2016;8: 120–126.

61. Lewis PW, Müller MM, Koletsky MS, Cordero F, Lin S, Banaszynski LA, et al. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. Science. 2013;340: 857–861.

62. Larson JD, Kasper LH, Paugh BS, Jin H, Wu G, Kwon C-H, et al. Histone H3.3 K27M Accelerates Spontaneous Brainstem Glioma and Drives Restricted Changes in Bivalent Gene Expression. Cancer Cell. 2019;35: 140–155.e7.

63. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539: 309–313.

64. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell. 2019. doi:10.1016/j.cell.2019.06.024

65. Tan TZ, Miow QH, Miki Y, Noda T, Mori S, Huang RY-J, et al. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med. 2014;6: 1279–1293.

66. Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, William WN, et al. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clin Cancer Res. 2016;22: 609–620.

67. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. arXiv [cs.MS]. 2019. Available:

http://arxiv.org/abs/1907.10121

68. Lukas J, Petersen BO, Holm K, Bartek J, Helin K. Deregulated expression of E2F family members induces S-phase entry and overcomes p16INK4A-mediated growth suppression. Mol Cell Biol. 1996;16: 1047–1057.

69. Holness CL, Simmons DL. Molecular cloning of CD68, a human macrophage marker related to lysosomal glycoproteins. Blood. 1993;81: 1607–1613.

70. Nakamura Y, Iwamoto R, Mekada E. Expression and distribution of CD9 in myelin of the central and peripheral nervous systems. Am J Pathol. 1996;149: 575–583.

71. Weng Q, Chen Y, Wang H, Xu X, Yang B, He Q, et al. Dual-mode modulation of Smad signaling by Smad-interacting protein Sip1 is required for myelination in the central nervous system. Neuron. 2012;73: 713–728.

72. Kagawa T, Mekada E, Shishido Y, Ikenaka K. Immune system-related CD9 is expressed in mouse central nervous system myelin at a very late stage of myelination. J Neurosci Res. 1997;50: 312–320.

73. Richardson WD, Pringle N, Mosley MJ, Westermark B, Dubois-Dalcq M. A role for platelet-derived growth factor in normal gliogenesis in the central nervous system. Cell. 1988;53: 309–319.

74. Zeisberg M, Neilson EG. Biomarkers for epithelial-mesenchymal transitions. J Clin Invest. 2009;119: 1429–1437.

75. Sancisi V, Gandolfi G, Ragazzi M, Nicoli D, Tamagnini I, Piana S, et al. Cadherin 6 is a new RUNX2 target in TGF-β signalling pathway. PLoS One. 2013;8: e75489.

76. Vallath S, Sage EK, Kolluri KK, Lourenco SN, Teixeira VS, Chimalapati S, et al. CADM1 inhibits squamous cell carcinoma progression by reducing STAT3 activity. Sci Rep. 2016;6: 24006.

77. Sakurai-Yageta M, Masuda M, Tsuboi Y, Ito A, Murakami Y. Tumor suppressor CADM1 is involved in epithelial cell structure. Biochem Biophys Res Commun. 2009;390: 977–982.

78. Kim J, Kang HS, Lee Y-J, Lee H-J, Yun J, Shin JH, et al. EGR1-dependent PTEN upregulation by 2-benzoyloxycinnamaldehyde attenuates cell invasion and EMT in colon cancer. Cancer Lett. 2014;349: 35–44.

79. Xu J, Lamouille S, Derynck R. TGF-beta-induced epithelial to mesenchymal transition. Cell Res. 2009;19: 156–172.

80. Tsukita S, Furuse M. Occludin and claudins in tight-junction strands: leading or supporting players? Trends Cell Biol. 1999;9: 268–273.

81. Zhang Y, Feng XH, Derynck R. Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF-beta-induced transcription. Nature. 1998;394: 909–913.

82. Barrett CSX, Millena AC, Khan SA. TGF-β Effects on Prostate Cancer Cell Migration and Invasion Require FosB. Prostate. 2017;77: 72–81.

83. Lv Q-L, Huang Y-T, Wang G-H, Liu Y-L, Huang J, Qu Q, et al. Overexpression of RACK1 Promotes Metastasis by Enhancing Epithelial-Mesenchymal Transition and Predicts Poor Prognosis in Human Glioma. Int J Environ Res Public Health. 2016;13. doi:10.3390/ijerph13101021

84. Berndt A, Richter P, Kosmehl H, Franz M. Tenascin-C and carcinoma cell invasion in oral and urinary bladder cancer. Cell Adh Migr. 2015;9: 105–111.

85. Turunen SP, Tatti-Bugaeva O, Lehti K. Membrane-type matrix metalloproteases as diverse effectors of cancer progression. Biochim Biophys Acta Mol Cell Res. 2017;1864: 1974–1988.

86. Jing P, Zhao N, Xie N, Ye M, Zhang Y, Zhang Z, et al. miR-24-3p/FGFR3 Signaling as a Novel Axis Is Involved in Epithelial-Mesenchymal Transition and Regulates Lung Adenocarcinoma Progression. J Immunol Res. 2018;2018: 2834109.

87. Liu J, Li J, Ren Y, Liu P. DLG5 in cell polarity maintenance and cancer development. Int J Biol Sci. 2014;10: 543–549.

88. Zhang X, Song Q, Wei C, Qu J. LRIG1 inhibits hypoxia-induced vasculogenic mimicry formation via suppression of the EGFR/PI3K/AKT pathway and epithelial-to-mesenchymal transition in human

glioma SHG-44 cells. Cell Stress Chaperones. 2015;20: 631–641.

89. Liu X, Li C, Zhang R, Xiao W, Niu X, Ye X, et al. The EZH2- H3K27me3-DNMT1 complex orchestrates epigenetic silencing of the wwc1 gene, a Hippo/YAP pathway upstream effector, in breast cancer epithelial cells. Cell Signal. 2018;51: 243–256.

90. Castel D, Philippe C, Calmon R, Le Dret L, Truffaux N, Boddaert N, et al. Histone H3F3A and HIST1H3B K27M mutations define two subgroups of diffuse intrinsic pontine gliomas with different prognosis and phenotypes. Acta Neuropathol. 2015;130: 815–827.

91. Szenker E, Ray-Gallet D, Almouzni G. The double face of the histone variant H3.3. Cell Res. 2011;21: 421–434.

92. Goldberg AD, Banaszynski LA, Noh K-M, Lewis PW, Elsaesser SJ, Stadler S, et al. Distinct factors control histone variant H3.3 localization at specific genomic regions. Cell. 2010;140: 678–691.

93. Nagaraja S, Quezada MA, Gillespie SM, Arzt M, Lennon JJ, Woo PJ, et al. Histone Variant and Cell Context Determine H3K27M Reprogramming of the Enhancer Landscape and Oncogenic State. Mol Cell. 2019. doi:10.1016/j.molcel.2019.08.030

94. Lin GL, Monje M. A Protocol for Rapid Post-mortem Cell Culture of Diffuse Intrinsic Pontine Glioma (DIPG). J Vis Exp. 2017. doi:10.3791/55360

95. Kim MY, Kaduwal S, Yang DH, Choi KY. Bone morphogenetic protein 4 stimulates attachment of neurospheres and astrogenesis of neural stem cells in neurospheres via phosphatidylinositol 3 kinase-mediated upregulation of N-cadherin. Neuroscience. 2010;170: 8–15.

96. Jones C, Karajannis MA, Jones DTW, Kieran MW, Monje M, Baker SJ, et al. Pediatric high-grade glioma: biologically and clinically in need of new thinking. Neuro Oncol. 2017;19: 153–161.

97. Hargrave D, Bartels U, Bouffet E. Diffuse brainstem glioma in children: critical review of clinical trials. Lancet Oncol. 2006;7: 241–248.

98. Meel MH, Schaper SA, Kaspers GJL, Hulleman E. Signaling pathways and mesenchymal transition in pediatric high-grade glioma. Cell Mol Life Sci. 2018;75: 871–887.

99. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. Nat Med. 2013;19: 1438–1449.

100. Puget S, Philippe C, Bax DA, Job B, Varlet P, Junier M-P, et al. Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. PLoS One. 2012;7: e30313.

101. Ott M, Litzenburger UM, Sahm F, Rauschenbach KJ, Tudoran R, Hartmann C, et al. Promotion of glioblastoma cell motility by enhancer of zeste homolog 2 (EZH2) is mediated by AXL receptor kinase. PLoS One. 2012;7: e47663.

102. Vajkoczy P, Knyazev P, Kunkel A, Capelle H-H, Behrndt S, von Tengg-Kobligk H, et al. Dominant-negative inhibition of the Axl receptor tyrosine kinase suppresses brain tumor cell growth and invasion and prolongs survival. Proc Natl Acad Sci U S A. 2006;103: 5799–5804.

103. Yin Y, Qiu S, Peng Y. Functional roles of enhancer of zeste homolog 2 in gliomas. Gene. 2016;576: 189–194.

104. Jolly MK, Mani SA, Levine H. Hybrid epithelial/mesenchymal phenotype(s): The “fittest” for metastasis? Biochim Biophys Acta Rev Cancer. 2018;1870: 151–157.

105. Pan M-R, Hsu M-C, Chen L-T, Hung W-C. Orchestration of H3K27 methylation: mechanisms and therapeutic implication. Cell Mol Life Sci. 2018;75: 209–223.

106. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555: 321–327.

107. Sun L, Fang J. Epigenetic regulation of epithelial-mesenchymal transition. Cell Mol Life Sci. 2016;73: 4493–4515.

108. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35: 314–316.

109. Fang Z. GSEApy.

110. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010;5: e13984.

111. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352: 189–196.

Chapter 4: Large genomics datasets for analyzing the utility of preclinical cancer models

4.1 Chapter Introduction

Preclinical models, including cancer-derived cell lines, are essential for drug development in pediatric cancer. While it is well known that cancer cell lines do not completely recapitulate the original tumor biology, the environmentally induced molecular differences between tumor and cell line have not been well characterized. It is important to know which signaling pathways are not representative of tumor biology, especially in the context of precision drug testing.

For example, a screen of DIPG cell lines identified the histone deacetylase inhibitor panobinostat as a promising agent both in cell lines and xenografts[21]. Unfortunately, in a subsequent clinical trial where 11/15 patients were treated with panobinostat in combination with other therapies, all 11 patients were deceased by the end of the trial[30]. While this example partly speaks to the overall difficult nature of DIPG, it is still puzzling that the *in vitro* cell line response did not better predict the clinical response.

In 2016, an excellent review article from the labs of Drs. Joe Gray and James Costello called for large-scale genomic analysis of publicly available cell line and tumor data to provide a comprehensive understanding of tumor vs. cell line biology, which they noted is becoming particularly important in the era of precision medicine[31]. As a student at the UCSC Genomics Institute, where huge amounts of genomic cancer data are uniformly processed to remove technical batch effects and made publicly available on the UCSC Xena platform[32], I was well positioned to undertake this analysis. Accordingly, I motivated the reprocess of the Cancer Cell Line Encyclopedia (CCLE) raw mRNA sequencing data through the UCSC TOIL RNA-seq pipeline, so that it would be comparable with The Cancer Genome Atlas (TCGA) expression dataset which had also been processed through TOIL[33]. Because of this effort, the processed CCLE dataset is now publicly available on Xena for the larger scientific community to use Xena's many visualization and analysis tools, or download the dataset.

This study involves not only the comparison of CCLE and TCGA mRNA expression data, but protein and miRNA data as well. In order to comprehensively evaluate the utility of the most widely used preclinical cancer models, I also included publicly available cancer organoid and patient derived xenograft gene expression data. This study identifies certain cancer driver pathways which are not representative of tumor biology and provides a roadmap for preclinical cancer researchers to evaluate the appropriateness of their models, avoid potentially non-translatable experiments, and develop improved models.

I designed and led this study. The contributions of others are as follows: the CCLE reprocess was completed by Jacob Pfeil, Rob Currie and Ellen Kephart. The original support vector machine analysis was performed by Rahul Chandra (second author) with oversight from Dr. David Vengerov.

This manuscript has been submitted to *Communications Biology,* June 2020.

Comparative multi-omic analysis of cancer cell lines and primary tumors from large public compendia reveals deregulation of key cancer driver pathways

Lauren M. Sanders[1*], Rahul Chandra[2], Ellen Towle Kephart[1], Jacob Pfeil[1], Allison Cheney[6], Katrina Learned[1], Rob Currie[1], Leonid Gitlin[3], David Vengerov[4], David Haussler[1&], Sofie R. Salama[1,5&], Olena M. Vaske[6&]

[1] Department of Biomolecular Engineering, UC Santa Cruz Genomics Institute

[2] Paul G. Allen School of Computer Science and Engineering, University of Washington

[3] Current Address: Gritstone Oncology, Incorporated, Emeryville, CA

[4] Oracle Labs, Oracle Corporation

[5] Howard Hughes Medical Institute, University of California Santa Cruz

[6] Department of Molecular, Cell and Developmental Biology, UC Santa Cruz Genomics Institute

[*]Corresponding author

E-mail: lmsh@ucsc.edu

[&]Co-senior author

Abstract

Cancer cell lines have been widely used for decades to study biological processes driving cancer development, and to identify biomarkers of response to therapeutic agents. Advances in genomic sequencing have made possible large-scale genomic characterizations of collections of cancer cell lines and primary tumors, such as the Cancer Cell Line Encyclopedia (CCLE) and The Cancer Genome Atlas (TCGA). These studies allow for the first time a comprehensive evaluation of the comparability of cancer cell lines and primary tumors on the genomic and proteomic level. Here we employ bulk mRNA and micro-RNA sequencing data from thousands of samples in CCLE and TCGA, and proteomic data from partner studies in the MD Anderson Cell Line Project (MCLP) and The Cancer Proteome Atlas (TCPA), to characterize the extent to which cancer cell lines recapitulate tumors. We identify dysregulation of a long non-coding RNA and microRNA regulatory network in cancer cell lines, associated with differential expression between cell lines and primary tumors in four key cancer driver pathways: KRAS signaling, NFKB signaling, IL2/STAT5 signaling and TP53 signaling. Our results emphasize the necessity for careful interpretation of cancer cell line experiments, particularly with respect to therapeutic treatments targeting these important cancer pathways, and highlight the importance of 3D culturing techniques, which our analysis suggests may lack some of these differences.

Introduction

Tumor-derived cell lines provide a robust model environment for testing treatment hypotheses, identifying biomarkers of response to therapies, and studying underlying cancer biology. Cancer cell line models grow quickly, are comparatively cost effective, and are readily available. Their integration into pre-clinical research has led to remarkable advances in cancer characterization and treatment[1].

However, additional genomic characterization of cancer cell lines has indicated that the transition from *in vivo* to *in vitro* may introduce key genomic alterations. One of the first groups to compare tumor and cell line used microarray gene expression profiles to identify breast cancer cell lines that seemed to be genetically inappropriate models for breast carcinoma[2]. As mutation detection has become more accurate, multiple studies have reported that head and neck cancer cell lines tend to harbor more mutations than their tumors of origin[3,4]. A recent study showed that colorectal cancer cell lines recapitulate colorectal tumor subtypes, but that cell lines have more mutations than tumors[5]. In general, the current consensus concerning cancer cell lines as primary tumor models is that cell lines share many of the original tumor characteristics, but can harbor genetic changes of poorly characterised significance; and that some cancer cell lines may not even be molecularly appropriate or representative models for their tumor of origin[6].

Nevertheless, cancer cell lines continue to be widely used in cancer research and therapeutic discovery. As the focus on molecularly targeted therapeutics grows, so does the need to thoroughly characterize how cancer cell lines diverge phenotypically from tumors due to their *in vitro* growth environment[6]. It is essential that preclinical researchers know which biological pathways behave similarly *in vivo* and *in vitro*, and even more importantly, which pathways demonstrate altered activity as a result of alterations in environmental signals and stressors in the *in vitro* growth setting. If a key pathway behaves differently in cell lines as compared to primary tumors, preclinical testing of a drug targeting that pathway will not accurately predict patient tumor response.

In order to characterize the specific pathway alterations that occur between primary tumors and tumor-derived cell lines, we analyze three types of high-throughput molecular data from The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE). We perform transcriptomic analysis of bulk RNA sequencing of TCGA and CCLE samples and identify a set of differentially expressed genes. We integrate micro-RNA sequencing from the same projects and identify an interaction network of microRNA, long non-coding (lncRNA) and protein coding genes that is aberrantly expressed in cell lines compared to tumors. This network implicates 4 key cancer driver pathways that are often the subject of preclinical drug evaluation in cell lines, but whose activity in cell lines is not representative of original tumors. We use proteomic quantification data from many of the same samples to demonstrate that the aberrant cancer driver pathway expression observed in cell lines extends to the proteomic level. Finally, we demonstrate that in some tumor types, tumor-derived organoids and PDX models can serve as more accurate representations of tumor biology.

Results

A. Support vector machine classifier identifies a set of genes differentially expressed between primary tumors and tumor-derived cell lines.

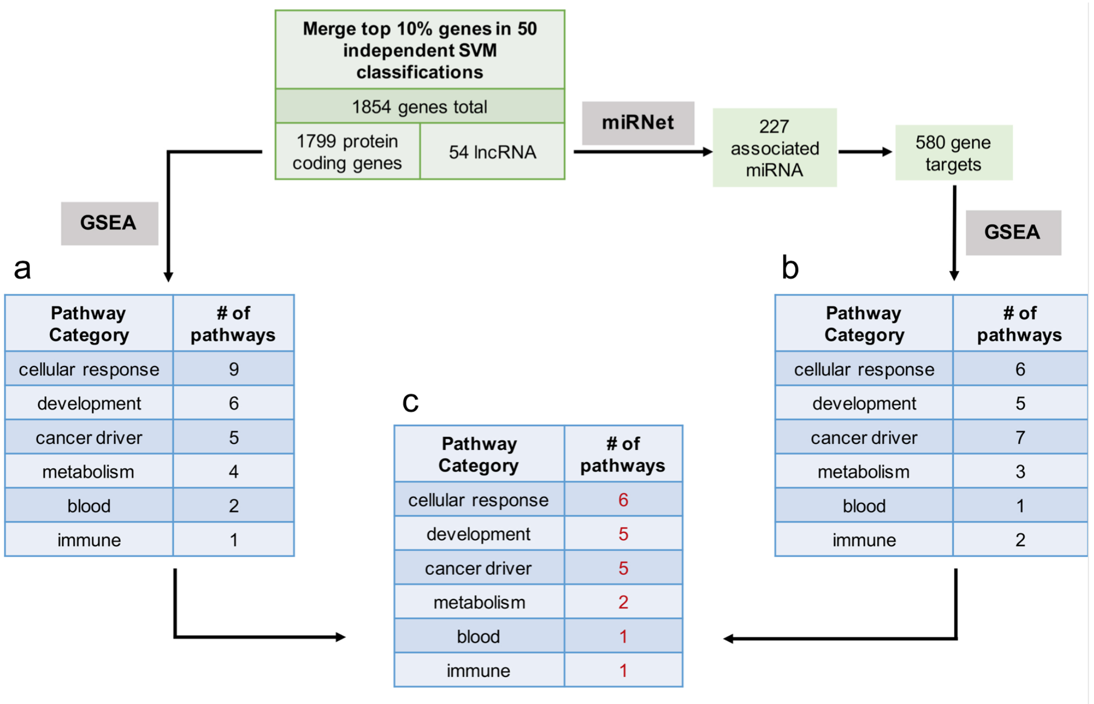
We hypothesized that genes with differential expression between the TCGA and CCLE datasets would represent differences in biological pathway activity. In order to identify novel sources of variation within these datasets, we eliminated immune-related genes because it is already known that cancer cell lines are unable to recapitulate the immune signatures of the primary tumors[7] (see Methods, Supplementary Fig. 1, Supplementary Table 1).

We then used a support vector machine (SVM) linear classifier within the Python *sklearn* module to identify genes (features) which are the most useful and important for classifying a new tumor or cell line based on a trained model[8]. In order to ensure robust results, we repeated the classification on fifty different random 80/20 test/training splits of the data. After sorting the genes by their SVM-assigned feature importance coefficients, we merged the top 10% of genes from all fifty classifications, resulting in 1854 genes that were in the top 10% most important genes for each classification (Supplementary Table 1). These genes included 54 long non-coding RNA (lncRNA) and 1799 protein-coding genes.

In order to characterize the functional significance of our SVM-derived gene set, we performed gene set enrichment analysis (GSEA) on the 1799 protein-coding genes using the Hallmark cancer pathway set from the Molecular Signatures Database (mSigDB)[9]. We found 27 gene sets with significant enrichment in the SVM-derived differentially expressed protein-coding genes (Fig. 1a, Supplementary Table 1, *tab 3*).

Since lncRNA play known regulatory roles in normal tissue and in cancer, we hypothesized that the 54 differentially expressed lncRNA may be involved in regulating the differentially expressed coding genes, and may compose an interaction network with aberrant expression in cell culture. In order to characterize the functional interactions of these lncRNA, we employed miRNet, a tool that integrates multiple interaction databases for identification of lncRNA-miRNA and miRNA-gene regulatory networks[10]. miRNet identified 227 miRNA with known interactions to our 54 lncRNA. In turn, these 227 miRNA had 580 known gene targets among the 1799 differentially expressed coding genes (pvalue<7.8^{-27}, hypergeometric test). GSEA of the 580 coding genes revealed 24 Hallmark gene sets with significant enrichment (Fig. 1b, Supplementary Table 1, *tab 3*). Strikingly, 20 of these Hallmark pathways overlapped with the enriched pathways from the coding genes GSEA (Fig. 1c, Supplementary Table 1, *tab 3*), indicating that the set of SVM-derived important lncRNAs is closely involved in many of the same pathways as the set of SVM-derived important coding genes.

We categorized the pathways into 6 categories (cellular response, development, cancer driver, metabolism, blood, and immune) and were particularly interested in the 5 cancer driver pathways which overlapped between the coding genes GSEA and the lncRNA-derived GSEA.


Merge top 10% genes in 50 independent SVM classifications
1854 genes total
1799 protein coding genes
54 lncRNA
miRNet
227 associated miRNA
580 gene targets
GSEA
GSEA
a
b
c

Pathway Category	# of pathways
cellular response	9
development	6
cancer driver	5
metabolism	4
blood	2
immune	1

Pathway Category	# of pathways
cellular response	6
development	5
cancer driver	7
metabolism	3
blood	1
immune	2

Pathway Category	# of pathways
cellular response	6
development	5
cancer driver	5
metabolism	2
blood	1
immune	1

Fig. 1. Workflow for support vector machine (SVM) classification of samples from TCGA and CCLE to assign feature importance to all genes. a) GSEA pathway enrichment results of 1799 protein coding genes in the overlap of the top 10% most important genes in 50 independent SVM classifications. b) GSEA pathway enrichment results of 580 protein coding genes linked by miRNet databases to the 54 lncRNAs found in the overlap of the top 10% most important genes in 50 independent SVM classifications. c) Overlap of pathways from the protein coding gene GSEA and the lncRNA-based GSEA.

B. Four main types of cancer driver pathways exhibit differential expression and protein levels in cancer cell lines compared to primary tumors.

The 5 cancer driver pathways with significant enrichment in both SVM-derived coding genes and lncRNA-related genes are KRAS Signaling Up and Down, P53 Pathway, IL2/STAT5 Signaling, and TNFA Signaling Via NFKB (Fig. 2a). With respect to KRAS signaling, since both pathways are a result of activated KRAS signaling, all subsequent analysis focuses on KRAS Signaling Up, which represents genes upregulated as a result of activated KRAS signaling.

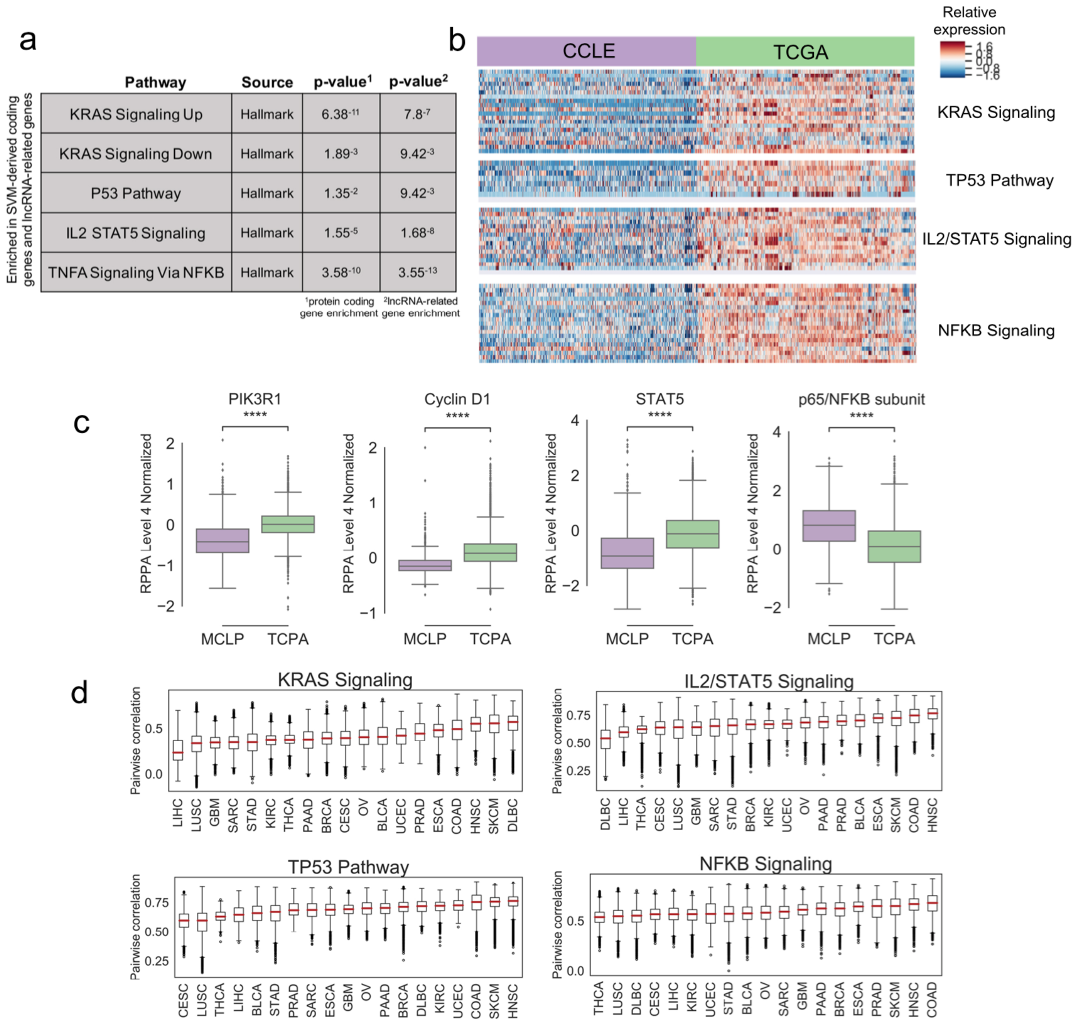
Interestingly, all 4 pathways show much higher overall expression in tumors than in cell lines, indicating that these genes are downregulated as a result of the transition from tumor to cell culture dish (Fig. 2b). As a control, we also examined the gene expression of the Hallmark PI3K-AKT-mTOR pathway, a cancer driver pathway that was not significantly enriched in SVM-derived genes (pvalue < 0.426). This pathway did not show differential expression between CCLE and TCGA (Supplementary Fig. 2). We also verified that this signal was not a disease-specific artifact by repeating the SVM after subsetting the data to the disease with the largest number of samples in TCGA (BRCA) and the smallest number of samples (DLBC). The same 4 cancer driver pathways were identified in both analyses (Supplementary Table 1).

We next examined whether the downregulation of these pathways extended beyond gene expression into protein activity. Proteomics quantification of many CCLE and TCGA samples was performed using Reverse Phase Protein Array (RPPA) in the MD Anderson Cell Lines Project (MCLP) and The Cancer Proteome Atlas (TCPA)[11,12]. Using the RPPA Level 4 Normalized data, we identified proteins that are normally highly expressed downstream of each cancer driver pathway, and examined their levels in the cell line and tumor data (Fig. 2c). PIK3R1 (antibody PI3KP85) is activated subsequent to KRAS signaling, and Cyclin D1 (antibody CYCLIND1) is activated downstream of the P53 pathway[13,14]. STAT5 (antibody STAT5ALPHA) represents the protein counterpart of the *STAT5* gene. The antibody NFKBP65_pS536 binds to phosphorylated p65, one of the two protein subunits of NFKB. Phosphorylation of p65 is one of several molecular mechanisms known to activate the NFKB pathway[15,16].

We noted significantly lower protein expression of PIK3R1, Cyclin D1 and STAT5 in the cell line data, consistent with our gene expression results in the corresponding pathways. This carries important implications for the applicability of preclinical drug tests against these targets in cancer cell lines. Interestingly, the phosphorylation level of p65 is higher in cell lines than tumors, opposite the gene expression of the NFKB signaling pathway. This suggests that p65 phosphorylation may be playing a different role in cell lines, and underscores the importance of examining multiple types of data to elucidate complex molecular interactions.

To explore disease-specific pathway expression differences, we calculated pairwise gene expression correlation scores between all TCGA and CCLE samples of each disease type (Fig. 2d), as well as mean correlation scores between all disease types in both datasets (Supplementary Fig. 3). We observed that the correlations between tumor and cell line were consistently higher in certain tumor types, across all four cancer driver pathways. HNSC, SKCM and COAD ranked highest, with tumor-cell line correlation scores in the top 4 most correlated diseases for all four cancer driver pathways. Conversely, LIHC and LUSC consistently ranked lowest and had tumor-cell line correlation scores in the bottom 5 least correlated diseases for each pathway. These results suggest that while global dysregulation of these cancer driver pathways occurs consistently across cancer cell lines, the HSNC, SKCM and COAD derived cell lines most closely resemble their primary tumors while LIHC and LUSC cell lines are least similar to their primary tumors. This analysis suggests the need for caution when interpreting preclinical results, particularly in disease types with low tumor-cell line correlation in the relevant signaling pathways.

Because activation of the KRAS and TP53 pathways are associated with mutations in the *KRAS* and *TP53* genes[17,18], we investigated whether there is a correlation between diseases with heavy mutation burden in these genes and diseases with higher correlation between tumor and cell line, with the assumption that cell lines derived from mutated tumors maintain those mutations. We found that there is no correlation between mutational burden and tumor-cell line correlation; in fact, in the case of the TP53 pathway, there is a slight inverse correlation between the two factors (Supplementary Fig. 4). This indicates that dysregulation of these pathways occurs in cancer cell lines regardless of the mutational status of the primary tumor, and is unrelated to the activating DNA mutations.



Pathway	Source	p-value[1]	p-value[2]
KRAS Signaling Up	Hallmark	6.38^{-11}	7.8^{-7}
KRAS Signaling Down	Hallmark	1.89^{-3}	9.42^{-3}
P53 Pathway	Hallmark	1.35^{-2}	9.42^{-3}
IL2 STAT5 Signaling	Hallmark	1.55^{-5}	1.68^{-8}
TNFA Signaling Via NFKB	Hallmark	3.58^{-10}	3.55^{-13}

[1] protein coding gene enrichment [2] lncRNA-related gene enrichment

Fig. 2. KRAS Signaling, TP53 Pathway, IL2/STAT5 Signaling and NFKB Signaling are significantly enriched for genes with reduced expression in CCLE compared to TCGA. a) GSEA results for the cancer driver pathways which overlap with SVM-derived genes. P values are shown for significance of gene overlap with SVM-derived protein coding genes, and for genes linked by miRNet to SVM-derived lncRNA. b) Heatmaps showing expression of SVM-identified genes in 4 cancer driver pathways in TCGA compared to CCLE. Samples shown are a random subset with equal representation from each dataset in each disease. c) Boxplots showing overall protein quantification of representative proteins from each of the

4 cancer driver pathways in tumor and cell line datasets. (Mann-Whitney significance test; * pvalue < 0.05, ** pvalue < 0.01, *** pvalue < 0.001) d) Boxplots show pairwise Spearman correlation scores between all CCLE and TCGA RNA-seq samples in each disease type, for all 4 cancer driver pathways. Plots are sorted by mean correlation.

C. Dysregulation of a lncRNA-miRNA regulatory network in cancer cell lines is associated with underexpression of key cancer pathways.

Because the four cancer driver pathways were derived in the context of lncRNA-related gene expression, we hypothesized that cell-line-specific dysregulation of lncRNA-based regulation programs may cause aberrant pathway-level gene expression. lncRNA control gene expression in a tissue-specific manner, and one of their key regulatory mechanisms is by sequestering or “sponging” microRNA (miRNA) through base pairing interactions[19–21]. miRNA directly affect gene expression by binding mRNA and targeting them for degradation[22–24]. In this method of expression control, lncRNA regulate miRNA, while miRNA regulate gene expression (Fig. 3a).

To investigate potential non-coding RNA dysregulation in cell lines as compared to tumors, we focused on the 54 lncRNA identified as differentially expressed through the SVM classification (Fig. 1, Supplementary Table 1). We used miRNet databases to link the 54 differentially expressed lncRNA to the 4 differentially expressed cancer driver pathways via shared miRNA interactions (Supplementary Table 2). Via miRNet, we found that 77 miRNA have known interactions both with genes in the four cancer driver pathways, and with 11 of the differentially expressed lncRNA (11 lncRNA: *LBX2-AS1, CERS6-AS1, DLGAP1-AS1, H19, IQCH-AS1,* *LINC00240, LINC00665, LINC00707, LINC00847, LINC00622, LIMD1-AS1*). With the exception of *LINC00707*, 10 of the 11 lncRNA are significantly underexpressed in CCLE (Fig. 3b,C). We hypothesized that the reduced cell line expression of these lncRNA may cause subsequent expression changes in the downstream miRNA regulatory network, which in turn would cause aberrant expression of the four cancer driver pathways being controlled by the miRNA network.

In order to investigate this hypothesis, we leveraged publicly available miRNA sequencing (miRNAseq) data from CCLE and TCGA. We used ComBat correction to remove experimental batch effects (see Methods, Supplementary Fig. 5)[25]. Sixty-nine of the 77 miRNA were quantified in both miRNAseq datasets, so we used these miRNA for all downstream analyses (Supplementary Table 3). We calculated the log fold change (LFC) in expression between CCLE and TCGA for these 69 miRNA. Notably, over half of the miRNA (n=43) are more highly expressed in cell lines than tumors. Cytoscape was used to visualize the lncRNA-miRNA-coding gene network colored by gene type or by LFC (Fig. 4a,b, Supplementary Table 3)[26].

In keeping with the lncRNA "sponge" regulatory model, the lncRNA are underexpressed in cancer cell lines, which in turn allows the observed overexpression of a majority of the miRNA whose expression is normally kept in check by these lncRNA. The end result is the observed underexpression of key genes in 4 important cancer driver pathways in cancer cell lines, due to aberrant overexpression of inhibitory miRNA. Indeed, we noted several miRNA with higher expression in cell lines that are known to play roles in regulation of the four cancer driver pathways. *mir-497, mir-195, mir-148a* and *mir-152* directly inhibit genes in the KRAS/RAF/MEK/ERK pathway[27,28]. The TP53 pathway is repressed by *mir-339*, and TP53-associated gene

TP53INP1 is regulated by *mir-92*[29,30]. *mir-519d* directly represses *STAT3*, a key gene in the IL2/STAT5 signaling pathway[31]. The NFKB pathway is activated by *mir-301a*, which has lower expression in cell lines compared to tumors, in keeping with lower NFKB activity in cell lines[32].

Because lncRNA and miRNA are known for cell-type-specific expression, we hypothesized that the observed dysregulation of lncRNA-miRNA expression networks is caused by biological selection for a subset of cancer cells which are more likely to survive the cell line derivation process and thrive in a cell culture setting. Consistent with this hypothesis, both stem cell and epithelial cell specific lncRNA and miRNA display reduced expression in cancer cell lines. Specifically, CCLE samples have reduced expression of *H19*, a lncRNA strongly associated with the cancer stem cell state[33], but show increased expression of *mir-1* and *mir-206,* which promote cellular differentiation by blocking anti-differentiation signaling targets[34,35]. Additionally, CCLE samples show reduced expression of *CERS6-AS1, IQCH-AS1* and *LINC00240*, lncRNA implicated in mediating tight junctions or extracellular matrix interactions, which are features of epithelial and endothelial cells[36–39]. At the same time, CCLE samples have comparatively high expression of *mir-9*, which directly represses E-cadherin, a well known epithelial marker[40]. E-cadherin repression is known to induce the epithelial-mesenchymal-transition, a process which plays a role in cancer progression from an epithelial state to a motile and invasive metastatic state[41]. CCLE samples display reduced expression of E-cadherin/*CDH1*, and higher expression of mesenchymal markers including N-cadherin/*CDH2*, *MUC1*, and claudins *CLDN1, CLDN2, CLDN3*[42] (Supplementary Fig. 6).

The observed reduced epithelial and stem cell expression in cancer cell lines suggests that cancer cell culture conditions select for the subset of cancer cells with a mesenchymal, invasive and metastatic phenotype. Overall, these results indicate that selection against specific cancer cell types in tumor-derived cell lines may cause global downregulation of key cell-type-specific lncRNAs, which in turn allows overexpression of a variety of miRNA, many of which play important roles in regulating cancer signaling pathways.

In light of recent research identifying a panel of 110 CCLE cell lines with highest correlation to their primary tumor samples, the TCGA-110-CL[7], we examined whether these cell lines show more representative expression of the 4 cancer driver pathways. We repeated the SVM after subsetting the CCLE dataset to the TCGA-110-CL and the TCGA dataset to the tumor types in the TCGA-110-CL (Supplementary Table 1). Interestingly, the same 4 cancer driver pathways were again identified as differentially expressed, although for IL2/STAT5, TP53 and NFKB signaling the overall correlations were higher (Supplementary Fig. 7). However, several metabolic, cellular response and developmental pathways that were identified in the original analysis were not identified here, including Hedgehog Signaling, Apical Junction and Fatty Acid Metabolism (Supplementary Table 1). Overall, these results indicate that while the TCGA-110-CL cell line panel is indeed more representative of its primary tumors by overall gene expression, our pathway-level examination reveals that caution must still be used when interpreting results involving targeting these 4 cancer driver pathways. This result is consistent with our hypothesis that the dysregulation of cancer driver signaling is driven by a loss of cellular heterogeneity overall in cancer cell lines.

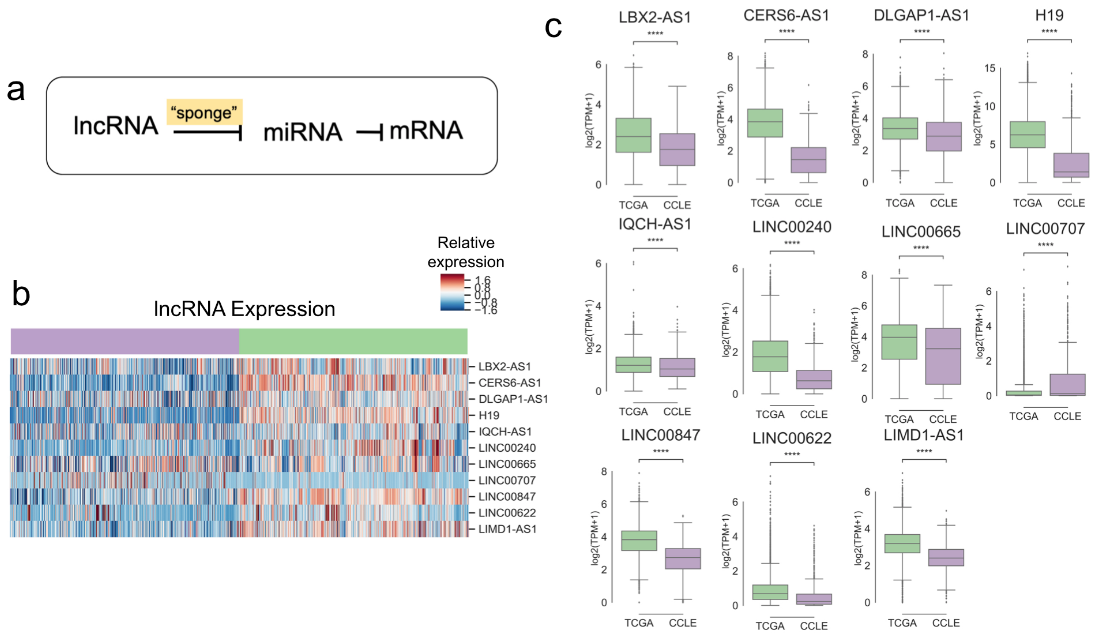


Fig. 3. Long non-coding RNA associated with 4 cancer driver pathways are significantly underexpressed in cancer cell lines. a) In the "sponge" model of lncRNA gene expression regulation, lncRNA competitively inhibit miRNA which would otherwise be responsible for inhibiting mRNA. b) Heatmap showing expression of 11 lncRNAs with miRNA-dependent associations to protein coding genes in the four cancer driver pathways. Samples shown are a random subset with equal representation from each dataset in each disease. c) Boxplots showing expression of the 11 lncRNAs associated with four cancer driver pathways. All samples from both datasets are shown. (Mann-Whitney significance test; * pvalue < 0.05, ** pvalue < 0.01, *** pvalue < 0.001)

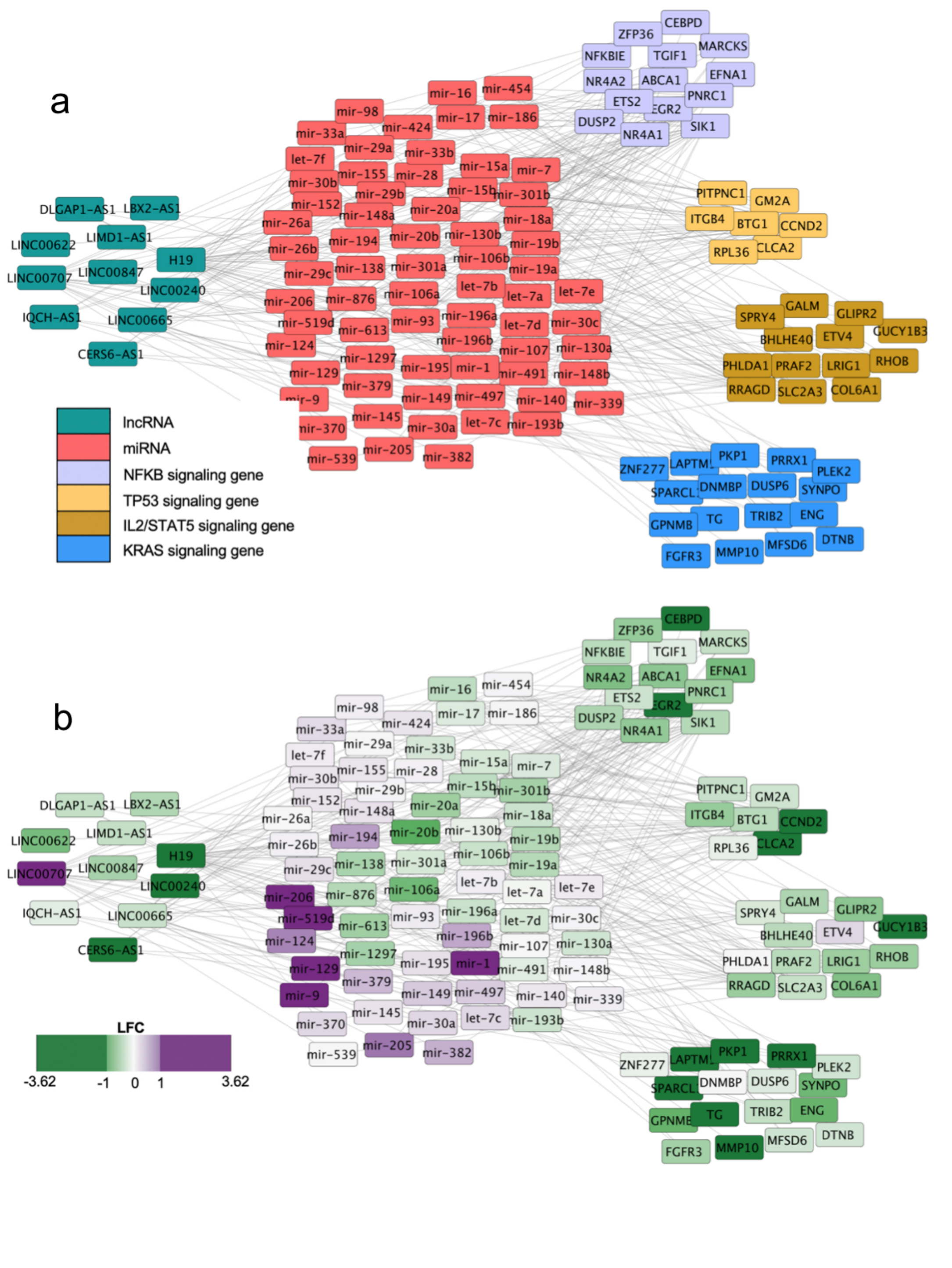


Fig. 4. Dysregulation of lncRNA-miRNA regulatory network causes downregulation of key cancer driver pathways in tumor derived cell lines. a) Types of genes are identified

by color and positioning in the Cytoscape graph. Gene interactions from miRNet databases are denoted by grey lines. lncRNA are on the left, miRNA in the center and differentially expressed protein coding genes from each of the four cancer driver pathways are on the right side of the graph. b) Positive LFC (purple) denotes higher expression in CCLE. Negative LFC (green) denotes higher expression in TCGA.

D. Tumor-derived organoids and patient-derived xenografts have lowest tumor correlation in KRAS signaling.

Finally, we investigated how well two other types of cancer models (organoids and patient-derived xenografts) recapitulate cancer driver pathway expression as compared to primary tumors.

Tumor-derived organoids have been shown to recapitulate several properties of the primary tumor including cellular heterogeneity, hypoxic gradient, activated molecular pathways and specific molecular aberrations such as mutations and fusions[43,44]. Notably, organoid cultures have the capacity to maintain the cancer stem cell population of the original tumor[44]. In contrast, patient-derived xenograft (PDX) models are intrinsically capable of recapitulating host-tumor interactions such as maintenance of substantial populations of stromal cells[45]. However, PDX cancer models are comparatively time intensive and costly, motivating significant advances in cancer organoid modeling technology in the last decade.

We investigated RNA sequencing data from both model types in order to determine which model best recapitulates the expression of KRAS, TP53, NFKB and IL2/STAT5 signaling. We hypothesized that both types of models would maintain the stem cell population lost in the 2D cell culture, thus preserving the biological activity of the lncRNA-miRNA regulatory network and the four cancer driver pathways. To test this hypothesis, we leveraged publicly available gene expression data from two recent publications on tumor-derived organoids from bladder cancer (n organoids=34, n tumors=8) and liver cancer (n organoids=15, n tumors=10)[46,47]. We also utilized rhabdomyosarcoma PDX gene expression data (n=7) and disease-matched patient samples (n=40) from the St. Jude Cloud[48], accessed through the UC Santa Cruz Treehouse cancer compendium v10 (treehousegenomics.soe.ucsc.edu/public-data/; Supplementary Table 4).

We performed Mann-Whitney significance test to compare the expression of representative genes from each cancer driver pathway between tumor and organoid or PDX (Fig. 5a,c,e). As compared to cancer cell lines, both model types are overall more similar to their associated tumors. However, in both organoid models, the *KRAS* gene is significantly under-expressed compared to tumors, the same pattern observed in cancer cell lines. The *NFKB* gene is also under-expressed in liver organoids, but not bladder organoids. Interestingly, PDX models show slight overexpression of the *KRAS* and *TP53* genes as compared to tumors.

To look more broadly at pathway-level expression differences, we calculated pairwise correlation between all tumor and organoid or PDX samples in each dataset (Fig. 5b,d,f). We noted disease-specific in pathway correlation in the organoid models. Notably, IL2/STAT5 Signaling has highest correlation in bladder cancer organoids (mean=0.7) but second to lowest expression in liver cancer organoids (mean=0.25). This indicates that the degree of success for organoid cancer modeling may be disease-specific, and more work is needed to ensure consistent recapitulation of cancer signaling. Additionally, the significant underexpression of *NFKB* in liver organoids indicates a potential loss of intrinsic immune signaling in some types of organoids.

In keeping with our previous observations, KRAS Signaling has the lowest overall correlation between tumors and organoids or PDX models. Even in the rhabdomyosarcoma PDX dataset, in which the other 3 cancer driver pathways have greater than 0.67 mean correlation, the KRAS Signaling pathway has only 0.59 mean correlation. The KRAS Signaling pathway appears to be uniquely dependent on the original host microenvironment, and is not fully recapitulated even in an animal model such as a PDX. This may help to explain the marked lack of effective RAS inhibitor therapeutics[49].

Overall, these analyses suggest that because tumor-derived organoids and PDX models better recapitulate cell type heterogeneity and cellular interactions of primary tumors, gene signaling and cancer driver pathway activity are also more accurately preserved than in 2D cancer cell lines. However, additional research efforts are needed towards including immune signaling in organoids, as well as efforts to ensure KRAS signaling in both model types is accurately recapitulated.

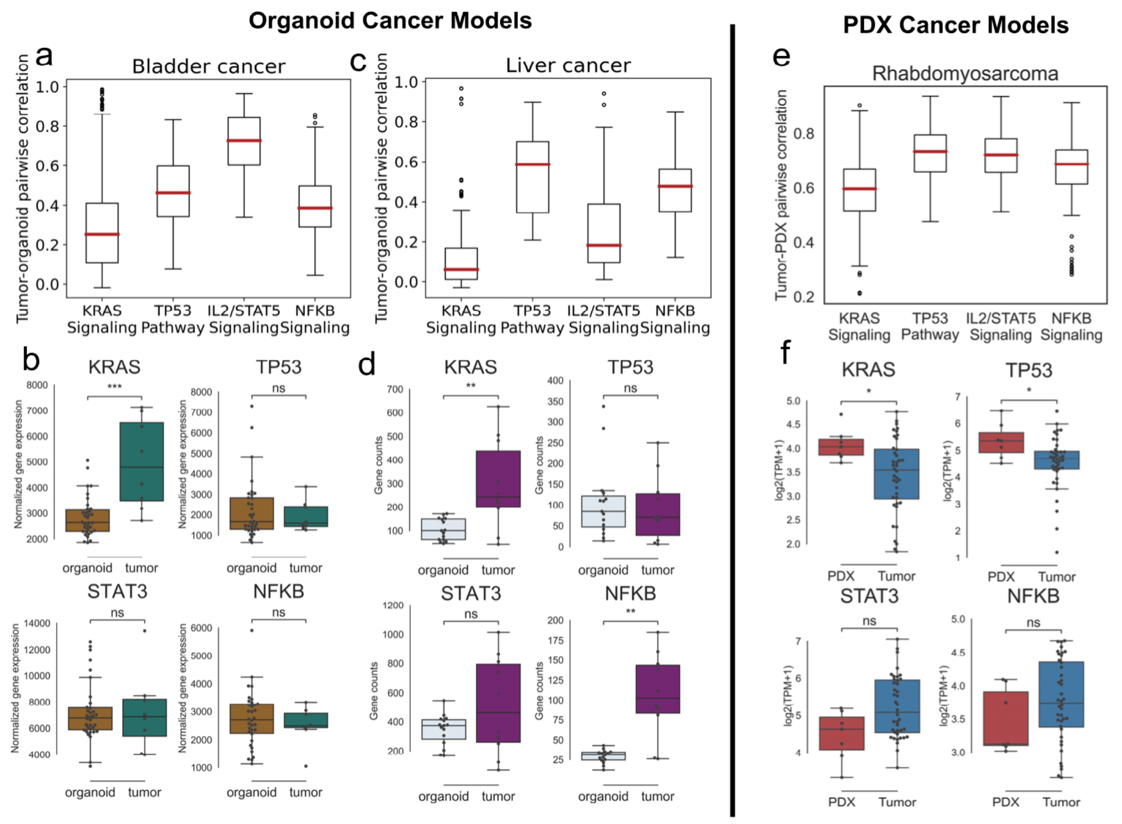


Fig. 5. Cancer driver pathway expression in 3D cancer-derived organoids and PDX models. a) Normalized gene expression from bladder cancer samples (n=36) and matched cancer-derived organoids (n=8). b) Pairwise correlation of all bladder tumor and organoid samples using all genes in each pathway gene set. c) Gene counts from liver cancer samples (n=10) and matched cancer-derived organoids (n=15). d) Pairwise correlation of all liver tumor and organoid samples using all genes in each pathway gene set. e) Normalized gene expression from rhabdomyosarcoma cancer samples (n=40) and sarcoma-derived PDX models (n=7). f) Pairwise correlation of all sarcoma cancer and PDX samples using all genes in each pathway gene set. (Mann-Whitney significance test; * pvalue < 0.05, ** pvalue < 0.01, *** pvalue < 0.001)

Discussion

The ability to model and manipulate cancer cells has empowered therapeutic discovery since the derivation of the first cancer cell line[50]. Two dimensional cancer cell cultures have enabled researchers to discover how cancers arise, characterize cancer cell types and growth patterns, and identify effective pharmaceuticals through drug screens[1]. Today, personalized tumor-derived 2D and 3D cultures are increasingly in use for identification of precision therapies for individual patients[51–53]. The rise of precision medicine and small molecule inhibitor development brings an urgent need for molecular characterization of the cell culture models used widely for therapeutic development. Large-scale genomic efforts such as The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) have enabled comprehensive comparison of cancer cell cultures and primary tumors.

Here, we provide a comparative multi-omic analysis of cancer cell lines and primary tumors by leveraging several types of genomic data from large public compendia. Our gene expression analysis reveals reduced expression of key cancer driver pathways including KRAS signaling, TP53 pathway, IL2/STAT5 signaling and NFKB signaling in cancer cell lines. These results are recapitulated in comparative analysis of protein levels for key proteins from each pathway. Our analysis indicates the need for caution when interpreting *in vitro* preclinical testing results of inhibitors targeting these pathways.

In fact, the consequences of preclinical testing in cell lines against these pathways have already been felt. For example, during preclinical testing several MDM2/TP53 small molecule inhibitors displayed potent cancer cell line inhibitory activity, but could only achieve partial tumor regression in xenograft models[54].

Subsequent optimization of these compounds led to success in clinical trials, but initial results in cell lines did not predict *in vivo* results. Targeting both wild-type and mutant KRAS in cancer has been notably unsuccessful; in particular, several high-throughput screens of KRAS mutant cancer cell lines identified compounds which subsequently only partially reduced tumor volumes in xenograft models[49,55]. In contrast, targeting the cholesterol biosynthesis pathway, which is not differentially expressed in cancer cell lines by our analysis, has been promising both *in vitro* and *in vivo* through the usage of the statin family of drugs[56,57]. In addition, targeted inhibition of cyclin-dependent kinases (CDKs) by agents such as PD-0332991/palbociclib has been promising in phase I and II clinical trials, and this drug was originally identified in cancer cell lines[58,59]. CDKs are active during cell cycle entry from G0 and during the G2M checkpoint pathway, which were not differentially expressed in cancer cell lines by our analysis. We note that apart from the four cancer driver pathways identified as dysregulated in cell lines, preclinical testing in cell lines can in many cases reliably predict *in vivo* responses.

To identify potential causes of dysregulation of cancer driver signaling in cell lines, we analyzed the expression of lncRNA and miRNA implicated in regulation of these pathways. In cell lines, we found reduced expression of a set of lncRNA predicted to regulate a downstream network of regulatory miRNA, which are in turn overexpressed. Several of these miRNA are directly involved in specific inhibition of these cancer driver pathways, linking their overexpression to the observed reduced expression of cancer driver pathways.

We speculate that our results may indicate a partial loss of cancer stem cells in cancer cell culture due to the presence of serum in culture media. All CCLE cell lines were cultured in RPMI or DMEM media with 10% fetal bovine serum[60]. It is well known that the presence of serum in culture media encourages cellular differentiation, and cancer cell lines grown in serum-free conditions contain larger populations of cancer stem cells[61,62]. This hypothesis is supported by the markedly lower expression of stem-cell-specific lncRNAs in cell culture, and higher expression of pro-differentiation miRNAs. Because cancer stem cell populations are known for their chemoresistance and even small populations are thought to be capable of tumor recurrence[63–65], it is essential that preclinical models accurately model the response of cancer stem cells to potential therapeutics.

We investigated the recently identified TCGA-110-CL cancer cell line panel[7], which has higher overall cell line-tumor gene expression correlation. The same four cancer driver pathways were identified by an SVM comparing these cell lines to disease-matched primary tumors, and in particular the KRAS pathway had the lowest overall cell line-tumor correlation. This indicates that even this subset of cancer cell lines does not fully recapitulate the cancer driver signaling of primary tumors and studies on these pathways must be interpreted with caution.

We find that in certain tumor types, 3D cancer-derived organoids and PDX models, which are known to better recapitulate tumor heterogeneity including cancer stem cell populations, are more similar to their primary tumor counterparts with respect to expression of the four cancer driver pathways. However, future work needs to focus on the preservation of cancer-immune microenvironment in organoids. Notably, all three cancer models we studied had lower model-tumor correlations in the KRAS pathway than the other three cancer driver pathways. KRAS signaling depends on certain signal transduction and cell-cell interaction events such as ligand-dependent

EGFR activation[66] and TWIST-dependent senescence bypass[67], which may be uniquely dependent on the original tumor microenvironment. This indicates that future model development must emphasize preservation of intrinsic KRAS signaling in order to develop effective anti-KRAS therapeutics. Indeed, genetic manipulation via overexpression or knockdown of certain epigenetic regulators such as lncRNA or miRNA may improve the generation of cancer cell lines or organoids from primary tumors.

Taken together, our results underscore the need for caution when interpreting preclinical cancer testing results in multiple model types, and point to specific signaling networks which can serve as litmus tests for the accuracy of past and future cancer laboratory models. We suggest several potential solutions to improve the efficacy of tumor-derived cell lines and organoids. Cancer cell culture in serum-free conditions may improve the maintenance of tumor stem cell populations and reverse the dysregulation of important regulatory gene networks. Specific efforts to model the immune microenvironment in cancer-derived organoids may improve cancer driver pathway expression related to the tumor microenvironment. A potential solution may be genetic manipulation of tumor-derived models with an emphasis on preserving or rescuing the intrinsic cancer driver pathway expression which is most at risk for dysregulation. Overall, this study provides much-needed genomics-based guidelines for future preclinical cancer model development.

Methods

RNA sequencing data

Gene expression transcripts per million (TPM) matrices from TCGA (n samples = 10535) and CCLE (n samples = 933) were downloaded from the UCSC Xena browser. These data were processed uniformly through the TOIL UCSC RNA sequencing data processing pipeline to remove technical batch effects[68]. Both datasets were normalized by log2(TPM+1) and duplicate genes were averaged. Genes not expressed in 80% of samples were removed, and 20% of the lowest varying remaining genes were removed, leaving 46865 remaining genes. Both datasets were subset to the 19 overlapping cancer types for subsequent analysis (BRCA, LUSC, LIHC, DLBC, THCA, PRAD, OV, STAD, BLCA, KIRC, UCEC, COAD, SARC, CESC, SKCM, PAAD, HNSC, ESCA, GBM). All heatmaps use a random subset of samples from each dataset with equal numbers from each disease.

micro-RNA sequencing data

TCGA micro-RNA (miRNA) Illumina sequencing read counts data was downloaded from the Genomic Data Commons Data Portal[69]. CCLE Nanostring probe miRNA quantification data was downloaded from the Broad Institute CCLE database: https://portals.broadinstitute.org/ccle/data[70]. For dataset comparability, miRNA naming formats were harmonized and duplicates averaged. Because different miRNA sequencing methods were used in each dataset, ComBat was used to batch correct the data[25]. Pre- and post-batch effect correction data was then log2(count+1) normalized for downstream visualization and analysis. Supplementary Figure 5 shows pre- and post-batch effect correction expression distributions of several housekeeping genes to validate successful correction[71,72].

RPPA data

Level 4 Reverse Phase Protein Array (RPPA) data for the TCGA and CCLE samples were downloaded from the The Cancer Proteome Atlas (TCPA) portal (https://tcpaportal.org/tcpa/download.html and http://tcpaportal.org/mclp/#/download). Both datasets were subset to the 16 overlapping cancer types for subsequent analysis (BLCA, BRCA, COAD, DLBC, HNSC, KIRC, LGG, LIHC, LUAD, OV, PAAD, PRAD, SARC, SKCM, STAD).

Organoid and PDX RNA sequencing data

RNA sequencing normalized counts data from 36 bladder cancer samples and 8 bladder cancer-derived organoid samples were downloaded from accession GSE103990[46]. RNA sequencing gene counts data from 10 liver cancer samples and 15 liver cancer-derived organoid samples were downloaded from accession GSE84073[47]. Gene expression TPM data from sarcoma PDX samples (n=7) and disease-matched patient samples (n=40) were downloaded from the UC Santa Cruz Treehouse cancer compendium v10 (treehousegenomics.soe.ucsc.edu/public-data/), and normalized by log2(TPM+1). The cancer compendium contains gene expression data generated from the raw data in the St. Jude Cloud[48].

Statistical gene selection via support vector machine

In Python (v3.6.8), the sklearn module (v0.21.2) was used with linear kernel to train a support vector machine (SVM) on 50 random 80/20 splits of the merged TCGA-CCLE gene expression dataset. The top 10% of genes from each training run based on their feature weights coefficients were merged in a non-duplicate manner, resulting in 1858 genes. Gene set enrichment from this analysis revealed 26/100 enriched pathways were immune-related, so a non-redundant immune gene list was created by merging all genes from the 26 enriched immune pathways (Supplementary Table 1). A second SVM analysis was conducted on a set of 50 random 80/20 splits of the merged TCGA-CCLE gene expression dataset with the immune-related gene list removed, and the top 10% of genes from each run based on feature importance coefficients were merged, resulting in 1854 genes which were used in all downstream analysis. We verified that the cancer driver pathway signal identified in the second SVM analysis is not related to immune signaling by calculating pairwise correlation between SVM-identified cancer driver genes and the immune genes which were removed from the analysis (mean=0.1, Supplementary Fig. 1).

Statistical methods

Expression comparisons of individual genes between datasets were performed using a two-sided Mann-Whitney significance test with significance defined as pvalue < 0.05. Significantly enriched gene sets were identified using gene set enrichment analysis[9] with FDR q-value < 0.05.

Supplementary Figures

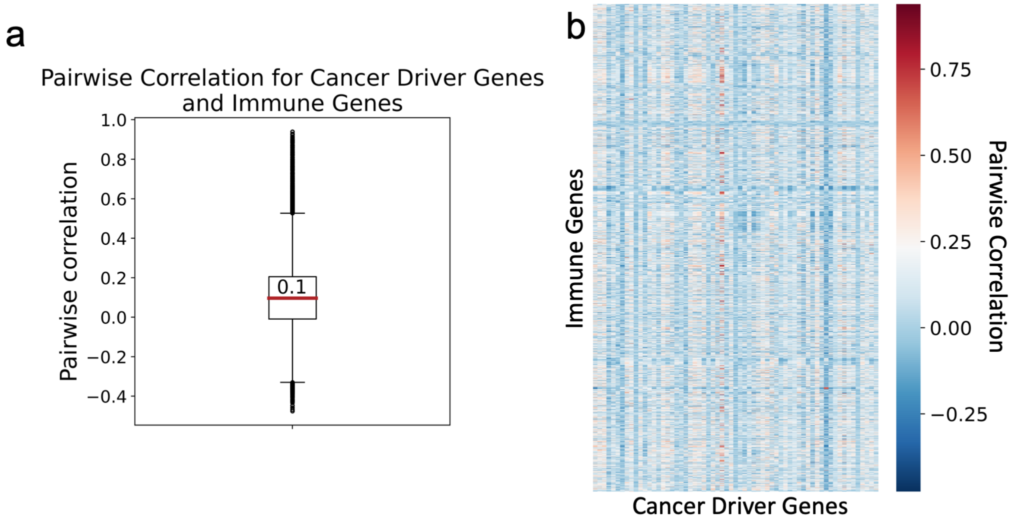


Figure S1. Pairwise correlation of expression of immune genes and cancer driver genes. a) Pairwise correlation between the immune genes removed from the SVM analysis and the cancer driver genes identified by the SVM analysis (mean correlation=0.1). b) Heatmap showing pairwise correlation scores between immune genes and cancer driver genes.

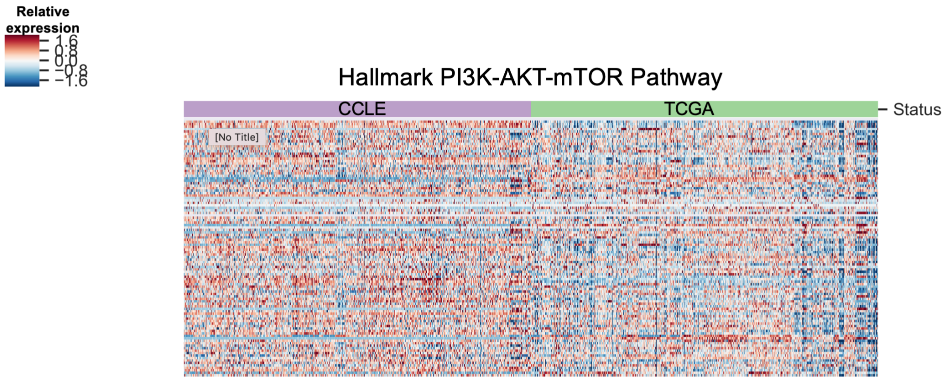


Figure S2. Hallmark PI3K-AKT-mTOR Pathway shows non-significant expression change between TCGA and CCLE. Heatmap showing expression of genes in the PI3K-AKT-mTOR pathway in cancer cell lines and tumors. Samples shown are a random subset with equal representation from each dataset in each disease.

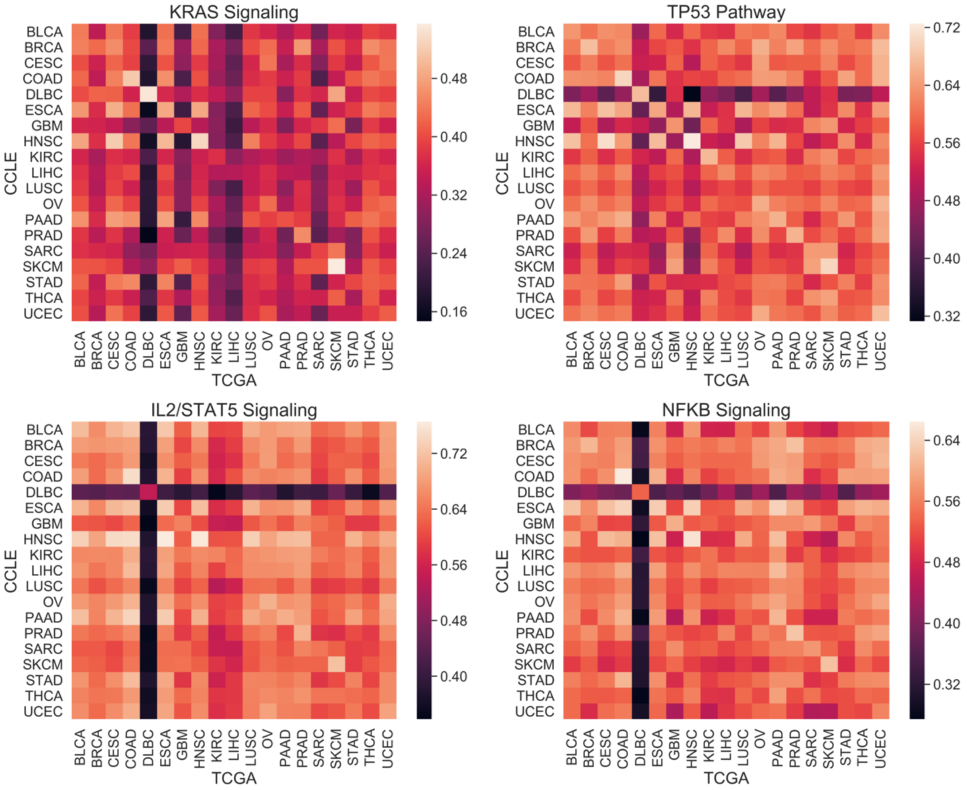


Figure S3. Heatmaps showing mean correlation scores between all disease types in each cancer driver pathway.

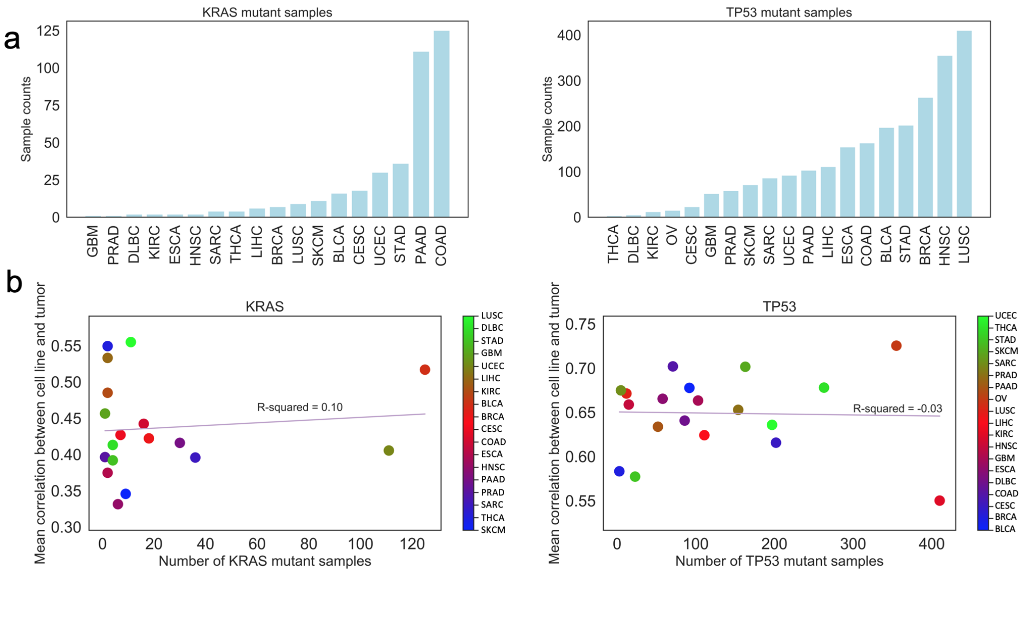


Figure S4. Relationship between KRAS or TP53 mutational burden and tumor-cell line correlation by disease type. a) Barplot of KRAS mutant TCGA samples and TP53 mutant TCGA samples by disease type. b) Mean correlation between cell line and tumor in the KRAS pathway (left) and TP53 pathway (right) colored by disease.

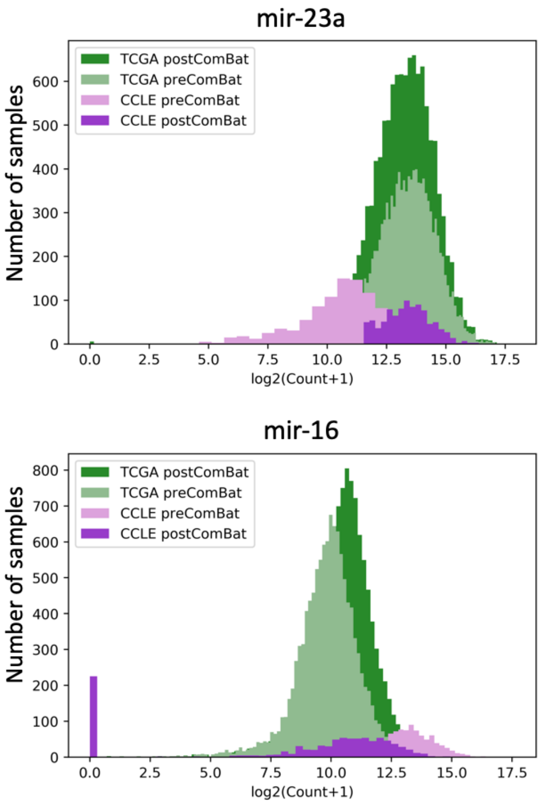


Figure S5. Expression of housekeeping miRNA mir-23a and mir-16 pre- and post-ComBat correction in all CCLE or TCGA samples.

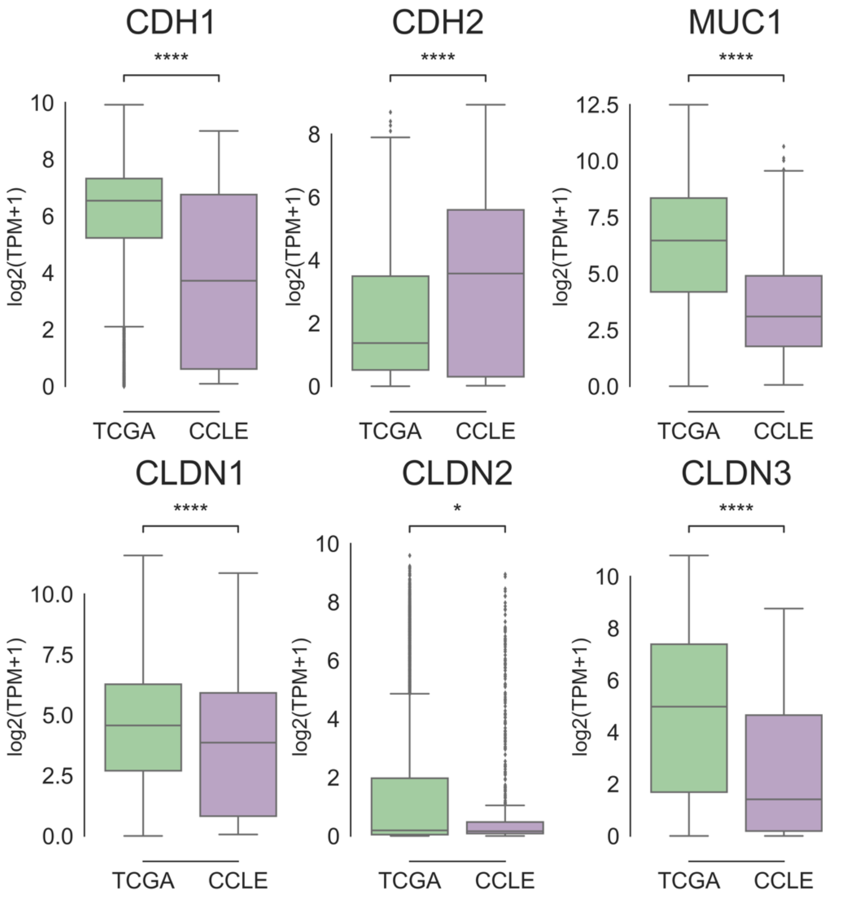


Figure S6. Expression of cell type markers in TCGA vs. CCLE. Boxplots showing expression of the E-cadherin/*CDH1* (an epithelial cell marker) and N-cadherin/*CDH2*, *MUC1*, and claudins *CLDN1, CLDN2, CLDN3* (mesenchymal cell markers) in TCGA and CCLE. All samples from both datasets are shown. (Mann-Whitney significance test; * pvalue < 0.05, **** pvalue < 0.0001)

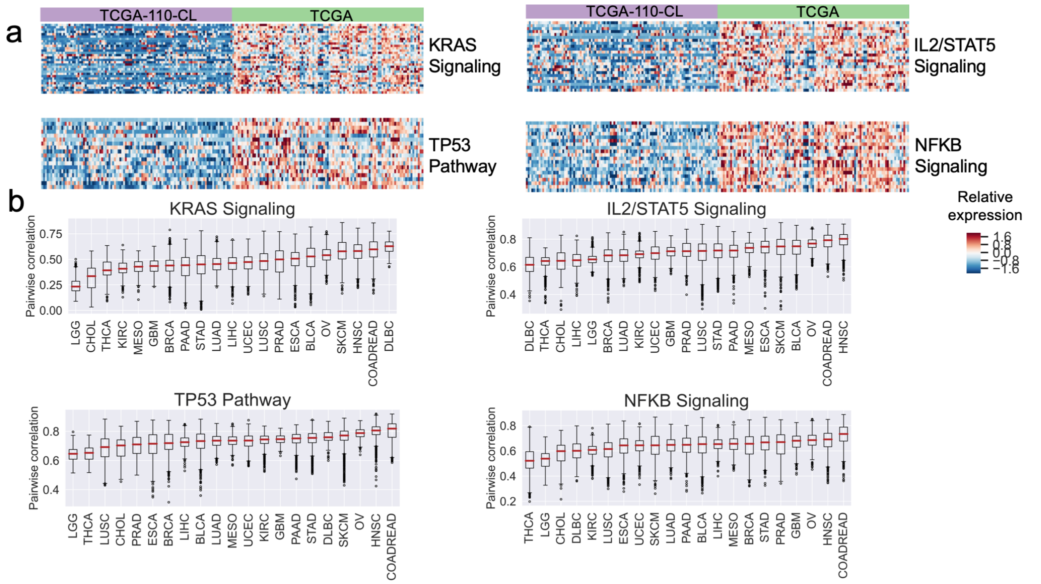


Figure S7. Expression of cancer driver pathways in TCGA-110-CL. a) Heatmaps showing expression of SVM-identified genes in 4 cancer driver pathways in TCGA-110-CL cell lines as compared to TCGA samples. Samples shown are a random subset with equal representation from each dataset in each disease. b) Boxplots show pairwise Spearman correlation scores between all TCGA-110-CL and TCGA RNA-seq samples in each disease type, for all 4 cancer driver pathways. Plots are sorted by mean correlation.

References

1. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).

2. Ross, D. T. & Perou, C. M. A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. *Dis. Markers* **17**, 99–109 (2001).

3. Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* **4**, 2126 (2013).

4. Li, H. *et al.* Genomic analysis of head and neck squamous cell carcinoma cell lines and human tumors: a rational approach to preclinical model selection. *Mol. Cancer Res.* **12**, 571–582 (2014).

5. Mouradov, D. *et al.* Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res.* **74**, 3238–3247 (2014).

6. Goodspeed, A., Heiser, L. M., Gray, J. W. & Costello, J. C. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Mol. Cancer Res.* **14**, 3–13 (2016).

7. Yu, K. *et al.* Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* **10**, 3574 (2019).

8. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

9. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

10. Fan, Y. & Xia, J. miRNet—Functional Analysis and Visual Exploration of miRNA–Target Interactions in a Network Context. in *Computational Cell Biology: Methods and Protocols* (eds. von Stechow, L. & Santos Delgado, A.) 215–233 (Springer New York, 2018).

11. Li, J. *et al.* Characterization of Human Cancer Cell Lines by Reverse-phase Protein Arrays. *Cancer Cell* **31**, 225–239 (2017).

12. Li, J. *et al.* TCPA: a resource for cancer functional proteomics data. *Nat. Methods* **10**, 1046–1047 (2013).

13. Eser, S., Schnieke, A., Schneider, G. & Saur, D. Oncogenic KRAS signalling in pancreatic cancer. *Br. J. Cancer* **111**, 817–822 (2014).

14. Eser, S. *et al.* Selective requirement of PI3K/PDK1 signaling for Kras oncogene-driven pancreatic cell plasticity and cancer. *Cancer Cell* **23**, 406–420 (2013).

15. Sasaki, C. Y., Barberi, T. J., Ghosh, P. & Longo, D. L. Phosphorylation of RelA/p65 on Serine 536 Defines an IκBα-independent NF-κB Pathway. *J. Biol. Chem.* **280**, 34538–34547 (2005).

16. Wang, J. *et al.* Activation of NF-{kappa}B by TMPRSS2/ERG Fusion Isoforms through Toll-Like Receptor-4. *Cancer Res.* **71**, 1325–1333 (2011).

17. Gemignani, M. L. *et al.* Role of KRAS and BRAF gene mutations in mucinous ovarian carcinoma. *Gynecol. Oncol.* **90**, 378–381 (2003).

18. Zhu, J. *et al.* Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth. *Nature* **525**, 206–211 (2015).

19. Liz, J. & Esteller, M. lncRNAs and microRNAs with a role in cancer development. *Biochim. Biophys. Acta* **1859**, 169–176 (2016).

20. Militello, G. *et al.* Screening and validation of lncRNAs and circRNAs as miRNA sponges. *Brief. Bioinform.* **18**, 780–788 (2017).

21. Paraskevopoulou, M. D. & Hatzigeorgiou, A. G. Analyzing MiRNA–LncRNA Interactions. in *Long Non-Coding RNAs: Methods and Protocols* (eds. Feng, Y. & Zhang, L.) 271–286 (Springer New York, 2016).

22. Gregory, R. I. & Shiekhattar, R. MicroRNA biogenesis and cancer. *Cancer Res.* **65**, 3509–3512 (2005).

23. Huang, Y. *et al.* Biological functions of microRNAs: a review. *J. Physiol. Biochem.* **67**, 129–139 (2011).

24. Cai, Y., Yu, X., Hu, S. & Yu, J. A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* **7**, 147–154 (2009).

25. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

26. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular

interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

27. Li, D. *et al.* Analysis of MiR-195 and MiR-497 expression, regulation and role in breast cancer. *Clin. Cancer Res.* **17**, 1722–1730 (2011).

28. Xu, Q. *et al.* A regulatory circuit of miR-148a/152 and DNMT1 in modulating cell transformation and tumor angiogenesis through IGF-IR and IRS1. *J. Mol. Cell Biol.* **5**, 3–13 (2013).

29. Jansson, M. D., Damas, N. D., Lees, M., Jacobsen, A. & Lund, A. H. miR-339-5p regulates the p53 tumor-suppressor pathway by targeting MDM2. *Oncogene* **34**, 1908–1918 (2015).

30. Yeung, M. L. *et al.* Roles for microRNAs, miR-93 and miR-130b, and tumor protein 53-induced nuclear protein 1 tumor suppressor in cell growth dysregulation by human T-cell lymphotrophic virus 1. *Cancer Res.* **68**, 8976–8985 (2008).

31. Deng, X., Zhao, Y. & Wang, B. miR-519d-mediated downregulation of STAT3 suppresses breast cancer progression. *Oncol. Rep.* **34**, 2188–2194 (2015).

32. Lu, Z. *et al.* miR-301a as an NF-κB activator in pancreatic cancer cells. *EMBO J.* **30**, 57–67 (2011).

33. Peng, F. *et al.* H19/let-7/LIN28 reciprocal negative regulatory circuit promotes breast cancer stem cell maintenance. *Cell Death Dis.* **8**, e2569 (2017).

34. Sweetman, D. *et al.* Specific requirements of MRFs for the expression of muscle specific microRNAs, miR-1, miR-206 and miR-133. *Dev. Biol.* **321**, 491–499 (2008).

35. Goljanek-Whysall, K. *et al.* Regulation of multiple target genes by miR-1 and miR-206 is pivotal for C2C12 myoblast differentiation. *J. Cell Sci.* **125**, 3590–3600 (2012).

36. Wang, L. *et al.* Identifying the crosstalk of dysfunctional pathways mediated by lncRNAs in breast cancer subtypes. *Mol. Biosyst.* **12**, 711–720 (2016).

37. Yang, L. *et al.* Genome-wide identification of long non-coding RNA and mRNA profiling using RNA sequencing in subjects with sensitive skin. *Oncotarget* **8**, 114894–114910 (2017).

38. Yang, S. *et al.* Construction of differential mRNA-lncRNA crosstalk networks based on ceRNA hypothesis uncover key roles of lncRNAs implicated in esophageal squamous cell carcinoma. *Oncotarget* **7**, 85728–85740 (2016).

39. Salvador, E., Burek, M. & Förster, C. Y. Tight Junctions and the Tumor Microenvironment. *Curr. Pathobiol. Rep.* **4**, 135–145 (2016).

40. Ma, L. *et al.* miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nat. Cell Biol.* **12**, 247–256 (2010).

41. Yang, J. & Weinberg, R. A. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Dev. Cell* **14**, 818–829 (2008).

42. Redmer, T. *et al.* E-cadherin is crucial for embryonic stem cell pluripotency and can replace OCT4 during somatic cell reprogramming. *EMBO Rep.* **12**, 720–726 (2011).

43. Gao, D. *et al.* Organoid cultures derived from patients with advanced prostate cancer. *Cell* **159**, 176–187 (2014).

44. Hubert, C. G. *et al.* A Three-Dimensional Organoid Culture System Derived from Human Glioblastomas Recapitulates the Hypoxic Gradients and Cancer Stem Cell Heterogeneity of Tumors Found In Vivo. *Cancer Res.* **76**, 2465–2477 (2016).

45. Nicolle, R. *et al.* Pancreatic Adenocarcinoma Therapeutic Targets Revealed by Tumor-Stroma Cross-Talk Analyses in Patient-Derived Xenografts. *Cell Rep.* **21**, 2458–2470 (2017).

46. Lee, S. H. *et al.* Tumor Evolution and Drug Response in Patient-Derived Organoid Models of Bladder Cancer. *Cell* **173**, 515–528.e17 (2018).

47. Broutier, L. *et al.* Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. *Nat. Med.* **23**, 1424–1435 (2017).

48. Chen, X. *et al.* Targeting oxidative stress in embryonal rhabdomyosarcoma. *Cancer Cell* **24**, 710–724 (2013).

49. Wang, Y., Kaiser, C. E., Frett, B. & Li, H.-Y. Targeting mutant KRAS for anticancer therapeutics: a review of novel small molecule modulators. *J. Med. Chem.* **56**, 5219–5230 (2013).

50. Gillet, J.-P., Varma, S. & Gottesman, M. M. The clinical relevance of cancer cell lines. *J. Natl. Cancer Inst.* **105**, 452–458 (2013).

51. Phan, N. *et al.* A simple high-throughput approach identifies actionable drug sensitivities in patient-derived tumor organoids. *Commun Biol* **2**, 78 (2019).

52. Pauli, C. *et al.* Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. *Cancer Discov.* **7**, 462–477 (2017).

53. Kodack, D. P. *et al.* Primary Patient-Derived Cancer Cells and Their Potential for Personalized Cancer Patient Care. *Cell Rep.* **21**, 3298–3309 (2017).

54. Zhao, Y., Aguilar, A., Bernard, D. & Wang, S. Small-molecule inhibitors of the MDM2-p53 protein-protein interaction (MDM2 Inhibitors) in clinical trials for cancer treatment. *J. Med. Chem.* **58**, 1038–1052 (2015).

55. Kempf, E., Rousseau, B., Besse, B. & Paz-Ares, L. KRAS oncogene in lung cancer: focus on molecularly driven clinical trials. *Eur. Respir. Rev.* **25**, 71–76 (2016).

56. Sleijfer, S., van der Gaast, A., Planting, A. S. T., Stoter, G. & Verweij, J. The potential of statins as part of anti-cancer treatment. *Eur. J. Cancer* **41**, 516–522 (2005).

57. Cho, S.-J. *et al.* Simvastatin induces apoptosis in human colon cancer cells and in tumor xenografts, and attenuates colitis-associated colon cancer in mice. *Int. J. Cancer* **123**, 951–957 (2008).

58. Fry, D. W. *et al.* Specific inhibition of cyclin-dependent kinase 4/6 by PD 0332991 and associated antitumor activity in human tumor xenografts. *Mol. Cancer Ther.* **3**, 1427–1438 (2004).

59. Cicenas, J. & Valius, M. The CDK inhibitors in cancer research and therapy. *J. Cancer Res. Clin. Oncol.* **137**, 1409–1418 (2011).

60. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

61. Chin, A. C. P., Padmanabhan, J., Oh, S. K. W. & Choo, A. B. H. Defined and serum-free media support undifferentiated human embryonic stem cell growth. *Stem Cells Dev.* **19**, 753–761 (2010).

62. Lee, J. *et al.* Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9**, 391–403 (2006).

63. Clarke, M. F. *et al.* Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells. *Cancer Res.* **66**, 9339–9344 (2006).

64. Clevers, H. The cancer stem cell: premises, promises and challenges. *Nat. Med.* **17**, 313–319 (2011).

65. Sakaguchi, M. *et al.* miR-137 Regulates the Tumorigenicity of Colon Cancer Stem Cells through the Inhibition of DCLK1. *Mol. Cancer Res.* **14**, 354–362 (2016).

66. Ardito, C. M. *et al.* EGF receptor is required for KRAS-induced pancreatic tumorigenesis. *Cancer Cell* **22**, 304–317 (2012).

67. Lee, K. E. & Bar-Sagi, D. Oncogenic KRas suppresses inflammation-associated senescence of pancreatic ductal cells. *Cancer Cell* **18**, 448–458 (2010).

68. Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).

69. Chu, A. *et al.* Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Res.* **44**, e3 (2016).

70. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* (2019) doi:10.1038/s41586-019-1186-3.

71. Masè, M. *et al.* Selection of reference genes is critical for miRNA expression analysis in human

cardiac tissue. A focus on atrial fibrillation. *Sci. Rep.* **7**, 41127 (2017).

72. Shen, Y. *et al.* Identification of miR-23a as a novel microRNA normalizer for relative quantification in human uterine cervical tissues. *Exp. Mol. Med.* **43**, 358–366 (2011).

Chapter 5: Current technologies for modeling embryonic effects of histone H3 K27M mutation

5.1 Chapter Introduction

Much of the research on H3 K27M mutant glioma has been appropriately focused on understanding and treating the mature tumors. However, because my work (Chapter 3) and the work of others points to a K27M-associated stall in early neural cell differentiation, a better understanding of the developmental origins of this tumor may aid treatment and prevention. Some excellent experimental models of K27M gliomagenesis have already been developed; for example, Dr. Suzanne Baker's lab showed that induced K27M expression in neonatal mouse hindbrain accelerated DIPG formation[34]. But because it seems that the K27M mutation can arise prenatally[35] while the tumor itself does not arise until a median patient age of six[36], this suggests that inducible mouse models may not fully recapitulate the long-term epigenetic and developmental effects of the K27M mutation on the embryonic human brain. Additionally, previous studies have not reached a conclusion on the developmental timing and cell of origin for the K27M mutation event, which would contribute significantly to understanding the pathogenesis of the mature tumor. In a 2019 review, leading pediatric neuro-oncologists Drs. Mariella Filbin and Michelle Monje concluded that better biological understanding and therapies for K27M glioma and other developmental pediatric tumors, will come from studies on early developmental and epigenetic cell states[6].

Accordingly, I took advantage of the Haussler lab expertise in modelling early human brain development with human embryonic stem cell derived cerebral organoids. Because the cerebral organoid assay developed in the Haussler lab can currently model up to the first 70 days of embryogenesis (with ongoing improvements in durability), I was motivated to create a K27M-inducible version of this organoid assay to study the effects of K27M expression during human embryogenesis. Notable advantages of this novel assay are that it is in human cells, it is possible to induce K27M expression at any time point, and organoid culture is relatively scalable and cost-effective.

Because of the COVID-19 pandemic, some results from this project have been delayed. Here I report the successful development of a novel H3 K27M inducible human cerebral organoid model, and preliminary results which indicate K27M-associated epigenetic and transcriptomic effects, as well as potential observation of cellular differentiation stall in K27M cells.

I designed and led this project. The contributions of others are as follows: I trained and mentored Marissa Chen and Liam Tran as undergraduate assistants on this project. Marissa Chen helped write the Introduction and Methods sections and helped with the stem cell culture maintenance, lentiviral transduction, digital droplet PCR and Western blots. Liam Tran helped with the organoid culture maintenance, tissue slicing on the cryostat, and immunofluorescence staining. Dr. Sameer Agnihotri provided the viral vector reagents. Dr. Sofie Salama provided scientific oversight and mentorship.

This manuscript reflects the current state of this work. Future directions are described in Chapter 6.

Development of a 3D cerebral organoid model of the origins of histone H3 K27M mutant diffuse midline glioma

Lauren M. Sanders[1], Marissa Chen[2], Liam Tran[1], David Haussler[1,3], Sofie R. Salama[2,3]

[1]Department of Biomolecular Engineering, University of California Santa Cruz

[2]Department of Molecular, Cell and Developmental Biology, University of California Santa Cruz

[3]Howard Hughes Medical Institute,University of California Santa Cruz

Abstract

Histone H3 K27M mutant diffuse midline glioma are extremely aggressive pediatric brain tumors with universally poor prognosis and no known effective treatments. This histone mutation occurs in 80% of pediatric brainstem gliomas, and confers a significant survival disadvantage as compared to pediatric gliomas without the mutation. The H3K27M mutation causes global loss of the H3K27 trimethyl transcriptional repressive mark, resulting in widespread transcriptional dysregulation of developmental genes. This mutation is thought to be oncogenic only during specific developmental windows where correct transcriptional regulation is particularly important for neural cell differentiation. However, the developmental timing and cell of origin for the H3K27M mutation event and subsequent gliomagenesis have not been conclusively identified. Here we report the development of a novel human embryonic stem cell derived cerebral organoid model with inducible H3K27M expression. This model recapitulates known epigenetic and transcriptional changes associated with the presence of the H3K27M mutation. Using preliminary immunofluorescence and single cell RNA sequencing characterization, we observe a possible stall in neural stem cell differentiation, identifying a potential cell of origin for these lethal tumors.

Introduction

Diffuse intrinsic pontine gliomas (DIPG) are highly aggressive tumors, with peak incidence occurring in patients ages 6-8[1]. This incurable tumor resides in the pons and brainstem, making surgical resection impossible. DIPGs are also profoundly chemoresistant, with a median overall survival time of less than one year after diagnosis. The overall survival rate for this tumor type is approximately 30% at 1 year, 10% at 2 years, and less than 1% at 5 years[2–4].

Eighty percent of DIPG tumors harbor a lysine 27 to methionine (K27M) amino acid mutation[5]. The majority of these mutations occur in the *H3F3A* gene encoding histone H3.3, while a small portion occurs in the *HIST1H3B* gene encoding histone H3.1[6,7]. The H3K27M mutation inhibits Polycomb Repressive Complex 2 (PRC2) methyltransferase activity, leading to a global reduction of the transcriptional repressive trimethyl mark on histone H3 K27 (H3K27me3). The loss of H3K27me3 results in the uncontrolled transcription of thousands of genes whose expression is normally tightly controlled during early brain development[8,9]. Tumors characterized by the H3K27M mutation are correlated with a significantly worse prognosis and a reduced chance of survival. Accordingly, the World Health Organization in 2016 designated a molecular classification “diffuse midline glioma with H3K27M mutation” for this tumor type, one of the first such classifications for pediatric cancers[10].

The constricted developmental time of growth and the distinct midline location of H3K27M-driven DIPG formation indicates that the tumorigenic mutation event may occur early in development, specifically prenatally. Induction of the H3K27M mutation with *p53* loss was insufficient to generate tumors in the postnatal mouse brain, supporting the idea that the initial histone mutation takes place during embryogenesis[8]. Another study successfully generated an *in vivo* model of H3K27M-driven DIPG by expressing the mutation in early neural progenitor cells (NPCs), again promoting the idea that the initial stage of tumorigenesis occurs prenatally[11].

Attempts using human pluripotent stem cells (hPSCs) as a model for studying DIPG have shown promising results as the generation of embryonic stem cell derived

NPCs with the H3K27M mutation, p53 shRNA, and mutant PDGFRA with the D82V mutation gave rise to neuronal tumor cell formation. Conversely, when these genetically modified NPCs were transplanted into immunocompromised mice the results showed that these cells solely generated a lower grade glioma compared to the high-grade glioma that are present in patients. Furthermore, when mature astrocytes were transduced with the same combination as NPCs there was an absence of tumor formation concluding that tumorigenesis for H3K27M-driven DIPGs occurs within a specific cell type[12].

However, the specific cell of origin and developmental timing for the H3K27M tumorigenic mutation remains a subject of study. Studies have suggested that neuroepithelial cells (neural stem cells), radial glia (neural progenitor cells), and oligodendrocytes precursor cells are possible candidates for the initial origin of this mutation[13,14]. Previous work from our lab demonstrates that H3K27M cells are transcriptionally similar to early neural cells which have not yet undergone an epithelial-to-mesenchymal transition (EMT). Because the EMT occurs several times in embryonic brain development and is controlled by H3K27me3 deposition, our study suggests that pre-EMT cells are particularly susceptible to the H3K27M mutation.

Past models of H3K27M-driven tumorigenesis have been exclusively in cell culture or murine models. While these experiments have contributed invaluable insights, they are limited by the need to induce K27M expression in all cells in a particular population, and the inability to compare H3K27M induction at multiple time points. In addition, most studies have been in mice, which have notable differences in developmental timing and cell populations as compared to the embryonic human brain. These limitations have made it difficult to draw conclusions about the origins of

H3K27M gliomas, but fully understanding the cell of origin and developmental timing of gliomagenesis will be fundamental to successfully treating these tumors.

Therefore, we present here the first human embryonic stem cell based cerebral organoid model of H3K27M occurrence in embryonic brain development. Human brain organoids have been shown in recent years to recapitulate with unprecedented accuracy the self-organizing cellular layering and neural signaling that occurs during early corticogenesis[15,16]. 3D organoid models provide researchers with a flexible and manipulatable platform for interrogating cellular trajectories and neurodevelopmental subtyping seen in human prenatal embryogenesis[17,18]. Thus, cerebral organoids allowed to develop over time recapitulate many of the processes and cell types that have been implicated in H3K27M-driven gliomagenesis. While recent studies have shown the successful generation of cancer organoids using glioma cells taken from patient tumors, these models are limited to studies of mature tumor growth, rather than allowing study of the initial oncogenic events[19,20]. Here, we generated a model which allows induction of H3K27M expression in isolated cells in an otherwise normal organoid, mimicking the initial H3K27M mutation event.

Results

A. hESCs with doxycycline-inducible H3K27M-GFP display reduced H3K27 trimethyl levels and robust GFP expression.

In order to recapitulate the H3K27M mutation event in early brain development, we generated an H9-derived hESC line in which a single genomic copy of *H3F3A*-mutant K27M is expressed from a TRE promoter in the presence of consistent doxycycline. This system allows for flexible induction of the K27M mutation in order to study its effects at any developmental time point, and also allows turning off the mutation to evaluate whether its effects can be reversed.

An eGFP sequence was included immediately downstream of the mutant histone sequence to allow detection of the mutant cells. A control line was generated in which wild-type *H3F3A*-GFP is expressed in the same manner. Two lentiviral transductions of H9 hESCs were necessary to generate these lines; the first transduction introduced a vector with the tTS/rtTA sequence which controls the TRE promoter (Figures 1A, S2) and the second transduction introduced the TRE-H3K27M-eGFP or TRE-H3F3A-eGFP vector (Figures 1B, S1). Digital droplet PCR was used to verify single genomic integration of each construct (Figures S2, S3).

We demonstrate expression of K27M-mutant histone H3 only in H3K27M-GFP hESC cells, after growing for 120 hours in 10 ug/mL doxycycline (Figure 1C). We also observe decreased levels of the transcriptional repressive mark H3K27 trimethyl (me3) in the H3K27M-GFP hESCs as compared to the H3WT-GFP (Figure 1C). We quantified the comparative decrease in H3K27me3 using ImageJ, and observed a nearly 50% decrease in trimethylation in the H3K27M cells compared with the H3WT cells (Figure 1D). This demonstrates that our experimental system recapitulates the epigenetic phenotype of H3K27me3 loss found in H3K27M-mutant gliomas. Finally, fluorescence imaging shows robust GFP expression after growing for 96 hours in 10 ug/mL doxycycline, with no GFP expression without doxycycline (Figure 1E). This indicates that there is no leaky expression of the histone mutation in non-doxycycline conditions.

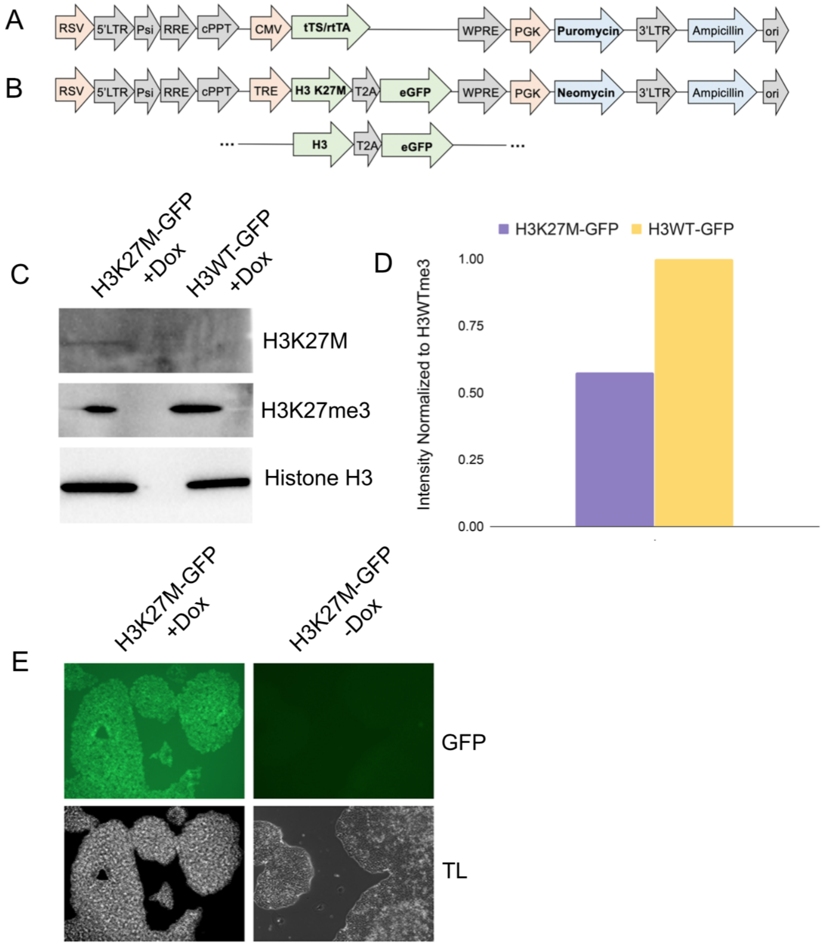


Figure 1. Generation of doxycycline-inducible H3K27M-GFP and H3WT-GFP expressing hESC lines. A) Map for the tTS/rtTA lentiviral vector. B) Map for the TRE promoter controlled H3K27M-GFP and H3WT-GFP lentiviral vectors. C) Western blots on protein lysate from cells incubated with Dox for 120 hours showing H3K27M expression and decreased H3K27me3 levels only in the H3K27M-GFP line. D) ImageJ quantification of H3K27me3 levels in both cell lines, normalized to the H3WT-GFP levels. E) GFP fluorescence in H3K27M-GFP cells after incubation with Dox for 48 hours. (TL, transmitted light)

B. Cerebral organoids seeded with H3K27M-expressing hESCs show preliminary alterations in cell type.

To study the effects of H3K27M expression on the developing embryonic brain, we generated normal hESC-derived cerebral organoids seeded with either 0.5% or 10% of H3K27M-GFP or control hESCs at aggregation (Figure 2A). The 0.5% organoids were used for immunofluorescence cell type marker staining, so that the smaller population of GFP+ cells would make it easier to visually discern cell interaction patterns. The 10% organoids were used for single cell RNA sequencing, so that the cell population of interest would be large enough to draw statistical conclusions.

These organoid cultures recapitulate the early weeks of human prenatal cerebral development, and generate several relevant neural cell types whose normal development is dependent on H3K27me3-driven transcriptional control. Neural epithelial cells are induced during week 1, and by week 2 have differentiated into radial glia cells and Cajal-Retzius neurons. Intermediate progenitors and early deep-layer neurons are generated between weeks 4 and 5. Differentiation into radial glia and intermediate progenitor cells from more primitive neural stem cells involves the epithelial-to-mesenchymal transition (EMT), which is controlled by H3K27me3-induced transcriptional repression of epithelial genes[21,22]. Because the H3K27M mutation causes global reduction of H3K27me3, we hypothesized that H3K27M-expressing cells would stall in differentiation at one of these developmental timepoints.

The organoids were allowed to develop normally over 5 weeks according to previously published protocols[23] (see Methods), except that on day 8 10 ug/mL doxycycline was added to the growth media. Doxycycline was added at every media change thereafter. Morphologically, the organoids appeared to develop normally, and no gross morphological changes were observed between the H3K27M-seeded or H3WT-seeded organoids (Figure 2A).

At week 5, we harvested 0.5% H3K27M and H3WT organoids for cryopreservation and antibody staining for cell type marker proteins. Preliminary results shown in Figure 2B demonstrate feasibility and illustrate trends, but the cryosectioning and staining protocols are still being optimized, so these results are not final. The cell marker protein antibodies were co-incubated with an anti-eGFP antibody to ensure robust GFP visualization even after fixation. Anti-PAX6 staining of the H3K27M-seeded organoid (Figure 2B left) shows almost no mixing between the H3K27M cells (green) and the PAX6+ radial glia (purple), consistent with an H3K27M-induced stall prior to differentiation into radial glia. In contrast, we observed significant co-staining (white) and mixing between the H3WT and PAX6+ populations (Figure 2B right).

We also stained for TBR2, a marker of intermediate progenitor (IP) cells, whose differentiation should be impacted if H3K27M stalls EMT in early neural development (Figure 2C). We observed mixing and co-staining between the GFP+ and TBR2+ populations in both the H3K27M and H3WT seeded organoids. We also observed mixing between GFP+ and CTIP2+ (early deep layer neurons) in both types of organoids. Notably, the CTIP2 antibody was not co-incubated with the anti-GFP antibody, demonstrating that robust GFP fluorescence is visible even after fixation. Subsequent optimization of the staining protocol may eliminate the anti-GFP antibodies to decrease background signal.

The TBR2 and CTIP2 staining results may be explained because in this experiment, doxycycline was erroneously left out of the organoid growth media during weeks 4-5. This would not have impacted the generation of most radial glia cells, as much of their development occurs prior to week 4. However, intermediate progenitor cells and deep layer neurons are developing during week 4, so we hypothesize that removal of doxycycline may have induced the “rescue” of the H3K27M-induced epigenetic and transcriptional phenotypes, allowing normal development of the more mature cell types. We plan to repeat this experiment with and without the doxycycline in order to study this result. Additionally, because deep layer neurons are only beginning to arise in week 5, we will stain week 10 organoids to further study the effect of H3K27M on the differentiation patterns of these cells.

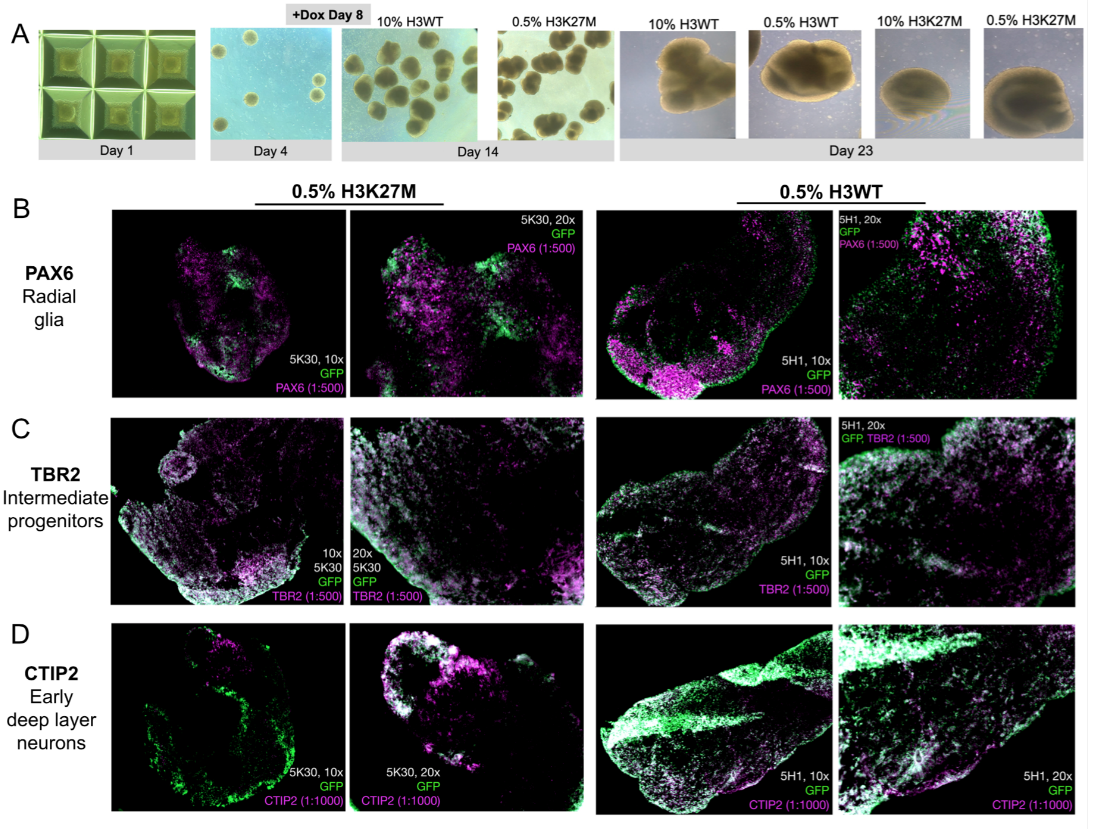


Figure 2. Developing human cerebral organoids with small populations of H3K27M cells. A) hESC-derived cerebral organoids were seeded with 0.5% or 10% of the experimental cell line (H3K27M-GFP or H3WT-GFP) and allowed to develop normally. B) PAX6 positive cells indicate radial glia cells. C) TBR2 positive cells indicate intermediate progenitors. D) CTIP2 positive cells indicate early deep layer neurons.

C. Preliminary single cell RNA sequencing of H3K27M-seeded cerebral organoids reveals H3K27M-driven transcriptional changes.

We performed single cell RNA sequencing (scRNA-seq) from week 2 +Dox cerebral organoids seeded with 10% H3K27M-GFP cells. Figure 3A shows confocal microscope imaging of whole week 2 and week 5 organoids (TL=transmitted light). We observed discrete populations of GFP+, H3K27M-expressing cells with limited mixing with the wild-type cells.

These results are preliminary since this was a low-depth sequencing run to check library quality. Higher-depth sequencing has been delayed due to the COVID-19 pandemic. However, stringent quality control measures were implemented to ensure only cells with robust expression were analyzed. We followed the MULTI-seq protocol[24] for multiplexing samples with lipid-tagged indices, before processing samples with 10X Genomics single cell kit. We took three replicates from each genotype, with approximately 4 organoids per replicate. Single cell suspensions of each replicate were labeled with unique lipid-tagged indices and then combined, and de-multiplexed after sequencing using the MULTI-seq de-multiplexing computational protocol. The data from the H3WT-seeded organoids were of too poor quality to analyze, meaning that we are not able to take into account any biological changes introduced by the transduction procedure. However, we are still able to compare the H3K27M-expressing cells with H3 wild-type cells in the same organoids.

The H3K27M scRNA-seq data were preprocessed and analyzed using the *scanpy* library in Python. Data from the 3 replicates were aggregated together for a total of 703 cells, which were reduced to 574 after filtering. H3K27M-transduced cells (n=27, 4%) were identified by implementing an *H3F3A* expression cutoff of 5, after observing a bimodal *H3F3A* expression distribution with the second mode starting at 5, and with the assumption that cells transduced with an extra copy of *H3F3A* would have much higher expression than the rest of the cells (Figure 3B "Inferred

Transduction”). This is a preliminary measure and future analysis will instead identify transduced cells by reads that map to the *eGFP* gene.

Leiden clustering found 8 clusters (Figure 3B “leiden”), and top ranked genes from each cluster were used to assign cell type based on cell type marker gene expression (Figure 3B “Cell Type”) using the scoreCT tool (github.com/LucasESBS/scoreCT). Substantial populations of neuroepithelial and radial glia cells were identified, along with a small population of intermediate progenitors which are most likely misassigned since week 2 is too early for this cell type. A few cells were unable to be assigned and are designated “NA”, likely a result of the low sequencing depth. Figure 3C shows expression of canonical early neural induction markers *NR2F1, NES* and *VIM*, indicating that our cerebral organoid model is undergoing neural induction as expected.

Finally, we investigated whether the inferred transduced H3K27M-expressing cells demonstrate altered gene expression associated with the H3K27M mutation. The H3K27M mutation is known to reduce the activity of the PRC2 methyltransferase enzyme so that it is unable to deposit transcriptional repressive trimethyl marks at the H3K27 residue[25]. Thus, in H3K27M mutant cells, many genes whose transcription is usually targeted for repression by PRC2 activity exhibit uncontrolled expression[25]. We observed that in our dataset, expression of the PRC2 catalytic subunit *EZH2* is comparatively low in the cluster containing most inferred H3K27M transduced cells (Figure 3D), consistent with reduced PRC2 activity. We next examined the expression of a list of PRC2 gene targets (BENPORATH PRC2 TARGETS from the Molecular Signatures Database), including only genes remaining in our dataset after filtering out lowly expressed genes. We observed overall higher expression of the PRC2 target genes in the inferred H3K27M-transduced genes, consistent with reduced PRC2 activity due to the H3K27M mutation (Figure 3E). As a comparison, we visualized the expression of a set of housekeeping genes[26–28] and observed similar overall expression between the inferred H3K27M-transduced cells and nontransduced cells (Figure 3F).

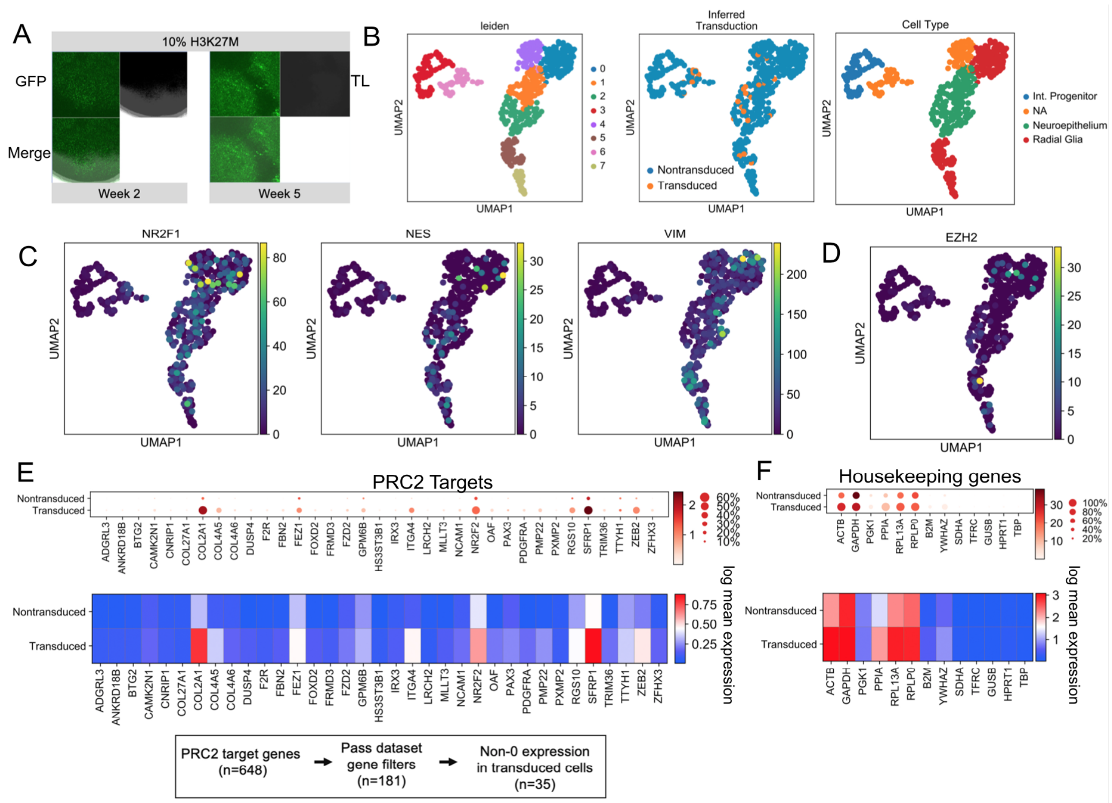


Figure 3. Single cell RNA sequencing of H3K27M-seeded cerebral organoids. A) Confocal organoid images of whole 10% H3K27M-GFP organoids at week 2 and week 5. B) *Scanpy* UMAP projection analysis of 574 single cells. 4% are inferred transduced with H3K27M-GFP. Cell type was assigned based on expression of cell type marker genes in the highest ranked

genes in each cluster (scoreCT method). C) Expression of neural induction marker genes in the single cell RNA sequencing dataset. D) Expression of EZH2. E) Expression of PRC2 targets by dot plot and log mean expression in the inferred transduced and nontransduced cell populations. In the dot plot, mean expression is shown by color, and fraction of cells expressing that gene in that category is shown by size of the dot. F) Expression of housekeeping genes by dot plot and violin plot in the inferred transduced and nontransduced cell populations.

Discussion

We demonstrate the successful generation of a cerebral organoid model of human embryonic brain development, with inducible H3K27M expression. We anticipate that this novel system will be very useful in characterizing the morphological, transcriptional and epigenetic effects of the histone mutation on early brain development. This system can be used to identify times in development which are most susceptible to the H3K27M mutation event, and can aid in identifying one or more potential cells of origin for deadly H3K27M-mutant gliomas.

Preliminary data from this experimental system indicates that H3K27M-mutant hESCs develop differently than H3 wild-type cells in the same organoid. Immunofluorescence data suggests that H3K27M-expressing neuroepithelial cells may fail to differentiate successfully into radial glia cells, consistent with a stall in the epithelial to mesenchymal transition caused by global loss of H3K27 trimethyl transcriptional control. Single cell RNA sequencing shows that inferred H3K27M-expressing cells tend to have higher expression of PRC2 target genes which would normally be repressed by the H3K27 trimethyl mark.

Overall, although these experiments will need to be repeated, we have established the necessary cell lines and optimized robust protocols. Our novel experimental organoid system holds tremendous promise for answering longstanding questions in the field about the developmental origins of H3K27M gliomas.

Materials and Methods

Cell Culture

Human embryonic stem cells (hESCs) were passaged every four to seven days at a splitting ratio of 1:3 to 1:6 to ensure that the cultures were 80 to 90% confluent on the day of the passage. The cells were passaged by washing thoroughly with DPBS (ThermoFisher Scientific) followed by treatment of 0.5 mM EDTA (Invitrogen) for 4-6 minutes then gently washed to ensure that all the cells were completely lifted off the plate but not broken down to single cells. The cell suspension was transferred to new culture vessels pre-coated with vitronectin (ThermoFisher Scientific) and maintained in StemFlex medium (ThermoFisher Scientific). After 24 hours of passaging, the media was changed followed by bi-daily media changes.

Lentiviral Transduction and Generation of Stable Cell Lines Expressing Doxycycline-inducible H3K27M-GFP or H3WT-GFP

H9 hESCs were plated at a density of $7x10^4$ cells per well of 24-well plate (Millipore Sigma). After 24 hours, the cells were transduced with a lentiviral vector carrying an expression construct with the tTS/rtTA open reading frame (ORF) and a puromycin resistance sequence (Figure S1) at 20, 30 and 40 infectious units (IU). These IU were used for transduction optimization to give the lowest number of integrated viral vector copies. The hESCs were incubated for 6 hours at 37℃, but after the first 6 hours the amount of StemFlex medium (ThermoFisher Scientific) was doubled per well. Medium was changed the next day. 48 hours post transduction, the cells were taken to single cell suspension using Accutase and seeded onto mouse embryonic fibroblasts (MEFs). Puromycin selection (10 ug/mL) was started 24 hours after seeding on MEFs and maintained at every media change. Colonies arising from a single cell were allowed to grow among MEFs for 4-7 days and were then transferred to vitronectin-coated culture vessels and grown and passaged normally. Digital droplet PCR was performed to quantify the number of genomic copies of the tTS/rtTA sequence in each cell clone (see next section). Two cell lines (named c5 and c8) were identified with single integration of the tTS/rtTa sequence. c5 was derived from the 30 IU transduction and c8 was derived from the 20 IU transduction. Gene and protein expression of the tTS/rtTA construct were validated by RT-PCR and Western blot (Figure S2).

c5 and c8 were each transduced with both the H3K27M-GFP virus and H3F3Awt-GFP (H3WT-GFP) virus separately, as described above except that the H3WT-GFP virus required incubation with 8ug/mL of polybrene (Millipore Sigma). Both plasmids contain a neomycin resistance sequence. Neomycin (ThermoFisher Scientific) selection (50 ug/mL) was started 24 hours after seeding on MEFs and maintained. Two independent cell lines will be derived from each tTS/rtTA line, for both H3K27M-GFP or H3WT-GFP. Currently, 2 H3WT-GFP and 1 H3K27M-GFP line have been derived from RTTA c5 transductions (Figure S3). Nineteen clones have been isolated from the c8 transductions and are frozen awaiting ddPCR quantification.

Reverse Transcriptase (RT) PCR

RNA was isolated and purified from 5x10^6 cells per cell line (Zymo Direct-Zol kit). cDNA was generated from 1 ug of total RNA using random hexamers and the SuperScript III First Strand Synthesis System (Invitrogen). PCR was performed using the following primers to the tTS/rtTA ORF: CCAGTTTGAACAAGCAGAGG (forward sense), CAGAGGTTCTCGCCTGAATA (reverse sense). cDNA from non-transduced hESC cells was used as a negative control; cDNA from the same cells with purified tTS/rtTA vector DNA spiked in was used as a positive control.

Digital Droplet Polymerase Chain Reaction (ddPCR)

For ddPCR, genomic DNA was isolated and purified from 5x10^6 cells per cell line (Zymo Quick-DNA Kit). Before the initial ddPCR assay, purified DNA samples were digested with ECORI in a reaction containing 5 µL 10x CutSmart Buffer (BioLabs cat# B7204S), 5 µL HF ECORI (BioLabs cat# R310RL) and 44 µL of gDNA at 37°C for 1 hour and 65°C for 20 mins in an Applied Biosystems Veriti 96 Well Thermal Cycler (Fisher Scientific). The ddPCR assays were carried out in a total volume of 30 µL containing 50 ng of DNA template, 40 µL 3 ×ddPCR Master Mix (Bio-Rad Laboratories), 3.6 µL of forward, 3.6 µL of reverse primer, 11.8 µL of water and 1 µL of probes. The primer and probe sequences are shown in Table 2. A Bio-Rad QX200 ddPCR droplet generator (Bio-Rad Laboratories) was used to divide the 20 µL mixture into approximately 20,000 droplets in a disposable DG8 Cartridge for the

QX100/QX200 Droplet Generator (Bio-Rad Laboratories). This was performed by 70 µL of Probes droplet generation oil (Bio-Rad Laboratories) being added to the bottom wells of the same cartridge. The final volume of droplets in oil was approximately 40 µL. The thermocycler was set to: 10 min at 95℃, followed by 40 cycles of 30 seconds at 94℃ and 1 minute at 60℃, followed by enzyme inactivation at 98℃ for 10 min and holding at 4℃. Finally, the amplified products were analyzed using a QX200 droplet reader (Bio-Rad Laboratories). For quantification, probes and primers for the *Ago* gene (2 genomic copies) were included.

Primer/Probe	**Sequences**
Ago - F	5’ TCTTGAGATGCCGGAACATAG 3’
Ago - R	5’ ACCAGCTGCGGAAGATTT 3’
Ago - P	5’-/56-FAM/CCAGGGTCA/ZEN/ACCTTGTTTCTGCAAATA/31ABkFQ/- 3’
tTS/rtTA - F	5’ TACCGTGAGGTGATGCTGGAG 3’
tTS/rtTA - R	5’ CACCTTTGGTTTGGTAAACAGG 3’
tTS/rtTA - P	5’ -/5HEX/TTACAGCAA/ZEN/CCTGGCCTCCATGGCA/3IABkFQ/- 3’
GFP - F	5’ ACTGGGTGCTCAGGTAGTGGTTGT 3’
GFP - R	5’ CAGCTCGCCGACCACTACCA 3’

GFP - P	5’- /5HEX/AACACCCCC/ZEN/ATCGGCGACGGCCCCGT/3ABkFQ/- 3’

Organoid Aggregation

hESC cultures were incubated with Accutase for 10 minutes at 37C and then washed into single cell suspension and counted. The cells were then aggregated into spheroids using AggreWell-800 plates (3x10^6 cells per AggreWell well), after coating with Anti-Adherence solution for 2 hours (STEMCELL Technologies). During the aggregation process, the H3K27M-GFP or H3WT-GFP hESCs were mixed with normal H9 hESCs at final percentages of either 0.5% (for immunofluorescence) or 10% (for single cell RNA-seq) based on initial cell counts. The organoids were transferred to a low-adherence 6 well plate (Corning) 48 hours after aggregation. For the first 18 days including aggregation, organoids were incubated in AggreWell media with 10 uM SB431542 and 3 uM IWR1. At day 18, the organoid plates were moved to an orbital shaker (100 rpm). For days 18-35, organoids were incubated in Sasai II media (see Supplemental Methods) with 10 ng/mL bFGF and 10 ng/mL EGF for the first week. For days 35-70, organoids were incubated in Sasai III media (see Supplemental Methods), and after day 70 incubated in Sasai IV media. Doxycycline (10 ug/mL) was added on day 7.

Western Blotting

Cell pellets (1.5x10^7 cells) were lysed with 1X RIPA buffer (abcam) and protease inhibitor (Millipore Sigma). 26 μL of total protein lysate was loaded in 12% NUPAGE Bis-Tris protein gel (ThermoFisher Scientific) and electrophoresed. Proteins were transferred onto nitrocellulose membranes by using a transfer apparatus iBlot 2 (ThermoFisher Scientific) at 20V for 4 minutes. Membranes were then blocked with 5% non-fat milk in PBST for 1 hour. The membranes were washed three times for 5 minutes in PBST and incubated with antibodies against H3 (1:5000, Abcam cat# 12079), H3K27M (1:250, Abcam cat# 190631), or H3K27me3 (1:500, Cell Signaling cat# 9733) at 4°C overnight in 5% non-fat milk. Membranes were again washed with PBST three times for 5 minutes and incubated with the appropriate secondary antibody containing chemiluminescence for 1 hour in 5% non-fat milk. Blots were washed with PBST three times for 5 minutes each, incubated in HRP Chemiluminescent Substrate (ThermoFisher) and visualized on ChemiDoc Imaging System (BioRad). Secondary antibodies were used against the appropriate primary antibodies shown in Table 1. Each experiment was repeated independently three times. Quantification of the western blots were completed using the Image J software.

1° Antibodies	**Dilutions**	**Company & Catalog Number**
TetR Monoclonal Antibody (mouse)	1:1000	Takara Bio cat# 631131
Histone H3 (goat)	1:5000	Abcam ab12079
H3K27M (rabbit)	1:250	Abcam ab190631-10

H3K27me3 (rabbit)	1:500	Cell Signaling Technology 9733S
2° Antibodies	**Dilutions**	**Company & Catalog Number**
Goat Anti-Mouse IgG-HRP	1:10,000	SCBT sc-2005
Donkey Anti-Goat IgG H&L (HRP)	1:12,500	ThermoFisher cat#15999
Goat Anti-Rabbit IgG H&L (HRP)	1:250 (for H3K27M) 1:1000 (for H3K27me3)	Abcam cat#6721

Single cell RNA sequencing

We followed the MULTI-seq protocol[24] for multiplexing samples with lipid-tagged indices before processing samples with the 10X Genomics Chromium Next GEM Single Cell 3’ Kit v3.1. Biological replicates (4-8 organoids per replicate) were treated with trypsin to generate single cell suspensions and filtered through cell strainers. Cell counts were generated to ensure 500,000 cells or fewer per replicate and final pellets were resuspended in 180 uL PBS. All subsequent MULTI-seq steps were performed on ice. Cell suspensions were incubated for 5 minutes with a 1:1 molar mixture of an “anchor” lipid modified oligonucleotide (LMO) with a unique MULTI-seq sample barcode per replicate. A “co-anchor” LMO was diluted 1.1 uL in 20 uL PBS and 20 uL of dilution was added to each replicate for 5 additional minutes before quenching with 1 mL 1% BSA. Cells were pelleted and washed with 1% BSA twice. All samples were combined, filtered through a cell strainer, counted, and diluted to 575 cells/uL to achieve 15,000 cells per each of 2 10X lanes. GEM generation, GEM-RT, and GEM-RT cleanup were performed according to 10X Chromium v3.1. cDNA amplification was performed according to 10X protocol except that the cDNA amplification mix included 1 uL of 2.5 uM MULTI-seq additive primer to ensure cDNA amplification of MULTI-seq barcodes. The barcode and endogenous cDNA fractions were separated during a 0.6X SPRI bead size selection, and the endogenous cDNA fraction was prepared according to the 10X protocol. The barcode fraction was prepared using a small RNA enrichment protocol with 3.2X SPRI and 1.8X isopropanol, before washing beads with 80% ethanol and resuspension in elution buffer. Finally, library PCR used 3.5 ng barcode DNA with 2.5 uL 10uM each Universal I5 primer and RPI primer (unique for each 10X lane). Endogenous cDNA and barcode libraries were sent for NGS and the resulting cDNA data were de-multiplexed using the barcode data.

Immunofluorescent staining of cryopreserved organoids

Individual organoids were fixed in 200 uL 4% PFA at room temperature for 30 minutes and then rinsed 3x in PBS. Each organoid was then incubated in 200 uL 30% sucrose in 4C for 24 hours. The organoid was then preserved in tissue freezing medium (Thomas Scientific) in a cryomold (Fisher Scientific) at -80C. Preserved organoids were sliced to 20 um and placed on microscope slides. Slides were stored at -80C. For staining, slides were warmed to room temperature and washed once in

PBST for 5 minutes without rocking to remove tissue freezing medium. We performed secondary fixation of the slides in 4% PFA for 30 minutes, followed by a second set of 3x PBST washes with rocking. Samples were then incubated in a permeabilizing solution (0.15% TritonX in PBS) for 15 minutes, and then blocked in 15% BSA (15% BSA in PBS) for 2 hours. Primary antibodies were diluted to either 1:500 or 1:1000 in blocking solution. Samples were then incubated overnight at 4C in primary antibody solution. The next day, samples were washed 3x in PBST with rocking. Secondary antibodies were diluted in blocking solution to either 1:500 or 1:1000. Samples were then incubated in secondary antibody solution at room temperature for 2 hours without exposure to light. Samples then went through another set of 3x PBST washes. A drop or about 20 uL of ProLong Gold Antifade Mountant (Thermo Fisher) was applied to each sample and a glass coverslip was placed atop each sample. The samples were then allowed to dry for 24 hours in the dark.

Supplementary Figures

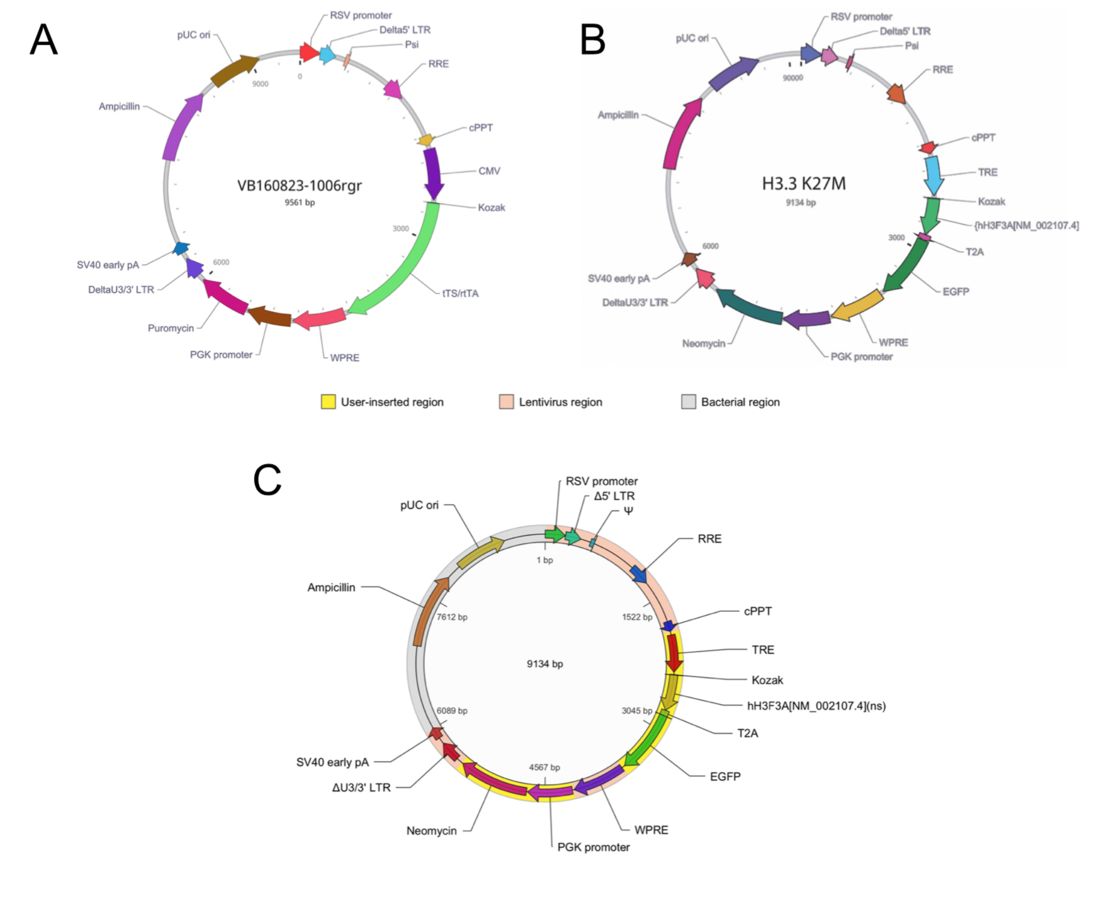


Figure S1. Vector maps used for lentiviral transductions. A) tTS/rtTA vector. B) H3.3 K27M vector. C) H3.3 wild type vector.

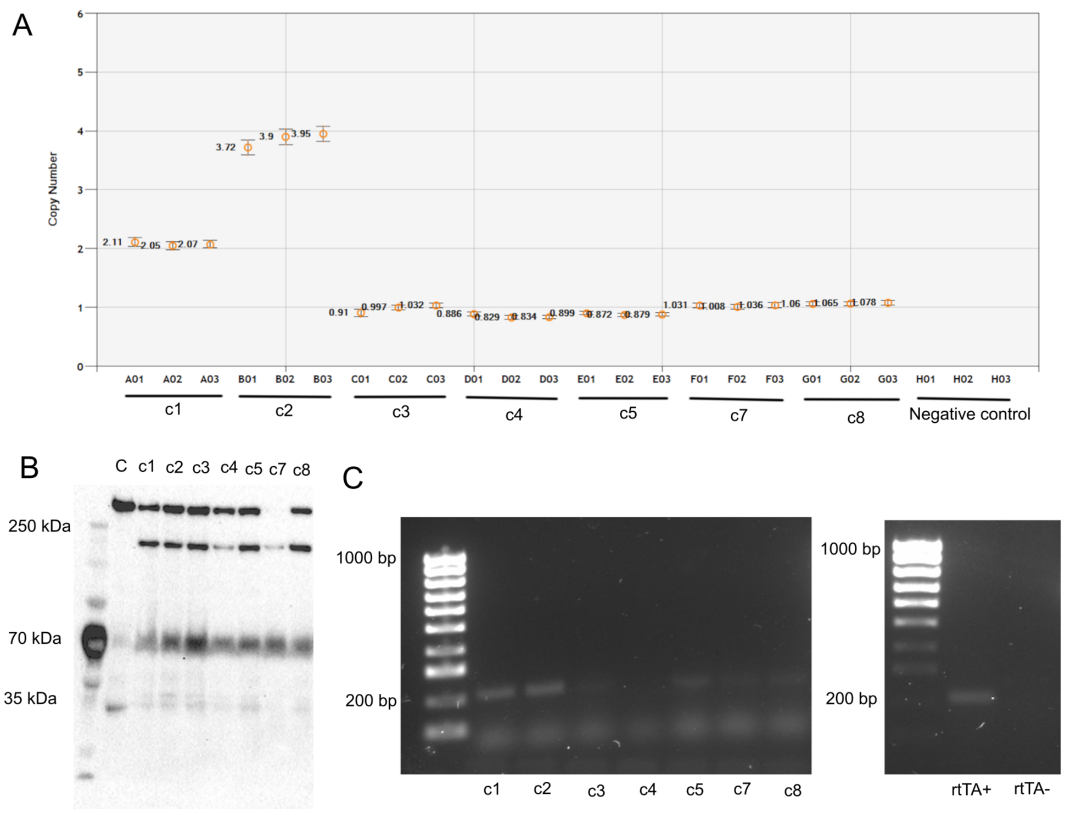


Figure S2. Digital droplet PCR, Western blot and RT-PCR validation of tTS/rtTA expressing hESC cell lines. A) ddPCR quantification of rtTA sequence genomic copies (normalized to Ago quantification). Clones 3-8 show single genomic integration. Negative control is gDNA from non-transduced hESC cells. B) Western blot (TetR mouse monoclonal antibody) showing expression of rtTA protein (35 kDa) in cell lysate. Clones 2, 3, 5 and 8 show robust expression. C = negative control lysate from non-transduced cells. C) RT-PCR showing RNA expression of the rtTA sequence (expected PCR product 200 bp). Of the single integrant clones with robust protein expression, only clones 5 and 8 show RNA expression. The rtTA+ positive control lane has cDNA from non-transduced hESCs with purified tTS/rtTA vector DNA spiked in, and the rtTA- negative control lane just has cDNA from the non-transduced hESCs.

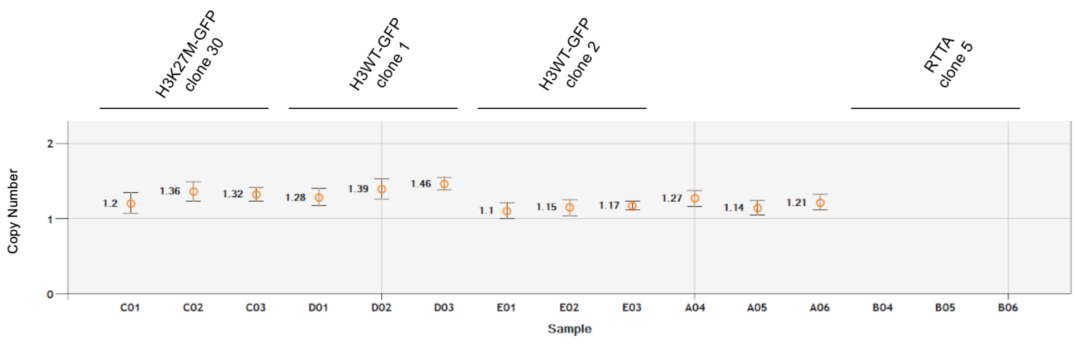


Figure S3. Digital droplet PCR validation of H3K27M-GFP and H3WT-GFP expressing hESC cell lines. ddPCR quantification of eGFP sequence genomic copies (normalized to Ago quantification). Shown here are the current successful cell lines with only one genomic copy of the eGFP sequence, resulting from lentiviral transduction. All 3 lines were derived from RTTA clone 5, which is used as the negative control here since it does not express eGFP.

References

1. Jones C, Baker SJ. Unique genetic and epigenetic mechanisms driving paediatric diffuse high-grade glioma. Nat Rev Cancer. 2014;14. doi:10.1038/nrc3811

2. Vanan MI, Eisenstat DD. DIPG in Children - What Can We Learn from the Past? Front Oncol. 2015;5: 237.

3. Dellaretti M, Reyns N, Touzet G, Dubois F, Gusmão S, Pereira JLB, et al. Diffuse brainstem glioma: prognostic factors. J Neurosurg. 2012;117: 810–814.

4. Mathew RK, Rutka JT. Diffuse Intrinsic Pontine Glioma : Clinical Features, Molecular Genetics, and Novel Targeted Therapeutics. J Korean Neurosurg Soc. 2018;61: 343–351.

5. Larson JD, Kasper LH, Paugh BS, Jin H, Wu G, Kwon C-H, et al. Histone H3.3 K27M Accelerates Spontaneous Brainstem Glioma and Drives Restricted Changes in Bivalent Gene Expression. Cancer Cell. 2019;35: 140–155.e7.

6. Khuong-Quang D-A, Buczkowicz P, Rakopoulos P, Liu X-Y, Fontebasso AM, Bouffet E, et al. K27M mutation in histone H3.3 defines clinically and biologically distinct subgroups of pediatric diffuse intrinsic pontine gliomas. Acta Neuropathol. 2012;124: 439–447.

7. Wu G, Broniscer A, McEachron TA, Lu C, Paugh BS, Becksfort J, et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. Nat Genet. 2012;44: 251–253.

8. Lewis PW, Müller MM, Koletsky MS, Cordero F, Lin S, Banaszynski LA, et al. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. Science. 2013;340: 857–861.

9. Bender S, Tang Y, Lindroth AM, Hovestadt V, Jones DTW, Kool M, et al. Reduced H3K27me3 and DNA hypomethylation are major drivers of gene expression in K27M mutant pediatric high-grade gliomas. Cancer Cell. 2013;24: 660–672.

10. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131: 803–820.

11. Pathania M, De Jay N, Maestro N, Harutyunyan AS, Nitarska J, Pahlavan P, et al. H3.3K27M Cooperates with Trp53 Loss and PDGFRA Gain in Mouse Embryonic Neural Progenitor Cells to Induce Invasive High-Grade Gliomas. Cancer Cell. 2017;32: 684–700.e9.

12. Funato K, Major T, Lewis PW, Allis CD, Tabar V. Use of human embryonic stem cells to model pediatric gliomas with H3.3K27M histone mutation. Science. 2014;346: 1529–1533.

13. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science. 2018;360: 331–335.

14. Jessa S, Blanchet-Cohen A, Krug B, Vladoiu M, Coutelier M, Faury D, et al. Stalled developmental programs at the root of pediatric brain tumors. Nat Genet. 2019;51: 1702–1713.

15. Kelava I, Lancaster MA. Dishing out mini-brains: Current progress and future prospects in brain organoid research. Dev Biol. 2016;420: 199–209.

16. Heide M, Huttner WB, Mora-Bermúdez F. Brain organoids as models to study human neocortex development and evolution. Curr Opin Cell Biol. 2018;55: 8–16.

17. Kadoshima T, Sakaguchi H, Nakano T, Soen M, Ando S, Eiraku M, et al. Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell-derived neocortex. Proc Natl Acad Sci U S A. 2013;110: 20284–20289.

18. Bhaduri A, Andrews MG, Mancia Leon W, Jung D, Shin D, Allen D, et al. Cell stress in cortical organoids impairs molecular subtype specification. Nature. 2020;578: 142–148.

19. Linkous A, Balamatsias D, Snuderl M, Edwards L, Miyaguchi K, Milner T, et al. Modeling Patient-Derived Glioblastoma with Cerebral Organoids. Cell Rep. 2019;26: 3203–3211.e5.

20. Ogawa J, Pao GM, Shokhirev MN, Verma IM. Glioblastoma Model Using Human Cerebral Organoids. Cell Rep. 2018;23: 1220–1229.

21. Itoh Y, Moriyama Y, Hasegawa T, Endo TA, Toyoda T, Gotoh Y. Scratch regulates neuronal migration onset via an epithelial-mesenchymal transition-like mechanism. Nat Neurosci. 2013;16: 416–425.

22. Ohayon D, Garcès A, Joly W, Soukkarieh C, Takagi T, Sabourin J-C, et al. Onset of Spinal Cord Astrocyte Precursor Emigration from the Ventricular Zone Involves the Zeb1 Transcription Factor. Cell Rep. 2016;17: 1473–1481.

23. Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, et al. Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and

Influence Cell-Type-Specific Genes. Stem Cell Reports. 2019;12: 245–257.

24. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat Methods. 2019;16: 619–626.

25. Stafford JM, Lee C-H, Voigt P, Descostes N, Saldaña-Meyer R, Yu J-R, et al. Multiple modes of PRC2 inhibition elicit global chromatin alterations in H3K27M pediatric glioma. Sci Adv. 2018;4: eaau5935.

26. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29: 569–574.

27. Curina A, Termanini A, Barozzi I, Prosperini E, Simonatto M, Polletti S, et al. High constitutive activity of a broad panel of housekeeping and tissue-specific cis-regulatory elements depends on a subset of ETS proteins. Genes Dev. 2017;31: 399–412.

28. Moein S, Javanmard SH, Abedi M, Izadpanahi MH, Gheisari Y. Identification of Appropriate Housekeeping Genes for Gene Expression Analysis in Long-term Hypoxia-treated Kidney Cells. Adv Biomed Res. 2017;6: 15.

Chapter 6: Future Directions and Other Work

6.1 Future directions: H3K27M-inducible cerebral organoid experiments

I have designed a plan for comprehensive characterization of the novel H3K27M-inducible cerebral organoid system. For these experiments, H3K27M-GFP or H3WT-GFP expression will be induced on day 8 as before, in organoids seeded with 0.5% or 10% of the appropriate transduced cell line. At weeks 2, 5 and 10, single cell RNA sequencing data (4 replicates each genotype) will be taken from 10% seeded organoids and processed using the MULTI-Seq protocol for sample multiplexing[37], followed by 10X Genomics Single Cell RNA-Seq Kit v3.1. At the same weekly timepoints, 3 organoids from each genotype will be cryopreserved and sectioned for staining. At all timepoints, each genotype will be stained for SOX2 (neural epithelium), PAX6 (radial glia), and nuclear DAPI. At weeks 5 and 10, each genotype will also be stained for TBR2 (intermediate progenitors) and CTIP2 (deep layer neurons). We will also take CHIP-seq data on the H3K27me3 mark to show its chromatin binding patterns in the H3K27M-mutant cells as compared to the wild-type cells.

If no interesting differentiation patterns are observed with H3K27M induction at day 8, we will try induction during week 3 when stem cells are differentiating into progenitor cells, and week 6 when deep layer neurons are developing.

For long-term future directions, this experimental platform lends itself to therapeutically-directed testing, and as the organoid technology evolves, may be improved by the addition of relevant immune or progenitor cells. For drug testing, the UCSC Chemical Screening Center (CSC) has several large compound libraries including both FDA-approved drugs and novel drug-like molecular compounds. A clinically-relevant future direction would be to subject the H3K27M-expressing organoids to high-throughput drug screening at the CSC to identify compounds with activity against pre-oncogenic H3K27M-expressing cells. Pro-differentiation therapies have shown promise in developmental pediatric cancers[38–40], so this approach has the potential to identify compounds which reverse the H3K27M-associated differentiation stall.

Additionally, in order to understand the role of the brain's immune system in the early development of H3K27M-associated gliomagenesis, involving microglia cells would be a very relevant addition to this model. Preliminary work in the Haussler lab has already shown the capacity for including microglia in cerebral organoids.

Lastly, because several groups have identified an oligodendrocyte-precursor-like (OPC) dominant malignant cell type in mature H3K27M gliomas, it would be worthwhile to investigate the consequences of inducing H3K27M at the OPC stage as compared to other developmental stages. Currently, our system shows a small cell population with early OPC-like signaling, but additional optimization and characterization is needed to promote differentiation along the oligodendrocytic lineage and show the presence of OPCs and mature oligodendrocytes. A recent study demonstrated the successful generation of organoids containing mature oligodendrocytes, astrocytes and neurons, and importantly shows that oligodendrocyte developmental stages are present and observable[41]. Future work incorporating OPCs into our organoids could follow a similar protocol.

6.2 Other work: Cholesterol biosynthesis as a novel therapeutic vulnerability in DIPG

In an effort to identify unique therapeutic vulnerabilities in DIPG, I identified genes with outlier expression in in 3 DIPG tumor-derived cell lines[21] as compared to the Treehouse cancer compendium (pan-cancer) or other gliomas (pan-disease) (github.com/UCSC-Treehouse/CARE, compendium v5). *HMGCR*, the rate-limiting enzyme in the cholesterol biosynthesis pathway[42], was the only targetable gene with both pan-cancer and pan-disease outlier expression in all 3 DIPG cell lines. Further investigation revealed that *HMGCR*, *HMGCS1*, and *IDI1*, key enzymes in the cholesterol biosynthesis pathway, all have significantly higher expression in DIPG patient samples (Treehouse compendium) as compared to TCGA adult glioblastoma (aGBM) or GTEx normal healthy cerebellum[43] (Figure 6.2.1), indicating a potential dependence of DIPG cells on cholesterol biosynthesis.

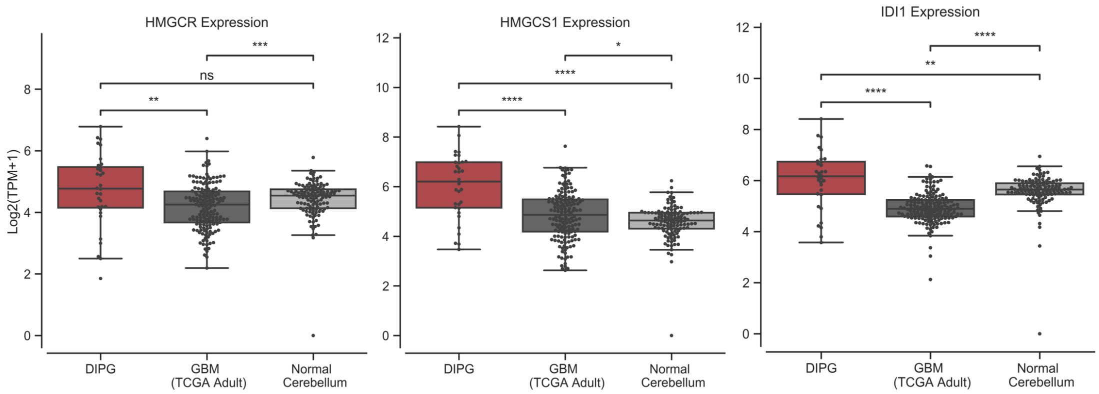


Figure 6.2.1. Expression of key cholesterol biosynthesis enzymes in samples from DIPG tumors, TCGA adult GBM tumors, and GTEx normal cerebellum from healthy deceased donors.

Notably, most of the brain's endogenous cholesterol is used for the production of myelin[44,45], which is interesting in the context of recent work showing that DIPG cells integrate into neural circuits and promote neuronal excitability for proliferation[46,47]. Such dependency on neuronal activity suggests that DIPG tumors may hijack normal myelination pathways for glioma cell proliferation. I found that myelin subunits *MBP* and *CNP* have significantly higher expression in DIPG compared to aGBM or normal cerebellum (Figure 6.2.2 A,B). Additionally, the myelination pathway displays higher overall expression in DIPG (Figure 6.2.2 C,D). These results suggest that if DIPG tumors hijack myelination for proliferation, this could introduce a metabolic dependence on production of endogenous cholesterol in the brain. Because decreased myelin is involved in multiple sclerosis and other myelin disorders[48], therapeutically targeting myelination in a young child can have long-term negative effects on cognitive development and function, but I hypothesized that targeting cholesterol biosynthesis using a statin may be effective and relatively harmless.

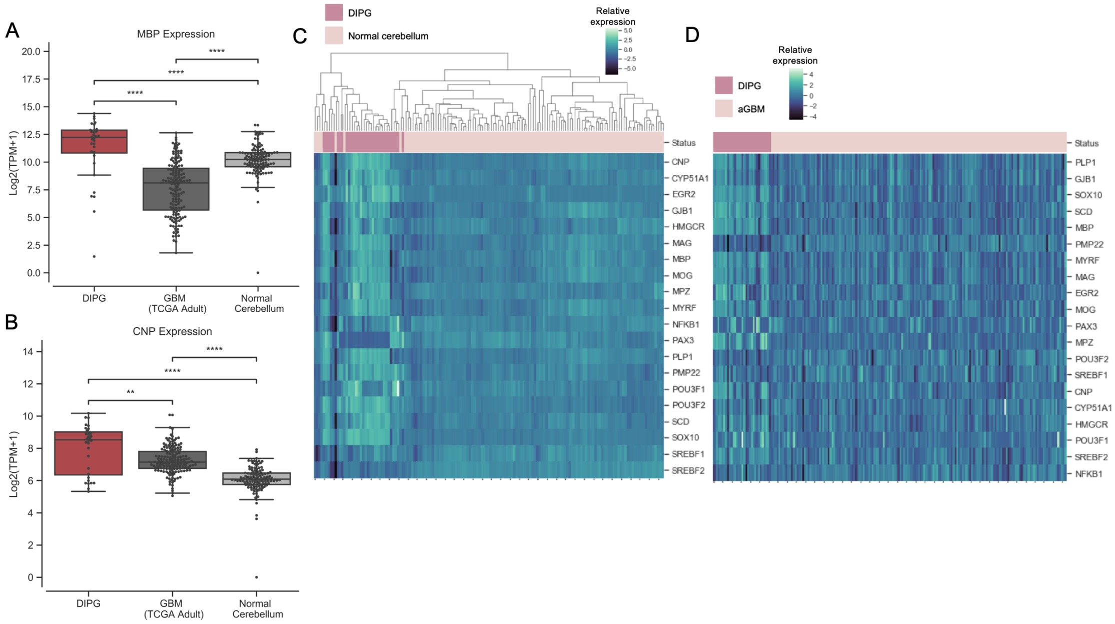


Figure 6.2.2 Expression of myelin genes and pathway in DIPG, aGBM and normal cerebellum. A) Expression of MBP (myelin subunit). B) Expression of CNP (myelin subunit). C) Heatmap of myelination pathway genes in DIPG and normal cerebellum. D) Heatmap of myelination pathway genes in DIPG and adult GBM.

To test this hypothesis, the Agnihotri lab (University of Pittsburgh) treated DIPG tumor-derived cell lines SU-DIPG-IV and SU-DIPG-VI with simvastatin, a well-characterized inhibitor of *HMGCR* and cholesterol biosynthesis[49]. As compared to vehicle control, simvastatin decreased DIPG cell viability nearly 50% in both cell lines, with almost no effect on normal human astrocytes or neural stem cells (Figure 6.2.3). This indicates that simvastatin is preferentially toxic to DIPG cells, and demonstrates that outlier gene expression can predict *in vitro* response.

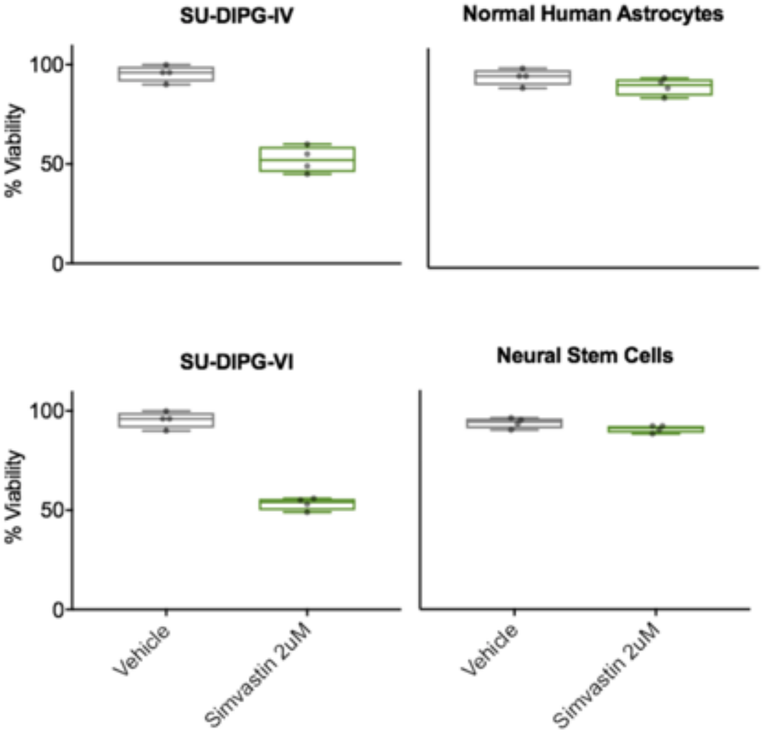


Figure 6.2.3. Simvastatin treatment of DIPG cell lines SU-DIPG-IV and SU-DIPG-VI, normal human astrocytes and neural stem cells. (Agnihotri lab)

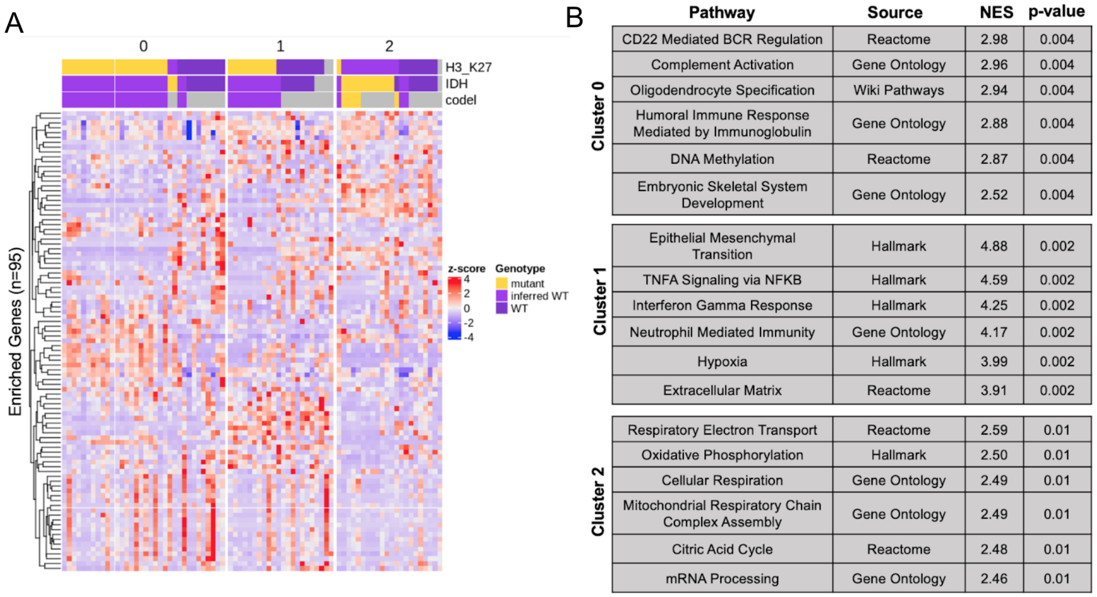
This project will be continued in the Vaske lab. Future directions include repeating simvastatin treatment on a larger (n=14) cohort of DIPG cell lines, evaluating whether simvastatin affects myelination activity using a myelination assay, and knocking out *HMGCR* with CRISPRi or shRNA to validate that DIPG cells are dependent on *HMGCR* activity.

6.3 Other work: *Hydra* Bayesian hierarchical clustering analysis identifies novel subtypes of pediatric high grade glioma

Identifying clinically relevant subtypes in pediatric cancer is an ongoing challenge, because while individual patients within a disease type may respond differently to therapy, often the small numbers of tumor cases at each institution make it difficult to assemble a cohort large enough to represent clinically meaningful diversity. Because the Treehouse cancer compendium contains pediatric high grade glioma (pHGG) RNA-seq samples from many institutions, I used the recently published *hydra* clustering method to search for clinically relevant pHGG subtypes[50].

Identifying relevant subtypes in high-dimensional gene expression data is difficult overall because the number of genes greatly exceeds the number of samples, making traditional unsupervised clustering methods underpowered. Therefore, *hydra* identifies clusters based on multimodal gene expression, based on previous work that shows that cancer subtypes fall into multimodal expression patterns[51]. *Hydra* has been shown to identify clinically relevant subtypes in other pediatric cancers such as neuroblastoma and small blue round cell tumor cohorts[50]. Therefore, I applied the *hydra enrich* unsupervised clustering method to the Treehouse pHGG cohort (n samples=78, compendium v9), and 3 novel clusters were identified (Figure 6.3.1).

Interestingly, clusters 0 and 1 were composed of a mixture of H3K27M-mutant and H3WT gliomas, whereas past pan-glioma analyses have typically found that H3K27M gliomas clustered together[3]. This indicates that *hydra* is capable of identifying subtle gene expression subtypes beyond the genes correlating with mutation status. Cluster 0 was enriched for BCR and complement immune activation, as well as oligodendrocytic and skeletal developmental signals which may be expected for H3K27M-mutant gliomas. Cluster 1 was enriched for interferon and neutrophil immune signaling, as well as other cancer pathways including NFKB signaling. Cluster 2, composed of a mixture of IDH-mutant and wild-type gliomas, was not enriched for immune signaling but was characterized by metabolism and RNA processing pathways.



	Pathway	Source	NES	p-value
Cluster 0	CD22 Mediated BCR Regulation	Reactome	2.98	0.004
	Complement Activation	Gene Ontology	2.96	0.004
	Oligodendrocyte Specification	Wiki Pathways	2.94	0.004
	Humoral Immune Response Mediated by Immunoglobulin	Gene Ontology	2.88	0.004
	DNA Methylation	Reactome	2.87	0.004
	Embryonic Skeletal System Development	Gene Ontology	2.52	0.004
Cluster 1	Epithelial Mesenchymal Transition	Hallmark	4.88	0.002
	TNFA Signaling via NFKB	Hallmark	4.59	0.002
	Interferon Gamma Response	Hallmark	4.25	0.002
	Neutrophil Mediated Immunity	Gene Ontology	4.17	0.002
	Hypoxia	Hallmark	3.99	0.002
	Extracellular Matrix	Reactome	3.91	0.002
Cluster 2	Respiratory Electron Transport	Reactome	2.59	0.01
	Oxidative Phosphorylation	Hallmark	2.50	0.01
	Cellular Respiration	Gene Ontology	2.49	0.01
	Mitochondrial Respiratory Chain Complex Assembly	Gene Ontology	2.49	0.01
	Citric Acid Cycle	Reactome	2.48	0.01
	mRNA Processing	Gene Ontology	2.46	0.01

Figure 6.3.1. *Hydra* clustering analysis reveals 3 novel clusters in a pHGG cohort. A) Gene expression heatmap of the 3 *hydra* clusters, showing the expression of 95 multimodally expressed genes which are enriched for coordinated expression of GO term genes, identified by the *hydra* enrich command. B) Representative GO terms significantly enriched in each cluster.

I next correlated *hydra* clusters with specific types of immune cells, because clusters 0 and 1 appear enriched in immune signaling, and cancer subtypes based on immune enrichment have prognostic significance[52]. First, I used ESTIMATE to infer the levels of immune cell infiltrate in each sample[53], confirming that clusters 0 and 1 have significantly higher immune infiltration than cluster 2 (Figure 6.3.3 A). Then, I used CIBERSORT and the leukocyte LM22 immune gene expression signatures to identify specific immune cell types with high expression in each cluster[54].

Monocytes have higher expression in clusters 0 and 1, while activated natural killer cells and eosinophils have higher expression in cluster 0 (Figure 6.3.3 B). A next step for this analysis would be to identify levels of microglia infiltration, since the CIBERSORT LM22 immune signatures dataset does not include microglia, and the microglia-glioma microenvironment has been the subject of much study on immunosuppression and potential immunotherapy development[55,56].

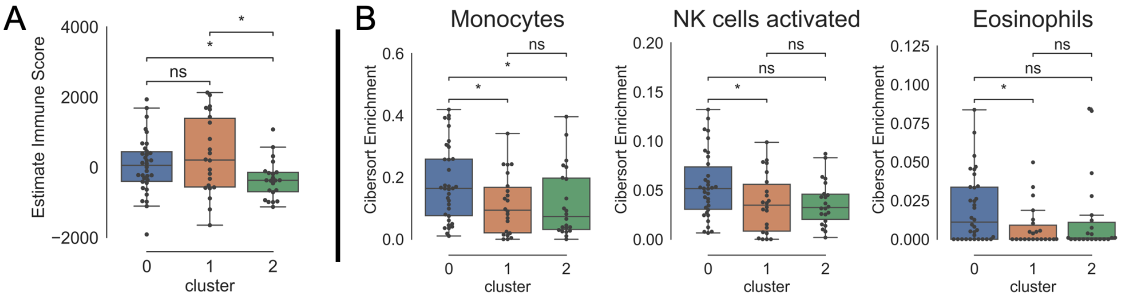


Figure 6.3.2. *Hydra* pHGG clusters are characterized by immune infiltration. A) Immune infiltration score from ESTIMATE for samples within *hydra* clusters. B) CIBERSORT enrichment for three types of immune cells with significant differences between *hydra* clusters.

Finally, I hypothesized that although the H3K27M mutation introduces a strong gene expression signal due to epigenetic dysregulation, K27M-mutant pHGG patient cohorts may in fact harbor diverse gene expression subtypes. I attempted *hydra enrich* on the H3K27M cohort (n=33) but only one cluster was identified, highlighting the known transcriptional similarity of H3K27M gliomas. Then, I applied the *hydra sweep* method to identify gene sets with differential expression within the H3K27M-only pHGG cohort. *Hydra sweep* results can be used to identify clusters with differential expression of each specific gene set. The *sweep* command identified 169 differentially expressed Gene Ontology (GO) terms in the H3K27M-only cohort. The majority of the GO terms had non-clinical relevance; for example, several gene sets were differentially expressed due to the presence of genes on the Y chromosome and so neatly separated the dataset into male and female patients.

However, 2 GO terms were particularly interesting: GO Stem Cell Differentiation and GO Mesenchymal Cell Differentiation. Both GO terms contained 2 transcription factors known for their roles in early brain cell type differentiation: *PAX3* and *SOX10*. Interestingly, the expression of these genes was nearly opposite each other: pHGG patient samples with high *PAX3* expression had lower expression of *SOX10*, and vice versa (Figure 6.3.3 A). A *PAX3*-high DIPG subtype has already been characterized as inhibiting apoptosis and enhancing *PDGFb*-induced brainstem gliomagenesis[57]. However, the *SOX10*-high subtype is novel, as is the observation that the expression of these transcription factors is mutually exclusive in a H3K27M-only cohort.

Notably, *PAX3* and *SOX10* have extremely time- and cell-type-restricted expression during early brain development (Figure 6.3.3 B). *PAX3* is exclusively expressed in neural stem cells, with no expression in mature astrocytes[58]. In contrast, *SOX10* is a marker of the oligodendrocytic lineage whose expression is required for formation of myelinating oligodendrocytes[59]. Most interestingly, both neural stem cells and oligodendrocyte precursors have been proposed as cells of origin for H3K27M glioma[14,60]. Therefore, I hypothesize that the expression of these transcription factors may represent different lineages deriving from each glioma's cell of origin, and that H3K27M-mutant gliomas may actually have at least 2 possible cells of origin. Further experimental study is needed to follow up on this hypothesis.

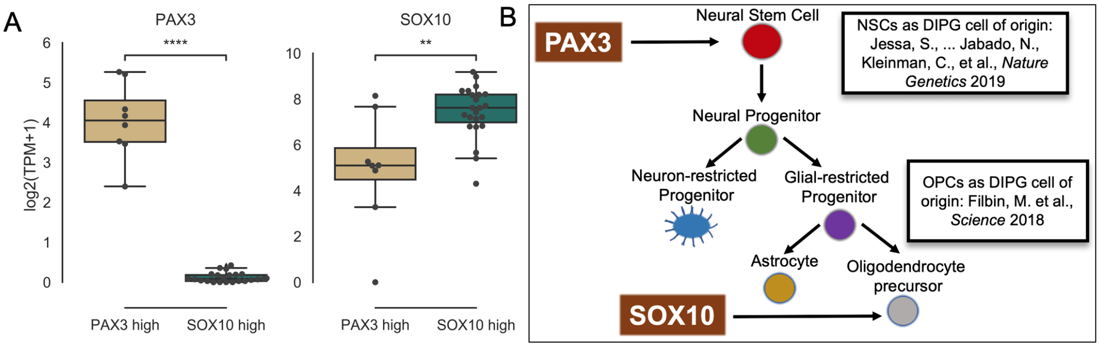


Figure 6.3.3. *Hydra* clustering analysis of H3K27M-only pHGG reveals novel developmental subtypes. A) Expression of *PAX3* and *SOX10* in clusters derived from *hydra sweep.* B) Illustration of normal expression timing of *PAX3* and *SOX10* during brain development.

As a preliminary *in vitro* verification, we performed RT-PCR for *PAX3* and *SOX10* in 10 H3K27M glioma patient-derived cell lines and 2 H3 wild-type (wt) cell lines (Figure 6.3.4). Several cell lines did not express either transcription factor, but SU-DIPG-4, 19 and 21 expressed *PAX3* but not *SOX10*, while SU-DIPG-6, 25, and 30 expressed *SOX10* but not *PAX3*. This partially recapitulates the computational analysis and provides a potential experimental framework for comparing *PAX3*+ vs *SOX10*+ glioma cells.

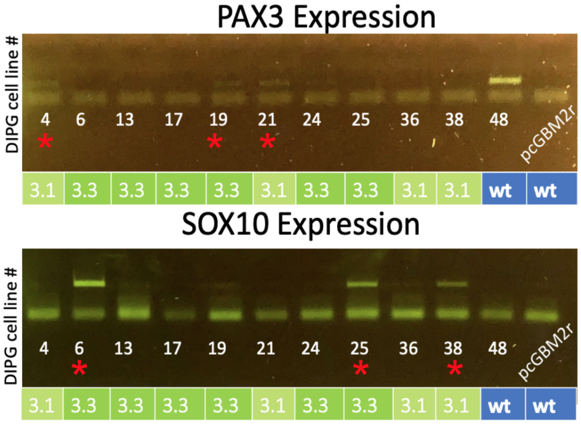


Figure 6.3.4. *Hydra* H3K27M pHGG subtypes are represented in DIPG patient-derived cell lines. All cell lines except pcGBM2r are "SU-DIPG" cell lines supplied by Dr. Michelle Monje, Stanford University. pcGBM2r is a pediatric glioblastoma patient-derived cell line.

This project will be continued in the Vaske lab. Future directions include repeating the *hydra* analysis with a larger cohort of pHGG recently included in the Treehouse cancer compendium (total=120, H3K27M=59), and correlating *hydra* clusters with co-mutation and clinical attributes.

Chapter 7: Summary of Doctoral Achievements

In addition to the projects detailed here, my achievements during my doctoral program involve authorship on other publications where I had a collaborative role, analysis of clinical pediatric cancer cases in the Treehouse Initiative, presentation of my work at research conferences, and mentoring and teaching younger students.

I have collaborated with Dr. Sameer Agnihotri's lab at University of Pittsburgh on several projects, providing computational analysis of *RAS* signaling, methionine metabolism, and developmental cell signaling in pediatric gliomas. As a result, I am a co-author on the 2019 *Cancer Research* publication "Identification of novel RAS signaling therapeutic vulnerabilities in Diffuse Intrinsic Pontine Gliomas"[61], and have contributed to other projects that are not yet published.

As a PhD student in the Treehouse Childhood Cancer Initiative, I served as the case analyst for many pediatric cancer cases from Stanford, UCSF and Children's Hospital of Orange County. Serving as a case analyst involves analyzing gene outlier and pathway enrichment data for an individual childhood cancer RNA-seq sample, to provide a summary of oncogenic overexpression and potential therapeutic directions. I have presented Treehouse analysis to oncologists in molecular tumor boards at these institutions, and I also served as the second reviewer for several cases. My involvement in Treehouse analysis led to my inclusion as a co-author on the 2019 *JAMA Network Open* publication "Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer"[12] and the *Scientific Data* publication "Barriers to accessing public cancer genomic data"[62].

Over the past several years, my work has been selected for presentation at several national and international research conferences. Less than a year into my doctoral program, I represented Treehouse at multiple events, including serving as a panelist at the UCSC Kraw Lecture Series on precision medicine in pediatric cancer, and giving an oral presentation at the Cancer Informatics for Cancer Centers Symposium. Additionally, I presented my research in poster sessions at the TGen Pediatric Precision Oncology Conference and the International Symposium on Pediatric Neuro-Oncology in 2018. I also presented the results of Treehouse comparative gene expression analysis in a Stanford-based clinical registry at the American Association for Cancer Research Advances in Pediatric Cancer Research Conference in 2019. My research abstract was also selected for an oral presentation at the Society of Neuro-Oncology Pediatric Research Conference in early 2019.

Finally, I have served as a mentor to several high school and undergraduate students during my PhD. I worked as a Graduate Student Teaching Assistant for two BME courses, and was awarded the 2016-2017 Outstanding Teaching Assistant Award. Additionally, during the summer of 2018 I was hired as an instructor to teach a workshop on cancer genomics for the Stanford Pre-Collegiate International Institutes (SPII). The SPII is designed to give a diverse, international group of high-school students from underrepresented populations the opportunity to experience American college life through academic workshops and social and cultural activities, and I contributed by teaching one of the few STEM workshops. I was selected for the 2017-

2018 UCSC T-32 Genome Sciences NIH Training Grant, and designed and led the NHGRI Bootcamp for incoming UCSC BME graduate students.

In conclusion, my doctoral work has addressed outstanding questions in clinical research, and particularly in pediatric neuro-oncology. My computational research has contributed to the field by demonstrating how gene expression can be used for identifying both known and novel molecular subtypes in pediatric cancers, and in characterizing therapeutic vulnerabilities and developmental origins in lethal pediatric brainstem gliomas. I have also developed a novel experimental organoid model for histone mutant gliomas, with unprecedented capabilities for intrinsically and flexibly characterizing tumorigenic origins, thus paving the way for finally understanding and effectively treating these deadly cancers.

References

1. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555: 371–376.

2. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555: 321–327.

3. Mackay A, Burford A, Carvalho D, Izquierdo E, Fazal-Salom J, Taylor KR, et al. Integrated Molecular Meta-Analysis of 1,000 Pediatric High-Grade and Diffuse Intrinsic Pontine Glioma. Cancer Cell. 2017;32: 520–537.e5.

4. Chen X, Pappo A, Dyer MA. Pediatric solid tumor genomics and developmental pliancy. Oncogene. 2015;34: 5207–5215.

5. Welby JP, Kaptzan T, Wohl A, Peterson TE, Raghunathan A, Brown DA, et al. Current Murine Models and New Developments in H3K27M Diffuse Midline Gliomas. Front Oncol. 2019;9: 92.

6. Filbin M, Monje M. Developmental origins and emerging therapeutic opportunities for childhood cancer. Nat Med. 2019;25: 367–376.

7. Hudson MM, Link MP, Simone JV. Milestones in the curability of pediatric cancers. J Clin Oncol. 2014;32: 2391–2397.

8. Nervi C, De Marinis E, Codacci-Pisanelli G. Epigenetic treatment of solid tumours: a review of clinical trials. Clin Epigenetics. 2015;7: 127.

9. Rodon J, Soria JC, Berger R, Batist G, Tsimberidou A, Bresson C, et al. Challenges in initiating and conducting personalized cancer therapy trials: perspectives from WINTHER, a Worldwide Innovative Network (WIN) Consortium trial. Ann Oncol. 2015;26: 1791–1798.

10. Mody RJ, Wu Y-M, Lonigro RJ, Cao X, Roychowdhury S, Vats P, et al. Integrative Clinical Sequencing in the Management of Refractory or Relapsed Cancer in Youth. JAMA. 2015;314: 913–925.

11. Chang W, Brohl AS, Patidar R, Sindiri S, Shern JF, Wei JS, et al. MultiDimensional ClinOmics for Precision Therapy of Children and Adolescent Young Adults with Relapsed and Refractory Cancer: A Report from the Center for Cancer Research. Clin Cancer Res. 2016;22: 3810–3820.

12. Vaske OM, Bjork I, Salama SR, Beale H, Tayi Shah A, Sanders L, et al. Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer. JAMA Netw Open. 2019;2: e1913968.

13. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539: 309–313.

14. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, et

al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science. 2018;360: 331–335.

15. Vladoiu MC, El-Hamamy I, Donovan LK, Farooq H, Holgado BL, Sundaravadanam Y, et al. Childhood cerebellar tumours mirror conserved fetal transcriptional programs. Nature. 2019. doi:10.1038/s41586-019-1158-7

16. Hovestadt V, Smith KS, Bihannic L, Filbin MG, Shaw ML, Baumgartner A, et al. Resolving medulloblastoma cellular architecture by single-cell genomics. Nature. 2019;572: 74–79.

17. Sanders LM, Rangaswami A, Bjork I, Lam DL, Beale HC, Kephart ET, et al. Comparative RNA-seq analysis aids in diagnosis of a rare pediatric tumor. Cold Spring Harb Mol Case Stud. 2019;5. doi:10.1101/mcs.a004317

18. Welcome to the Pan-Cancer Atlas. [cited 21 Apr 2020]. Available: https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html

19. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, et al. The genetic landscape of high-risk neuroblastoma. Nat Genet. 2013;45: 279–284.

20. Treehouse Public Data. [cited 21 Apr 2020]. Available: https://treehousegenomics.soe.ucsc.edu/public-data/

21. Grasso CS, Tang Y, Truffaux N, Berlow NE, Liu L, Debily M-A, et al. Functionally defined therapeutic targets in diffuse intrinsic pontine glioma. Nat Med.

2015;21: 555–559.

22. Moser HH, Panditharatna E, Packer RJ, Nazarian J. Diffuse Intrinsic Pontine Glioma: A Therapeutic Challenge. In: Agrawal A, editor. Neurooncology - Newer Developments. InTech; 2016.

23. Albright AL, Packer RJ, Zimmerman R, Rorke LB, Boyett J, Hammond GD. Magnetic resonance scans should replace biopsies for the diagnosis of diffuse brain stem gliomas: a report from the Children's Cancer Group. Neurosurgery. 1993;33: 1026–9; discussion 1029–30.

24. Puget S, Blauwblomme T, Grill J. Is biopsy safe in children with newly diagnosed diffuse intrinsic pontine glioma? Am Soc Clin Oncol Educ Book. 2012; 629–633.

25. Chan K-M, Fang D, Gan H, Hashizume R, Yu C, Schroeder M, et al. The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. Genes Dev. 2013;27: 985–990.

26. Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K, et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Res. 2017;77: e111–e114.

27. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158: 929–944.

28. Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, et al. Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. Stem Cell Reports. 2019;12: 245–257.

29. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science. 2017;355. doi:10.1126/science.aai8478

30. Mueller S, Jain P, Liang WS, Kilburn L, Kline C, Gupta N, et al. A pilot precision medicine trial for children with diffuse intrinsic pontine glioma - PNOC003: a report from the Pacific Pediatric Neuro-Oncology Consortium. Int J Cancer. 2019. doi:10.1002/ijc.32258

31. Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Mol Cancer Res. 2016;14: 3–13.

32. Goldman M, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. bioRxiv. 2019. p. 326470. doi:10.1101/326470

33. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35: 314–316.

34. Larson JD, Kasper LH, Paugh BS, Jin H, Wu G, Kwon C-H, et al. Histone H3.3

K27M Accelerates Spontaneous Brainstem Glioma and Drives Restricted Changes in Bivalent Gene Expression. Cancer Cell. 2019;35: 140–155.e7.

35. Lewis PW, Müller MM, Koletsky MS, Cordero F, Lin S, Banaszynski LA, et al. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. Science. 2013;340: 857–861.

36. Juratli TA, Qin N, Cahill DP, Filbin MG. Molecular pathogenesis and therapeutic implications in pediatric high-grade gliomas. Pharmacol Ther. 2018;182: 70–79.

37. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat Methods. 2019;16: 619–626.

38. Nowak D, Stewart D, Koeffler HP. Differentiation therapy of leukemia: 3 decades of development. Blood. 2009;113: 3655–3665.

39. Campos B, Wan F, Farhadi M, Ernst A, Zeppernick F, Tagscherer KE, et al. Differentiation therapy exerts antitumor effects on stem-like glioma cells. Clin Cancer Res. 2010;16: 2715–2728.

40. Reynolds CP. Differentiating agents in pediatric malignancies: retinoids in neuroblastoma. Curr Oncol Rep. 2000;2: 511–518.

41. Marton RM, Miura Y, Sloan SA, Li Q, Revah O, Levy RJ, et al. Differentiation

and maturation of oligodendrocytes in human three-dimensional neural cultures. Nat Neurosci. 2019;22: 484–491.

42. Tan JME, Cook ECL, van den Berg M, Scheij S, Zelcer N, Loregger A. Differential use of E2 ubiquitin conjugating enzymes for regulated degradation of the rate-limiting enzymes HMGCR and SQLE in cholesterol biosynthesis. Atherosclerosis. 2019;281: 137–142.

43. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550: 204–213.

44. Courtney R, Landreth GE. LXR Regulation of Brain Cholesterol: From Development to Disease. Trends Endocrinol Metab. 2016;27: 404–414.

45. Vitali C, Wellington CL, Calabresi L. HDL and cholesterol handling in the brain. Cardiovasc Res. 2014;103: 405–413.

46. Venkatesh HS, Morishita W, Geraghty AC, Silverbush D, Gillespie SM, Arzt M, et al. Electrical and synaptic integration of glioma into neural circuits. Nature. 2019. doi:10.1038/s41586-019-1563-y

47. Venkatesh HS, Johung TB, Caretti V, Noll A, Tang Y, Nagaraja S, et al. Neuronal Activity Promotes Glioma Growth through Neuroligin-3 Secretion. Cell. 2015;161: 803–816.

48. Fancy SPJ, Kotter MR, Harrington EP, Huang JK, Zhao C, Rowitch DH, et al. Overcoming remyelination failure in multiple sclerosis and other myelin disorders. Exp Neurol. 2010;225: 18–23.

49. Kim JH, Cox ME, Wasan KM. Effect of simvastatin on castration-resistant prostate cancer cells. Lipids Health Dis. 2014;13: 56.

50. Pfeil J, Sanders LM, Anastopoulos I, Lyle AG, Weinstein AS, Xue Y, Blair A, Beale HC, Lee A, Leung SG, Dihn PT, Shah AT, Breese MR, Devine WP, Bjork I, Salama SR, Sweet-Cordero EA, Haussler D, Vaske OM. Hydra: A mixture modeling framework for subtyping pediatric cancer cohorts using multimodal gene expression signatures. PLoS Comput Biol. 2020.

51. Lenz M, Müller F-J, Zenke M, Schuppert A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Sci Rep. 2016;6: 25696.

52. Angell H, Galon J. From the immune contexture to the Immunoscore: the role of prognostic and predictive immune markers in cancer. Curr Opin Immunol. 2013;25: 261–267.

53. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4: 2612.

54. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust

enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12: 453–457.

55. Yang I, Han SJ, Kaur G, Crane C, Parsa AT. The role of microglia in central nervous system immunity and glioma immunology. J Clin Neurosci. 2010;17: 6–10.

56. Gieryng A, Pszczolkowska D, Walentynowicz KA, Rajan WD, Kaminska B. Immune microenvironment of gliomas. Lab Invest. 2017;97: 498–518.

57. Misuraca KL, Barton KL, Chung A, Diaz AK, Conway SJ, Corcoran DL, et al. Pax3 expression enhances PDGF-B-induced brainstem gliomagenesis and characterizes a subset of brainstem glioma. Acta Neuropathol Commun. 2014;2: 134.

58. Liu Y, Zhu H, Liu M, Du J, Qian Y, Wang Y, et al. Downregulation of Pax3 expression correlates with acquired GFAP expression during NSC differentiation towards astrocytes. FEBS Lett. 2011;585: 1014–1020.

59. Stolt CC, Rehberg S, Ader M, Lommes P, Riethmacher D, Schachner M, et al. Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10. Genes Dev. 2002;16: 165–170.

60. Jessa S, Blanchet-Cohen A, Krug B, Vladoiu M, Coutelier M, Faury D, et al. Stalled developmental programs at the root of pediatric brain tumors. Nat Genet. 2019;51: 1702–1713.

61. Koncar RF, Dey BR, Stanton A-CJ, Agrawal N, Wassell ML, McCarl LH, et

al. Identification of Novel RAS Signaling Therapeutic Vulnerabilities in Diffuse Intrinsic Pontine Gliomas. Cancer Res. 2019;79: 4026–4041.

62. Learned K, Durbin A, Currie R, Kephart ET, Beale HC, Sanders LM, et al. Barriers to accessing public cancer genomic data. Sci Data. 2019;6: 98.