Machine Learning Applications for Neurological Diseases

by

Maxwell P. Gold

A.B. Molecular Biology, Princeton University (2014)

Submitted to the Graduate Program of Computational and Systems Biology
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Computational and Systems Biology

at the

Massachusetts Institute of Technology

September 2022

Author .....
Maxwell P Gold
Computational and Systems Biology Program
July 22, 2022

Certified by .....
Ernest Fraenkel
Professor of Biological Engineering
Thesis Supervisor

Accepted by.....
Christopher B. Burge
Professor of Biology
Director, Computational and Systems Biology Graduate Program

Machine Learning Applications for Neurological Diseases

by

Maxwell P. Gold

Submitted to the Graduate Program in Computational and Systems Biology
on July 22nd, 2022 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

Abstract

Neurological conditions affect the brain and other parts of the nervous system. This includes neurodegenerative diseases like Huntington's Disease, psychiatric conditions like schizophrenia, and brain cancers like glioblastoma. These conditions are particularly challenging to study because they affect such a vital and complex organ system, making it difficult to understand disease etiology and to develop high-quality model systems.

Because of these challenges, experiments studying neurological diseases typically either contain very few patient samples or are collected from imperfect model systems. Machine learning approaches have proven helpful for processing these types of datasets and identifying relevant biological signal. In this thesis, I detail five examples of the utility of machine learning methods for analyzing neurological disease data. Some chapters focus primarily on the development of novel machine learning methods, while others discuss the implementation of established algorithms leading to significant advancements in our understanding of the given disease.

Chapter 2 details a novel gene set scoring algorithm that significantly improves upon existing methods. This new approach is particularly useful for analyzing single-cell transcriptomics assays, which are becoming increasingly common in neurological disease studies. In Chapter 3, I describe how multi-omic integration of ATAC-Seq, ChIP-Seq, and RNA-seq data revealed a novel population of cycling cells relevant to Huntington's Disease models. In Chapter 4, I discuss an improved multi-commodity flow algorithm for omics data integration and highlight its utility for understanding drug effects in glioblastoma. Chapter 5 highlights how clustering and the Prize-Collecting Steiner Forest algorithm led to a better understanding of proteomic subtypes in medulloblastoma tumors. Lastly, Chapter 6 expands upon the work in Chapter 5, and details how I used computational approaches to figure out that some medulloblastoma tumors contain cells recapitulating cerebellar granule neuron development.

In summary, this thesis showcases the value machine learning techniques for analyzing the small, complicated datasets typically found in neurological disease experiments. Throughout this work, I emphasize the importance of collecting and integrating multiple types of biological data to get a more complete understanding of these conditions.

Thesis supervisor: Ernest Fraenkel
Title: Professor of Biological Engineering

Acknowledgements

When I started college, I never imaged I'd pursue a PhD. In fact, I only worked with Dr. Jacques Fresco at Princeton because I was required to do an undergraduate thesis, but this experience changed my life. Jacques single-handedly taught me to love research with his dedication, passion, and curiosity and I cannot thank him enough for mentorship. When I graduated, I was hired to a research position at a biotech startup, but within weeks, I was asked to switch roles and learn how to code. I want to thank Alex Bisignano for the opportunity to work at Recombine and Nick De Vaux, Roman Shraga, and Charlotte Rivera for their patience and mentorship.

As I was learning about coding, I realized all I wanted to with my new computational skills was become a better biologist. So, I decided to pursue a PhD in computational biology and was fortunate that the Computational and Systems Biology (CSB) program took a chance on someone who had never taken a formal computer science class. Thank you to Chris Burge, Jacquie Carota, and the whole CSB community for their support over the last six years.

When I told Jacques I was pursuing a PhD, his one piece of advice was that my advisor was more important than my research topic, and I certainly took that to heart when joining the Fraenkel Lab. I am consistently in awe of Ernest's efficiency, rigorous standards, and clear communication. I want to thank him for being an incredible advisor/mentor and giving me the freedom to explore many many directions during my PhD. I also want to thank Doug Lauffenberger and Matt Vander Heiden for helping to significantly improve the quality of this work with their engagement and thoughtful comments during my thesis committee meetings. I have also had the fortune of collaborating with many other labs during my PhD and I really appreciate everyone who helped generate the data and ideas presented in this work. Specifically, I want to thank Scott Pomeroy, Jill Mesirov, and Rob Wechsler-Reya for being incredible role models for how to do collaborative science in a productive and fun manner.

I also want to thank the whole Fraenkel Lab for everything over the last six years. Specifically, I need to shout out Divya for being an incredible deskmate and friend and Tobi for his guidance on the medulloblastoma project. Additionally, I want to thank Mandy, Natasha, and Johnny for their consistent mentorship and Andrew for his invaluable help in finishing the wet lab experiments that let me graduate. To everyone I was unable to mention by name, I appreciate you making the Fraenkel Lab such an enjoyable place to work during my time. I also want to thank MIT for the many opportunities it offers outside of the classroom. Teaching through the educational studies program and playing club volleyball were integral parts of my PhD experience. Additionally, teaching "A Brief History of Kanye" was one of the highlights of graduate school and I am incredibly appreciative of the IAP program for allowing me to do so.

Furthermore, I need to thank the important people in my life outside of school. I am so appreciative of my incredible friends from high school, camp, college, and graduate school who remind me of the importance of life outside of work. I also need to thank my fiancé Kristin, who has a remarkable ability to make any experience better, and who somehow convinced me four months before my thesis was due would be a good time get a puppy. And finally, thank you to my family. There is no way I would be here without your constant support. My siblings Jason and Nicki are both smart, thoughtful, caring people who I look up to, even though they're younger. My parents David and Ruth are remarkable role models who have always encouraged me to pursue my passions and taught me the patience and dedication needed to finish my PhD.

I grew a tremendous amount as a person and scientist during these six years, and I am truly grateful that I was given this opportunity.

Table of Contents

Chapter 1– Introduction

1.1 Summary of relevant neurological diseases

Neurological disorders are conditions that affect any part of the nervous system (Patel et al., 2016). This includes a broad range of diseases, such as neurodegenerative disorders, psychiatric conditions, and brain cancers. These diseases are especially difficult to study because they involve complex, crucial organs. This thesis describes machine learning methods that are generally applicable to neurological disease datasets, but focuses on applications for three conditions (medulloblastoma, glioblastoma, and Huntington's disease), summarized in this section.

1.1.1 **Medulloblastoma (MB)**

Medulloblastoma (MB) is one of the most common malignant pediatric brain tumors. There are around 1000 new cases each year and the mean patient age is estimated to be between 3 and 7 years old (Fogarty et al., 2005). The standard treatment protocol of surgical resection, radiation, and chemotherapy has led to relative high survival rates (85% 5-year survival for average risk patients and 70% survival for high-risk patients). Despite these good outcomes, the toxic therapies unfortunately can also lead to severe neurological side effects and increased risks of secondary cancers (Archer et al., 2017; Gajjar et al., 2006). Thus, there is an urgent need for more targeted, less dangerous treatments.

Using precision medicine in medulloblastoma has been extremely challenging since these tumors are so heterogeneous. The World Health Organization recognizes both histological and molecular subtypes of MB (Louis et al., 2016). There are four primary histological patterns. Classic tumors are composed of tightly packed undifferentiated cells, while the large-call anaplastic (LCA) tumors are characterized by cytological pleomorphism and high rates of mitosis. The desmoplastic/nodular (DNMB) tumors have tightly packed cells interrupted by less dense nodules, typically filled with NeuN+ cells. Tumors with extensive amounts of nodularity (MBEN) are designated as their own histological category.

The four consensus molecular subgroups of MB are named WNT-activated, SHH-activated, Group 3, and Group 4 (Taylor et al., 2012). The WNT-activated and SHH-activated tumors have overexpression of the wingless and sonic hedgehog pathways respectively. Group 3 and Group 4 tumors are less well understood, but molecular studies have shown that MYC-activated Group 3 tumors have particularly poor survival and that Group 4 tumors are primarily composed of differentiated cells (Archer et al., 2018; Cavalli et al., 2017; Cho et al., 2011; Northcott et al., 2017).

Since the establishment of the four consensus molecular subgroups, it has been well documented that each individual subgroup contains its own heterogeneity. Studies using genomics (Northcott et al., 2017; Taylor et al., 2012), transcriptomics (Cavalli et

al., 2017; Cho et al., 2011) and proteomics (Archer et al., 2018; Forget et al., 2018) have highlighted more subdivisions within each large subgroup. Additionally, single-cell transcriptomic studies have characterized the undifferentiated and differentiated cells in these tumors (Hovestadt et al., 2019; Riemondy et al., 2022; Vladoiu et al., 2019). These efforts have led to new potential therapeutic directions targeted at specific subtypes, such as differentiation therapies for SHH MBs (Cheng et al., 2020).

This thesis details significant advancements in our understanding of MB heterogeneity. First, I describe our proteomic and phosphoproteomic analysis of MB tumors, which revealed novel, clinically relevant subtypes. Then, I discuss work showing that tumor cells in the SHH-activated subtype of MB can recapitulate cerebellar development. This finding has led to insights about tumor subtypes, mutations, metabolism, and histology.

1.1.2 **Glioblastoma (GBM)**

Glioblastoma (GBM) is a highly-aggressive brain tumor with an incidence of 5 per 100,000 in North America (Stupp et al., 2010). GBM has a terrible prognosis, with the 5-year survival rate being 5% and death typically occurring within the first 16 months after diagnosis (Holland, 2000; Oronsky et al., 2021). GBM is treated with a surgical resection, if possible, but this is often not feasible given the invasiveness of the tumor. Patients are then typically given the alkylating agent Temozolomide and radiation therapy (Stupp et al., 2005), neither of which are particularly effective.

Given the incredibly low survival rate of GBM, there is a significant need for efficacious therapeutics. Like medulloblastoma, researchers have tried to identify molecular subtypes that can inform drug target identification. The primary subdivision in GBM is between IDH-wildtype (IDH-wt) and IDH-mutant (IDH-mt) tumors. IDH-mt cases have a mutation in one of the isocitrate dehydrogenase (IDH) genes, leading to excessive conversion of alpha-ketoglutarate into 2-hydroxyglutarate. This change depletes carbohydrates from the TCA cycle and produces excessive amounts of the oncometabolite 2-hydroxyglutarate, which affects tumor metabolism and epigenetics (Han et al., 2020).

Molecular profiling of IDH-wt GBMs revealed three consensus subtypes named classical (CL), mesenchymal (MES), and proneural (PN) (Verhaak et al., 2010; Wang et al., 2017). Each subgroup has a distinct molecular profile and specific genomic events. For example, EGFR mutations occur almost exclusively in classical GBMs. Recent single-cell transcriptomics analyses revealed that IDH-wt glioblastomas contain four primary classes of malignant cell types: Oligodendrocyte Precursor-Like (OPC-like), Neural Precursor-Like (NPC-like), Astrocyte-Like (AC-like), and Mesenchymal-Like (MES-like) (Neftel et al., 2019). Each of the bulk subtypes are associated with specific cell types. The proneural tumors are primarily composed of NPC-like and OPC-like cells, while the classical tumors contain AC-like and MES-like cells. Additionally, the mesenchymal tumors are composed almost entirely of MES-like cells.

This thesis explores subtype-specific drug response in GBM and details computational approaches for understanding drug sensitivity and resistance in the proneural and mesenchymal subgroups.

1.1.3 **Huntington's Disease (HD)**

Huntington's disease (HD) is a neurodegenerative disorder that affects approximately 10.6 – 13.7 people per 100,000 (Ross and Tabrizi, 2011). HD is a monogenic dominant genetic disease caused by a mutation in the huntingtin (*HTT*) gene on chromosome 4 (MacDonald et al., 1993). This gene typically contains less than 36 CAG repeats and any person with greater than 39 CAG repeats is almost certain to develop HD (McColgan and Tabrizi, 2018; Ross and Tabrizi, 2011).

HD often presents between 35 and 45 years of age (Walker et al., 1981), but this age of onset is affected by the length of CAG repeats in the *HTT* gene and mutations in other genes (Djoussé et al., 2004; Valcárcel-Ocete et al., 2015). The first symptom is almost always involuntary jerky motions, known as chorea, which become more pronounced as the disease progresses. Additionally, HD is associated with cognitive impairments, such as decreased executive functioning, and neuropsychiatric symptoms like depression and psychosis.

It is hypothesized that HD symptoms are caused by the dysfunction and death of specific neurons in the brain. In HD, the first brain cells to die are the medium spiny neurons of the striatum, a region of the brain related to motion and reward processing (Hollerman et al., 2000). As HD progresses, cells in cerebral cortex and nucleus accumbens are also affected (Ross and Tabrizi, 2011). This pattern likely explains why motor symptoms occur first and cognitive effects appear later.

Despite HD being caused by a single genetic mutation, it is not well understood how extra CAG repeats in the *HTT* gene lead to neuron dysfunction, cell death, and eventually Huntington's disease. It is hypothesized that the molecular effects of *HTT* mutation are a combination of a loss of function (i.e. only having one healthy *HTT* gene) and gain of function (i.e. the mutant *HTT* gene developing new toxic properties). The mutant *HTT* expansion has been linked to many molecular pathologies, such as synaptic dysfunction, bioenergetic deficiencies and protein aggregates (McColgan and Tabrizi, 2018; Ross and Tabrizi, 2011), but it is still not clear how these specific molecular disturbances manifest as HD and why the age of onset is so late in life.

In this thesis, I highlight analysis of iPSC-derived neurons, whereby human fibroblasts are reprogrammed into stem cells and then differentiated into neurons. We found that the iPSC-derived neurons from HD patients contain a distinct population of cycling cells not present in neurons derived from healthy controls.

1.2 Overview of Disease-Specific Model Systems

It is particularly challenging to create model systems for neurological conditions because the human nervous system is so complex that it is difficult to faithfully model in an animal or cell line. Additionally, these diseases are typically multifactorial and their pathophysiology is not well understood. Still, there are model systems for many of these conditions and this section provides a brief overview of the most common models used to study the conditions discussed in this thesis.

1.2.1 **Medulloblastoma (MB)**

Most observational studies of MB use tumor tissue resected during surgery. This model allows researchers to study the diseased tissue directly, but unfortunately it can be challenging to obtain these samples given the low incidence of the tumor. The rareness of MB also makes it difficult to run assays requiring fresh cells, such as single-cell RNA-sequencing, and hard to justify performing experiments like bulk proteomics which require large amounts of tissue.

There are many cell lines used in MB research (Casciati et al., 2020), but it is debatable how well they model human MB. MYC-activated Group 3 cell lines (like D458) are considered the most reliable cell line models. Mouse models are also commonly used in MB research and they are usually either genetically engineered mouse (GEM) models or patient-derived xenograft (PDX) models (Roussel and Stripay, 2020). GEM models use a Cre system (Feil et al., 2009) to induce cell-type specific activation or repression of a gene to promote tumor formation. For example, PTCH1 gene knockouts in Math1+ granule precursor cells lead to MB formation in mice (Li et al., 2013). The best GEM models mimic SHH MB and MYC-activated G3 MB as these tumor models contain many of the same cell types identified in human MB (Riemondy et al., 2022).

Patient-derived xenograft (PDX) models of MB are created by injecting cells from human tumors into mice brains; these mice are typically immunocompromised to the raise the chances of tumor growth (Roussel and Stripay, 2020). The most successful PDX models come from aggressive forms of MB, like TP53-Mt SHH MB and MYC-activated G3 MB. PDX models are a great resource for studying aggressive MBs, but the lack of high-quality mouse models for other subgroups makes it challenging to perform interventional studies on the less aggressive forms of MB.

This work focuses primarily on observational data collected using resected tumor tissue and highlights one example of a validation study performed in a Group 3 cell line that shows the importance of PRKDC in MYC-activated MBs.

1.2.2 **Glioblastoma (GBM)**

The types of models used in GBM studies are quite similar to MB models. Most observational studies in GBM use surgically resected tumor tissue, while interventional experiments rely on cell lines or mouse models (Kijima and Kanemura, 2017). There are

many commercially available cell lines (such as U87 and T98G), but this thesis also discusses cell lines developed in the department of neurosurgery in Heidelberg: NCH711d, NCH705, NCH421K, and NCH644.

Like MB, there are GEM models and xenograft-derived models of GBM. These GEM models target specific genes that induce GBM, such as PDGF-B and KRAS (Huszthy et al., 2012). There are many patient-derived xenografts for GBM, but some studies also use cell-line derived xenografts, which are created by injecting well-established GBM cell lines into mice. These cell-line based xenografts have higher rates of engraftment and growth than PDX models, but the patient-derived xenografts more faithfully recapitulate human GBM tumor heterogeneity (Wakimoto et al., 2012).

In Chapter 4, I discuss drug treatment experiments performed in GBM cell lines developed at DKFZ in Heidelberg. These cell lines were associated with a given GBM subtype (CL, MES, or PN) through transcriptional profiling and then comparing the expression to human tumors with known subtypes.

1.2.3 **Huntington's Disease (HD)**

Like MB and GBM, HD experiments are typically performed with cell lines and mice, but there are key differences between modeling neurodegenerative diseases and brain cancers. Primarily, no tissue is surgically resected for HD treatment, making it difficult to directly study the human disease. Observational studies can be performed using postmortem brain tissue from HD patients, but this model has serious flaws. First, these samples only capture the latest stages of HD and thus are not great models for investigating disease etiology. Additionally, when an HD patient passes away, most of the striatal neurons have died, making it challenging to investigate the region of the brain most affected by HD.

Because of these limitations, animal and cell line models are widely used in HD research. The most common animal models are mice that contain a version of the human *HTT* or mouse *Htt* gene with high numbers of CAG repeats (Ramaswamy et al., 2007). Some mouse models contain the full human *HTT* gene, while others only have a small portion of the gene that contains the CAG repeats. The models with the full-length *HTT* gene more faithfully replicate human HD, but have slower progression and thus lead to significantly longer experiments (Shenoy et al., 2022). These mouse models have also been used to create cell line models of HD, like the STHdh^{Q111} model, which was developing using striatal neurons from the HdhQ111 mouse (Trettel et al., 2000).

In recent years, neurons derived from induced pluripotent stem cells (IPSC-derived neurons) have been used to study HD (Lim et al., 2017). For this model, fibroblasts or blood cells from HD patients are reprogrammed into pluripotent stem cells and induced to differentiate into neural cultures that contain neurons, neural progenitors, and glia. This model system allows for studying human neurons with the *HTT* gene expansion but can be hard to work with because of the challenging differentiation protocol and the heterogeneous samples. In this thesis, I present a multi-omic characterization of these

iPSC-derived neurons and discuss a novel cell type only observed in neuron cultures from HD patients.

1.3 Summary of relevant biological data types

"Omics" generally refers any to any study seeking to characterize or quantify large amounts of biological molecules at once (Hasin et al., 2017). In the last two decades, significant technological advancements have increased the quantity and quality of biological data produced. This omics revolution has been extremely exciting, but has also created new challenges for how to interpret and analyze these large and confusing datasets. In this section, I provide an overview of the primary omics data types discussed in this thesis.

1.3.1 **Genomics**

Genomics refers to the analysis of DNA sequences. There are two main technologies used for genomic studies. The first is the DNA microarray, a limited but cost-effective assay for collecting information about select positions on the genome. DNA microarrays use 1000s of user-specified oligonucleotides, called probes, which are designed to bind a specific spot in the genome. This technology allows for detecting the identity of a single nucleotide (adenine, cytosine, guanine, or thymine) at those specified locations. Additionally, the hybridization rates of positionally sequential probes can be used to infer copy number variations (CNVs), mutations that cause gains or losses in regions of the genome.

The biggest limitation of microarrays is that they can only detect the user-specified mutations, and thus cannot be used to identify de novo mutations. Next Generation DNA Sequencing (NGS) solves this problem because it can characterize the nucleotide sequences for entire genomes (Reis-Filho, 2009). There are many methods for performing NGS, but the most popular system was developed by Illumina (Quail et al., 2012). For this protocol, DNA is extracted, fragmented into smaller bits, and then amplified through PCR. This DNA is then sequenced using modified nucleotides that compete to bind with the extracted DNA (Mardis, 2008). These nucleotides contain a fluorescent tag for identification and a 3' blocker that prevents other nucleotides from being added to the 3' end. When a nucleotide successfully hybridizes to DNA, the fluorophore is imaged to get the nucleotide identity and then the fluorescent tag and 3' blocker are terminated allowing for the next nucleotide to bind. This cycle repeats for hundreds of bases to create a sequence of DNA that can be mapped to a reference genome. In this thesis, I discuss medulloblastoma genomics experiments that characterized single nucleotide polymorphisms and copy number variations in these tumors.

1.3.2 **Epigenomics**

The exact definition of “epigenomics” is often debated (Greally, 2018). The term is often used refer to describe the study of chromatin, a mixture of DNA and histone proteins, that controls the accessibility of DNA. Inaccessible DNA is wrapped tightly around histones and cannot be reached by proteins like transcription factors (TFs), which alter the transcription of other genes. When the DNA is needed for cellular processes like replication and transcription, the chromatin can become unwound, leaving the DNA accessible to the relevant factors. Epigenomics can also refer to the study of any modifications made to DNA or chromatin that can affect transcription. Methylation of specific nucleic acid bases can affect nearby gene transcription and post-translational modifications to histone proteins can control nearby DNA accessibility.

Epigenetic factors play a significant role in neurological conditions (Dubuc et al., 2012; Jakovcevski and Akbarian, 2013). Changes in chromatin accessibility can alter transcription patterns and even affect the developmental potential of a cell without directly altering transcription (Zhang et al., 2018). This section details the main assays used to collect epigenomic information (Mehrmohamadi et al., 2021).

1.3.2.1 ChIP-Seq

ChIP-Seq combines chromatin immunoprecipitation (ChIP) (Milne et al., 2009) and high-throughput DNA Sequencing to identify genomic sequences bound by particular proteins. This method uses antibodies to target proteins or chromatin modifications to pull down nearby DNA fragments for sequencing.

ChIP-Seq is frequently used to identify genomic regions bound by a specific TF in a given tissue or cell type. Additionally, this method can be used to target relevant histone modifications, such as methylation of histone protein 3 at its fourth lysine (H3K4), which indicates that a genomic region undergoing active transcription (Barski et al., 2007). In this thesis, we collected ChIP-Seq from iPSC-derived neurons derived from HD and control fibroblasts.

1.3.2.2 ATAC-Seq

ATAC-Seq is used to identify regions of accessible chromatin (Buenrostro et al., 2015). This method uses the Tn5 transposase which inserts into open regions of the genome and tags and fragments the DNA for sequencing. This method allows for the comparison of chromatin accessibility between conditions. Typically, the differentially accessible regions are further analyzed for proximity to relevant genes or enrichment for transcription factor binding motif sequences. We also collected ATAC-Seq data for HD and control iSPC-derived neurons and integrated this information with the ChIP-Seq data.

1.3.2.3 Methylomics

DNA methylation affects transcription in complicated ways, sometimes increasing expression levels of nearby genes and other times decreasing it (Wagner et al., 2014). Regions of methylation are typically assayed using a microarray or a special type of DNA-sequencing. Microarray-based methods are designed to detect 5-methylcytosine at specific locations in the genome. Additionally, bisulfite sequencing converts all unmethylated cytosines in DNA to uracil, which can then be run through DNA Sequencing to identify methylated nucleotides (Hayatsu, 2008). The resulting methylation data can reveal information about transcriptional regulation and can also be used to identify clinically relevant subtypes of tumors (Capper et al., 2018). In this thesis, I present a analyses of methylation data where we tried to better understand the epigenetic differences between novel subtypes of MB.

1.3.3 **Transcriptomics**

Transcriptomics assays measure RNA expression levels. Like genomics, these experiments are performed using a microarray or an NGS-based approach (Raghavachari and Garcia-Reyero, 2018). Microarray-based transcriptomics uses a similar principle to CNV detection in genomics; RNA is reverse transcribed into cDNA and a microarray with specified probes is used to infer relative RNA expression from hybridization rates (King and Sinha, 2001).

Like the genomics assays, microarrays are more cost-effective but are limited in their detection capabilities. RNA-sequencing (RNA-seq) (Wang et al., 2009) uses NGS technologies to collect more comprehensive RNA expression information for a sample. In this method, RNA is reverse transcribed into cDNA, which is then run through DNA-Seq and mapped to the appropriate regions of the genome. RNA-seq allows for the detecting novel features, such a new gene fusions or mRNA isoforms. In this thesis, we analyze RNA expression profiles from MB tumors, GBM tumors, and HD iPSC-derived neuron cultures.

1.3.3.1 Single Cell Transcriptomics

In recent years, technological advancements have allowed for the measurement of gene expression profiles from single cells (Chen et al., 2019). There are many methods to perform single-cell RNA-sequencing (scRNA-seq). The main protocols differ primarily in cell isolation and RNA capture methodology and there is typically a tradeoff between the number of sequenced cells and how many genes can be detected. The quality of RNA-seq for each cell is lower than for bulk tissue and there are known issues with non-biological zeros, whereby the RNA for a gene will not be measured even though it is present (Jiang et al., 2022). Still, there is tremendous value in observing the transcriptomic profiles from many individual cells in a sample. These technologies have helped characterize the intra-tumoral heterogeneity within MB tumors (Hovestadt et al., 2019; Riemondy et al., 2022; Vladoiu et al., 2019) and GBM tumors (Neftel et al., 2019) and revealed issues with astrocytes in HD that are not obvious when studying the expression profiles from bulk brain samples (Al-Dalahmah et al., 2020). In this thesis,

we generated novel scRNA-seq data for both MB and HD studies to better characterize the model systems of interest.

1.3.4 **Proteomics**

Proteomics assays are used for protein identification and quantification. These methods begin with protein digestion, leaving smaller fragments known as peptides that are separated through liquid chromatography (LC) (Karpievitch et al., 2010). The molecules are then ionized and sent into a mass spectrometer (MS), where the time of flight can be used to identify the mass to charge ratio (m/z) of the peptide. Sometimes multiple rounds of MS will be repeated to get more exact measurements. These digested peptides are then mapped to whole proteins by matching m/z values to known databases and through inferences about the fragmentation patterns.

There are many methods for quantifying protein levels. For label-free (LF) quantification, each sample is run through LC-MS individually and the relative abundances are determined by comparing the intensities in the mass spectrometer. Additionally, labeling methods allow for more accurate relative quantification between samples. The most common labeling techniques are stable isotope labeling by amino acids in cell culture (SILAC) (Chen et al., 2015; Oda et al., 1999) and tag-based systems like iTRAQ (isobaric tags for relative or absolute quantification) and TMT (tandem mass tags) (Thompson et al., 2003). SILAC is limited to cell culture experiments and works by labeling one condition with heavy amino acids; quantification is then performed by comparing the ratio of heavy to light peptides. Tagging methods chemically label each peptide, which allows for accurate relative quantification since all samples can be run through a single MS experiment. In this thesis, I detail novel proteomics datasets collected from MB tumors and GBM cell line models.

1.3.4.1 Phosphoproteomics

Protein activity is often regulated by the addition or subtraction of phosphate groups to specific amino acids (serine, threonine, and tyrosine). Phosphoproteomics dedicated to the detection and quantification of these phosphorylated proteins. These experiments use the same steps as proteomics except there is an initial enrichment for proteins with phosphorylated residues. Additionally, the downstream computational mapping uses specific databases generated for phosphorylated proteins. In this thesis, I also detail phosphoproteomic experiments from MB tumors and GBM cell lines. In addition to identifying differential phosphoproteins between conditions, we also analyzed specific phosphosites to better understanding subtype-specific kinase activity.

1.3.4.2 Spatial Proteomics

Proteomics and phosphoproteomics assays characterize and quantify proteins, but these assays lose all information about how these proteins are spatially organized in a sample. Proteins can be markers of cell type, like CD3 for T-cells, or functional state,

such as Ki67 for actively cycling cells, and for complicated neurological disorders, there is great interest in understanding how these proteins are co-expressed and spatially arranged. This information can help determine the intra- and inter-cellular forces driving disease progression.

Immunohistochemistry (IHC) is commonly used to detect the spatial expression of a small number of proteins (Coons et al., 1941; Ramos-Vara and Miller, 2014). IHC uses primary antibodies to bind targets of interest and then fluorescent secondary antibodies that allow for detection and quantification of the targeted protein. Since most microscopes can only distinguish a few fluorophores at once, this is assay can only detect a very small number of proteins at once (typically less than 5).

Many multiplexed immunohistochemistry techniques have been developed to address this limitation. Cyclic immunofluorescence (CyCIF) allows for the detection of 10s of proteins using cycles of imaging (Lin et al., 2016). For this assay, primary antibodies are directly conjugated to fluorophores, removing the need for secondary antibodies. After a round of imaging, the samples are photobleached to remove the initial fluorescence, allowing for another set of fluorescent antibodies to be added and imaged.

Another similar method is Co-Detecting by Indexing (CODEX) (Goltsev et al., 2018), which uses successive cycles of antibodies conjugated to nucleotides. The antibodies can be distinguished using special fluorescent nucleotides, where the fluorophores are removed each cycle. These multiplexed technologies allow for many proteins to be imaged at once, substantially increasing the potential for understanding the spatial organization of cell types and functional features. In Chapter 6 of this thesis, I highlight the power of CyCIF for analyzing the spatial organization of cell types in MB.

1.3.5 **Metabolomics**

Metabolomics is the study of small molecules, such as sugars, nucleic acid bases, amino acids, and fatty acids. These molecules play crucial roles in health and disease and studying them can reveal important changes in processes like bioenergetics and epigenomics.

Metabolites are analyzed through untargeted and targeted approaches (Roberts et al., 2012). Untargeted metabolomics is similar to the proteomics methods discussed above, whereby metabolites are extracted from a biological samples and LC-MS is used to collect metabolite information from the retention time in a mass spectrometer. This approach produces many m/z peaks, but unfortunately only a small portion of them can be mapped to known metabolites using the m/z value or spectral pattern. Targeted approaches rely on using analytical standards to validate metabolomic LC-MS patterns to ensure their identity and then measuring the intensity of sample metabolites that match those known profiles.

1.3.5.1 Spatial metabolomics

Like spatial proteomics, there is significant interest in understanding the spatial relationships among metabolites within a biological sample. MALDI-IMS (Caprioli et al., 1997; Stoeckli et al., 2001) uses a laser to vaporize and ionize molecules at specific locations and then sends those molecules to a mass spectrometer. Unfortunately, this assay cannot provide single-cell resolution, but current technologies allow for analyzing small spots that likely contain less than 5 cells. This technique is typically untargeted and relies on matching the m/z peaks to a known database. In this thesis, I present MALDI-IMS data which highlights the inter- and intra-tumoral metabolic heterogeneity of SHH-activated MBs.

1.4 Summary of relevant machine learning methods

As noted previously, neurological diseases are particularly difficult to study because it is challenging to develop good model systems and the diseases themselves are often multifactorial conditions with unknown etiology. Even in cases like HD where the disease is caused by a single mutation, the complexity of the nervous system makes it challenging to understand how this single genetic alteration leads to neurological symptoms. Machine learning techniques have proven helpful in finding relevant biological signal from these small, noisy datasets. This section provides background for many of the methods and algorithms discussed in this thesis.

1.4.1 **Autoencoders**

Autoencoders are used to learn a low dimensional representation of data. For omics studies, datasets are often represented as a matrix, where the rows are samples, the columns are biological features (like genes). These datasets typically contain far more features (~20,000 genes for bulk RNA-seq) than samples (typically 10s to 100s for bulk RNA-seq), which makes it difficult to find meaningful relationships between samples or identify biologically relevant features. Autoencoders reduce these complex features to a small number of latent variables that summarize much of the variability in the dataset.

The basic unit of an autoencoder is known as a neuron and it is composed of an input (x), a weights matrix (w), and a bias term (b). The output of this neuron is summarized as $\phi(w^Tx + b)$, where ϕ is some activation function. Φ is frequently a non-linear function, such as the rectified linear unit (ReLu), to allow for learning non-linear relationships between features. Autoencoders use layers of these neurons to learn an encoder function that maps the high-dimensional data onto a low-dimensional space and a decoder function that takes this low dimensional representation and seeks to recover the initial high-dimensional space. These models are trained by minimizing the reconstruction loss, which is some measure of distance between the original input data and the output reconstructed from the low-dimensional space.

Variational autoencoders (VAEs) are a type of autoencoder that learns to represent the input data using a continuous distribution (typically a multivariate gaussian). In a VAE, the encoder learns a function that maps to one set of neurons representing the means of multivariate gaussian and another set of neurons representing the standard deviations. Together, the mean and standard deviation vectors function like a

multivariate gaussian distribution, which can be sampled from by the decoder. In a VAE, the decoder learns a function that can reconstruct the input based on these samples from the encoder distribution. A VAE is optimized by considering the reconstruction loss and another loss term that uses the KL divergence to penalize the mean and standard deviation vectors if they differ too much from a prior distribution (typically the unit gaussian with means of 0 and standard deviations of 1). VAEs lead to a more continuous latent space than standard autoencoders because this loss function encourages encodings that are as close as possible while still distinguishing relevant features. These methods are widely used in biological analysis for dimensionality reduction and sample clustering and Chapter 2 of this thesis discusses a novel method using sparsely connected autoencoders to produce a biologically-informed latent space.

1.4.2 **Gene Set Scoring Methods**

In a typical omics experiment, large amounts of biological data are collected from disease and control models. It is then common to use statistical tests to identify differential features between conditions. In addition to analyzing differential genes, there is also great interest is considering sets of genes with related functions. The most common gene set analysis is Gene Ontology (GO) enrichment analysis (Harris et al., 2008), whereby differential genes are identified in an omics experiment and a hypergeometric test is used to determine if a biologically-related set of genes (e.g. genes localized at synapses) are statistically enriched for the set differential of genes. There are many other methods, such as Gene Set Enrichment Analysis (Subramanian et al., 2005), for identifying differential gene set enrichment between conditions.

In addition to these comparison methods, there are also gene set scoring algorithms. These approaches convert gene level data into gene set scores for each sample, which can then be used directly to identify differential gene sets or for sample clustering. This thesis introduces a new method for gene set scoring and compares it to four existing algorithms. The Z-score method calculates the z-score for each gene across samples and then considers the gene set score for a given sample as the mean of the z-scores of all genes in a given gene set (Lee et al., 2008). PLAGE also uses z-score normalization and then performs Singular Value Decomposition (SVD) for each gene set separately; here the gene set scores are the first right singular vector obtained from the SVD (Tomfohr et al., 2005). GSVA (Ramasamy et al., 2014) and ssGSEA (Barbie et al., 2009) are both rank-based enrichment algorithms that use distinct methods to perform a modified KS Test to compare the ranks of genes in a gene set to all other genes in a sample.

The new method presented in this work uses sparse autoencoders for gene set scoring, by treating latent variables in an autoencoder as gene sets that only receive inputs from associated genes. This method significantly improves upon the existing methods because the most commonly used scoring algorithms, GSVA and ssGSEA, are both rank-based methods that struggle with 0-inflated scRNA-seq data, and this is less of an issue for the novel auto-encoder based approach.

1.4.3 **Minimum Cost Flow Algorithms**

Minimum cost flow algorithms are used to identify optimal paths through directed networks where each edge has an associated cost and flow capacity. These methods are frequently used in logistics problems to identify optimal paths for transportation.

These algorithms have also been used to analyze biological networks. For example, ResponseNet utilizes the minimum cost flow method to identify highly relevant subsets of proteins that are related to hits from two distinct biological assays (Yeger-Lotem et al., 2009). For ResponseNet, a directed network connects hits from one assay (e.g. genetic screen) and to hits from another biological assay (e.g. RNA-seq). These hits are connected through some known biological network, like a protein-protein interaction (PPI) network, where edges represent physical interactions between proteins. For ResponseNet, every edge in the network has an associated capacity and cost. For any gene that was a hit in a biological assay, the capacity and cost are related to the strength of the hit, where stronger hits have higher capacities and lower costs. For the other proteins, the capacity and cost are related to the biological network itself; in ResponseNet, the edge costs in the PPI are based on confidence, where experimentally determined protein interactions have lower costs than computationally derived edges. ResponseNet then solves the minimum cost flow optimization problem through this network to reveal a small subset of highly relevant proteins that connect the hits from the two biological assays.

SamNet is another minimum cost flow algorithm that expands upon the ResponseNet framework to consider multiple conditions with one optimization problem (Gosline et al., 2012). In this method, there are condition-specific source nodes at the top of the network and condition-specific sink nodes at the bottom. Still, all conditions share the same intermediate nodes. Gosline *et al.* (2012) found that this framework revealed more condition-specific effects than running ResponseNet individually for each case. This thesis details an updated and significantly improved version of SamNet. In this new method, I penalize flow through hub nodes and utilize RNA expression data to promote more realistic output networks. Chapter 4 highlights the utility of this approach in studying drug response in GBM cell line models.

1.4.4 **Prize-Collecting Steiner Forest Algorithm**

OmicsIntegrator is a machine learning package used to integrate multiple types of biological data (Tuncbag et al., 2016). This method is based on the Prize-Collecting Steiner Forest (PCSF) algorithm and uses a known biological network like ResponseNet and SamNet. In this method, every edge has an associated cost, and a small percentage of nodes have prizes. The goal of the optimization problem is to find the connected subnetwork that maximizes the prize values while minimizing the edge costs.

For OmicsIntegrator, nodes are assigned prizes if they are hits from an omics experiment and the value of the prize is related to the strength of the hit. The edges are connections in the PPI network and the costs are related to the confidence of the edge.

The PCSF algorithm is then used to find the optimal network which retains high-prize nodes and connects them through low-cost edges. This results in a final subnetwork of the highly relevant hits from omics assays and other important proteins. In this thesis, I detail an example of OmicsIntegrator being used to better understand the drivers of proteomic subtypes in medulloblastoma.

1.4.5 **Motif Enrichment Algorithms**

ATAC-Seq experiments help identify regions of DNA with accessible chromatin. It is common to then test whether these open regions (or differentially open regions) are associated with specific transcription factors (TFs) binding sites. This is a computationally challenging problem that involves searching these oligonucleotide sequences for transcription factor binding motifs and comparing the frequency of those motifs to other regions of the genome.

The most commonly used package for motif enrichment is HOMER (Benner et al., 2017; Heinz et al., 2010), which searches for potential transcription factors using supervised and de novo approaches. In the supervised mode, the regions of interest are matched against a database of known motifs, and it is determined whether these regions contain more matches than the rest of the genome. Additionally, HOMER can discover new motifs by finding sequential patterns of nucleotides that are enriched in these regions of interest compared to background areas. These de novo motifs are then matched against a database of known motifs to help with annotation. In Chapter 3 of this thesis, I integrated ATAC-Seq and ChIP-Seq data to identify relevant genomic regions and used the HOMER package to find potentially relevant transcription factors for HD models.

1.5 Additional Computational Approaches for Similar Datasets

Section 1.4 provided background on the primary machine learning approaches discussed in this thesis, but there are many other methods that could have been used to analyze and integrate the data types collected for these projects. In this section, I discuss some alternative approaches that could have been employed for this thesis.

1.5.1 **Bayesian Approaches for Biologically Informed Networks**

In this thesis, I discuss methods for incorporating prior biological knowledge into autoencoders by using sparsely connected networks. Other groups have performed similar tasks using a Bayesian framework (Wang et al., 2018). In general, Bayesian Inference is a probabilistic way to learn about a dataset using a likelihood function, which measures how well the data matches a given model, and a prior distribution, which incorporates expectations about the dataset. Wang *et al.* (2018) used a Bayesian approach to model the interactions between genetic, epigenetic, and transcriptomic data and then used Markov Chain Monte Carlo (MCMC) sampling to infer the most important connections. The primary advantage of this approach is that it allows for encoding

complicated interdependences between the biological data types that are not easy to model in a neural network. On the other hand, the sampling methods needed to infer results are very challenging to employ and have incredibly long run times. The autoencoder based gene set scoring method discussed in Chapter 2 is very straightforward and thus a Bayesian approach would likely not be worth the computational challenges, but this Bayesian method would be very useful for future methods that try to calculate gene set scores based on multiple types of omics data at once.

1.5.2 **ATAC-Seq Footprinting for TF Inference**

For the HD project in Chapter 3, we collected epigenomic and transcriptomic data from iPSC-derived neurons from HD patients and controls. One of the primary analyses was identifying differentially accessible chromatin regions and then performing TF motif enrichment to identify disease relevant transcription factors. In 2019, Li *et al.* developed a method to help improve TF identification using a footprinting approach (Li et al., 2019). Their HINT-ATAC algorithm is based on the principle that the ATAC-Seq transposase makes specific cleavage patterns around bound transcription factors. They incorporate this knowledge into ATAC-Seq data processing to identify chromatin regions likely bound by TFs. This algorithm could be used on the HD ATAC-Seq data to filter the differential chromatin regions down to areas likely bound by TFs. Performing motif enrichment on only these genomic regions would provide a more confident assessment of disease-relevant TFs.

1.5.3 **Additional Network-Based Models for Omics Integration**

For the glioblastoma and medulloblastoma projects, we used network-based algorithms to integrate data from multiple biological assays to get a more comprehensive understanding of the disease. The GBM project utilized network flow methods, while the integration of MB datasets was based on the PCSF algorithm. Many other groups have also used network-based algorithms to analyze multiple biological data types at once.

For example, the HotNet2 method uses network diffusion principles to find relevant subnetworks for a given dataset (Leiserson et al., 2015). This method is based on how heat would diffuse through a network. It is implemented using a protein-protein interaction network where proteins are assigned an initial heat score and each node can receive and pass heat to its neighbors. The method then allows heat to diffuse between proteins until equilibrium is reached and uses statistical tests to identify subnetworks with significantly high amounts of heat. In the glioblastoma project, our goal was to identify potential combination therapies for the proneural subtype of GBM. We could have used the HotNet2 framework to study this question by assigning initial heat values to proteins that were hits from the shRNA viability screen or proteomics assay and analyzing significant subnetworks to see if any led to hypotheses about potential therapeutic targets.

Additionally, Hristov *et al.* developed the network-based algorithm, uKIN, which would be particularly useful for our MB data (Hristov et al., 2020). uKIN is similar HotNet2 and uses random walks through a PPI network to identify highly relevant subnetworks. The primary difference, however, is that uKIN allows for the incorporation of both prior knowledge and new information into the algorithm. Previously collected data is used to create the transition probabilities between nodes and hits from the new dataset are used as the starting points for the random walks. This framework could be very helpful for understanding the relevant properties of proteomic subtypes in MB. For example, when studying the new SHH-activated MB subtypes (SHHa and SHHb), prior genetic, epigenetic, and transcriptomic data collected about SHH-activated MBs could be used to create transition probabilities for the network. Then, to learn more about the SHHa proteomic subtype, uKIN can be run using proteomic hits that are upregulated in SHHa. This same procedure could be repeated to make inferences about the SHHb subtype.

1.6 Overview of Thesis

In this thesis, I present multiple examples of machine learning methods being used to improve our understanding of neurological diseases. I only discuss a small subset of neurological conditions, but the computational algorithms and analytical principles are applicable to similar datasets from other complicated disorders.

Chapter 2 focuses on a novel machine learning method that I developed for gene set scoring. The main algorithms for gene set scoring are ssGSEA and GSVA, but both of these methods are rank-based and struggle with zero-inflated scRNA-seq data. I present a new method to address this issue, which uses sparse autoencoders to convert gene-level data into gene set scores, while simultaneously creating a low-dimensional representation of the data. This allows for successful gene set projection that can be utilized for the many scRNA-seq datasets being produced to study neurological conditions (Al-Dalahmah et al., 2020; Hovestadt et al., 2019; Neftel et al., 2019; Riemondy et al., 2022; Vladoiu et al., 2019). Additionally, these methods are currently being employed to analyze omics data from Amyotrophic Lateral Sclerosis (ALS) models, but this specific work is not discussed in the thesis.

Chapter 3 details multi-omic analysis of iPSC-derived neuron models of HD. We performed ATAC-Seq, ChIP-Seq, bulk RNA-seq, and snRNA-seq on these models and found a novel population of cycling cells in the HD samples. My primary contributions to this work were integrating the epigenomic and transcriptomic data to better understand the molecular features of the cycling cells.

Chapter 4 describes analysis of drug response in glioblastoma cell lines. I analyzed proteomic and phosphoproteomic data to characterize subtype specific drug response. In this chapter, I also present significant improvements to the SamNet multi-commodity flow algorithm. This new method uses a penalty to reduce flow through hub nodes, and allows for the incorporation of RNA expression data, both of which lead to more realistic output networks.

Chapters 5 and 6 detail machine learning approaches for studying medulloblastoma. Chapter 5 focuses on proteomic and phosphoproteomic analysis of MB. We found that clustering protein data from MB tumors leads to novel subtypes not observed in methylation or RNA data. We also used the OmicsIntegrator method to better understand the molecular drivers of these subtypes and observed that the new G3 subtypes (named G3a and G3b) have significantly different outcomes.

Chapter 6 focuses on single-cell RNA-sequencing of medulloblastoma tumors with extensive nodularity. Clustering and pseudotime algorithms helped reveal that some tumors contain cells mimicking every stage of granule neuron development. With that observation, I developed a novel computational method to connect these results to published examples of heterogeneity in SHH-activated MB. I found that considering this developmental perspective helped yield novel insights about tumor subtypes, mutations, metabolism, and histology.

Some chapters detail specific improvements in machine learning methods applicable to a wide variety of disease data, while other sections emphasize novel biological insights found using advanced computational approaches. Together, this thesis highlights the value of machine learning for studying neurological diseases.

1.7 References

Al-Dalahmah, O., Sosunov, A.A., Shaik, A., Ofori, K., Liu, Y., Vonsattel, J.P., Adorjan, I., Menon, V., and Goldman, J.E. (2020). Single-nucleus RNA-seq identifies Huntington disease astrocyte states. Acta Neuropathol. Commun.

Archer, T.C., Mahoney, E.L., and Pomeroy, S.L. (2017). Medulloblastoma: Molecular Classification-Based Personal Therapeutics. Neurotherapeutics.

Archer, T.C., Ehrenberger, T., Mundt, F., Gold, M.P., Krug, K., Mah, C.K., Mahoney, E.L., Daniel, C.J., LeNail, A., Ramamoorthy, D., et al. (2018). Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups. Cancer Cell.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell.

Benner, C., Heinz, S., and Glass, C.K. (2017). HOMER - Software for motif discovery and next generation sequencing analysis.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol.

Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based

classification of central nervous system tumours. Nature.
Caprioli, R.M., Farmer, T.B., and Gile, J. (1997). Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. Anal. Chem.
Casciati, A., Tanori, M., Manczak, R., Saada, S., Tanno, B., Giardullo, P., Porcù, E., Rampazzo, E., Persano, L., Viola, G., et al. (2020). Human medulloblastoma cell lines: Investigating on cancer stem cell-like phenotype. Cancers (Basel).
Cavalli, F.M.G., Remke, M., Rampasek, L., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A.S., et al. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell.
Chen, G., Ning, B., and Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. Front. Genet.
Chen, X., Wei, S., Ji, Y., Guo, X., and Yang, F. (2015). Quantitative proteomics using SILAC: Principles, applications, and developments. Proteomics.
Cheng, Y., Liao, S., Xu, G., Hu, J., Guo, D., Du, F., Contreras, A., Cai, K.Q., Peri, S., Wang, Y., et al. (2020). NeuroD1 Dictates Tumor Cell Differentiation in Medulloblastoma. Cell Rep.
Cho, Y.J., Tsherniak, A., Tamayo, P., Santagata, S., Ligon, A., Greulich, H., Berhoukim, R., Amani, V., Goumnerova, L., Eberhart, C.G., et al. (2011). Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. J. Clin. Oncol.
Coons, A.H., Creech, H.J., and Jones, R.N. (1941). Immunological Properties of an Antibody Containing a Fluorescent Group. Proc. Soc. Exp. Biol. Med.
Djoussé, L., Knowlton, B., Hayden, M.R., Almqvist, E.W., Brinkman, R.R., Ross, C.A., Margolis, R.L., Rosenblatt, A., Durr, A., Dode, C., et al. (2004). Evidence for a modifier of onset age in Huntington disease linked to the HD gene in 4p16. Neurogenetics.
Dubuc, A.M., MacK, S., Unterberger, A., Northcott, P.A., and Taylor, M.D. (2012). The epigenetics of brain tumors. Methods Mol. Biol.
Feil, S., Valtcheva, N., and Feil, R. (2009). Inducible cre mice. Methods Mol. Biol.
Fogarty, M.P., Kessler, J.D., and Wechsler-Reya, R.J. (2005). Morphing into cancer: The role of developmental signaling pathways in brain tumor formation. J. Neurobiol.
Forget, A., Martignetti, L., Puget, S., Calzone, L., Brabetz, S., Picard, D., Montagud, A., Liva, S., Sta, A., Dingli, F., et al. (2018). Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling. Cancer Cell.
Gajjar, A., Chintagumpala, M., Ashley, D., Kellie, S., Kun, L.E., Merchant, T.E., Woo, S., Wheeler, G., Ahern, V., Krasin, M.J., et al. (2006). Risk-adapted craniospinal radiotherapy followed by high-dose chemotherapy and stem-cell rescue in children with newly diagnosed medulloblastoma (St Jude Medulloblastoma-96): long-term results from a prospective, multicentre trial. Lancet Oncol.
Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. Cell.
Gosline, S.J.C., Spencer, S.J., Ursu, O., and Fraenkel, E. (2012). SAMNet: A network-based approach to integrate multi-dimensional high throughput datasets. Integr. Biol. (United Kingdom).
Greally, J.M. (2018). A user's guide to the ambiguous word "epigenetics." Nat. Rev. Mol.

Cell Biol.
Han, S., Liu, Y., Cai, S.J., Qian, M., Ding, J., Larion, M., Gilbert, M.R., and Yang, C. (2020). IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. Br. J. Cancer.
Harris, M.A., Deegan, J.I., Ireland, A., Lomax, J., Ashburner, M., Tweedie, S., Carbon, S., Lewis, S., Mungall, C., Day-Richter, J., et al. (2008). The Gene Ontology project in 2008. Nucleic Acids Res.
Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. Genome Biol.
Hayatsu, H. (2008). Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis - A personal account. Proc. Japan Acad. Ser. B Phys. Biol. Sci.
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell.
Holland, E.C. (2000). Glioblastoma multiforme: The terminator. Proc. Natl. Acad. Sci. U. S. A.
Hollerman, J.R., Tremblay, L., and Schultz, W. (2000). Involvement of basal ganglia and orbitofrontal cortex in goal-directed behavior. In Progress in Brain Research, p.
Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.K.L., Baumgartner, A., DeWitt, J.C., Groves, A., Mayr, L., Weisman, H.R., et al. (2019). Resolving medulloblastoma cellular architecture by single-cell genomics. Nature.
Hristov, B.H., Chazelle, B., and Singh, M. (2020). uKIN Combines New and Prior Information with Guided Network Propagation to Accurately Identify Disease Genes. Cell Syst.
Huszthy, P.C., Daphu, I., Niclou, S.P., Stieber, D., Nigro, J.M., Sakariassen, P.O., Miletic, H., Thorsen, F., and Bjerkvig, R. (2012). In vivo models of primary brain tumors: Pitfalls and perspectives. Neuro. Oncol.
Jakovcevski, M., and Akbarian, S. (2013). Epigenetic mechanisms in neurodevelopmental and neurodegenerative disease. Nat. Med.
Jiang, R., Sun, T., Song, D., and Li, J.J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol.
Karpievitch, Y. V., Polpitiya, A.D., Anderson, G.A., Smith, R.D., and Dabney, A.R. (2010). Liquid chromatography mass spectrometry-based proteomics: Biological and technological aspects. Ann. Appl. Stat.
Kijima, N., and Kanemura, Y. (2017). Mouse Models of Glioblastoma. In Glioblastoma, p.
King, H.C., and Sinha, A.A. (2001). Gene expression profile analysis by DNA microarrays: Promise and pitfalls. J. Am. Med. Assoc.
Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. PLoS Comput. Biol.
Leiserson, M.D.M., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J. V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet.

Li, P., Du, F., Yuelling, L.W., Lin, T., Muradimova, R.E., Tricarico, R., Wang, J., Enikolopov, G., Bellacosa, A., Wechsler-Reya, R.J., et al. (2013). A population of Nestin-expressing progenitors in the cerebellum exhibits increased tumorigenicity. Nat. Neurosci.
Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., and Costa, I.G. (2019). Identification of transcription factor binding sites using ATAC-seq. Genome Biol.
Lim, R.G., Salazar, L.L., Wilton, D.K., King, A.R., Stocksdale, J.T., Sharifabad, D., Lau, A.L., Stevens, B., Reidling, J.C., Winokur, S.T., et al. (2017). Developmental alterations in Huntington's disease neural cells and pharmacological rescue in cells and mice. Nat. Neurosci.
Lin, J.R., Fallahi-Sichani, M., Chen, J.Y., and Sorger, P.K. (2016). Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. Curr. Protoc. Chem. Biol.
Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., and Ellison, D.W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol.
MacDonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., Groot, N., et al. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell.
Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet.
McColgan, P., and Tabrizi, S.J. (2018). Huntington's disease: a clinical review. Eur. J. Neurol.
Mehrmohamadi, M., Sepehri, M.H., Nazer, N., and Norouzi, M.R. (2021). A Comparative Overview of Epigenomic Profiling Methods. Front. Cell Dev. Biol.
Milne, T.A., Zhao, K., and Hess, J.L. (2009). Chromatin immunoprecipitation (ChIP) for analysis of histone modifications and chromatin-associated proteins. Methods Mol. Biol.
Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell.
Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V.A., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. Nature.
Oda, Y., Huang, K., Cross, F.R., Cowburn, D., and Chait, B.T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. Proc. Natl. Acad. Sci. U. S. A.
Oronsky, B., Reid, T.R., Oronsky, A., Sandhu, N., and Knox, S.J. (2021). A Review of Newly Diagnosed Glioblastoma. Front. Oncol.
Patel, V., Chisholm, D., Dua, T., Laxminarayan, R., and Medina-Mora, M.E. (2016). Disease Control Priorities, Third Edition (Volume 4): Mental, Neurological, and Substance Use Disorders.
Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq

sequencers. BMC Genomics.
Raghavachari, N., and Garcia-Reyero, N. (2018). Overview of gene expression analysis: Transcriptomics. In Methods in Molecular Biology, p.
Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Coin, L., De Silva, R., Cookson, M.R., et al. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci.
Ramaswamy, S., McBride, J.L., and Kordower, J.H. (2007). Animal models of Huntington's disease. ILAR J.
Ramos-Vara, J.A., and Miller, M.A. (2014). When Tissue Antigens and Antibodies Get Along: Revisiting the Technical Aspects of Immunohistochemistry-The Red, Brown, and Blue Technique. Vet. Pathol.
Reis-Filho, J.S. (2009). Next-generation sequencing. Breast Cancer Res.
Riemondy, K.A., Venkataraman, S., Willard, N., Nellan, A., Sanford, B., Griesinger, A.M., Amani, V., Mitra, S., Hankinson, T.C., Handler, M.H., et al. (2022). Neoplastic and immune single-cell transcriptomics define subgroup-specific intra-tumoral heterogeneity of childhood medulloblastoma. Neuro. Oncol.
Roberts, L.D., Souza, A.L., Gerszten, R.E., and Clish, C.B. (2012). Targeted metabolomics. Curr. Protoc. Mol. Biol. *1*.
Ross, C.A., and Tabrizi, S.J. (2011). Huntington's disease: From molecular pathogenesis to clinical treatment. Lancet Neurol.
Roussel, M.F., and Stripay, J.L. (2020). Modeling pediatric medulloblastoma. Brain Pathol.
Shenoy, S.A., Zheng, S., Liu, W., Dai, Y., Liu, Y., Hou, Z., Mori, S., Tang, Y., Cheng, J., Duan, W., et al. (2022). A novel and accurate full-length HTT mouse model for Huntington's disease. Elife.
Stoeckli, M., Chaurand, P., Hallahan, D.E., and Caprioli, R.M. (2001). Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. Nat. Med.
Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J.B., Belanger, K., Brandes, A.A., Marosi, C., Bogdahn, U., et al. (2005). Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. N. Engl. J. Med.
Stupp, R., Tonn, J.C., Brada, M., and Pentheroudakis, G. (2010). High-grade malignant glioma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann. Oncol.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.
Taylor, M.D., Northcott, P.A., Korshunov, A., Remke, M., Cho, Y.J., Clifford, S.C., Eberhart, C.G., Parsons, D.W., Rutkowski, S., Gajjar, A., et al. (2012). Molecular subgroups of medulloblastoma: The current consensus. Acta Neuropathol.
Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem.
Tomfohr, J., Lu, J., and Kepler, T.B. (2005). Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics *6*.

Trettel, F., Rigamonti, D., Hilditch-Maguire, P., Wheeler, V.C., Sharp, a H., Persichetti, F., Cattaneo, E., and MacDonald, M.E. (2000). Dominant phenotypes produced by the HD mutation in STHdh(Q111) striatal cells. Hum. Mol. Genet. *9*, 2799–2809.
Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput. Biol. *12*.
Valcárcel-Ocete, L., Alkorta-Aranburu, G., Iriondo, M., Fullaondo, A., García-Barcina, M., Fernández-García, J.M., Lezcano-García, E., Losada-Domingo, J., Ruiz-Ojeda, J., De Arcaya, A.Á., et al. (2015). Exploring genetic factors involved in huntington disease age of onset: E2F2 as a new potential modifier gene. PLoS One.
Verhaak, R.G.W., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell.
Vladoiu, M.C., El-Hamamy, I., Donovan, L.K., Farooq, H., Holgado, B.L., Sundaravadanam, Y., Ramaswamy, V., Hendrikse, L.D., Kumar, S., Mack, S.C., et al. (2019). Childhood cerebellar tumours mirror conserved fetal transcriptional programs. Nature.
Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol.
Wakimoto, H., Mohapatra, G., Kanai, R., Curry, W.T., Yip, S., Nitta, M., Patel, A.P., Barnard, Z.R., Stemmer-Rachamimov, A.O., Louis, D.N., et al. (2012). Maintenance of primary tumor phenotype and genotype in glioblastoma stem cells. Neuro. Oncol.
Walker, D.A., Harper, P.S., Wells, C.E.C., Tyler, A., Davies, K., and Newcombe, R.G. (1981). Huntington's chorea in South Wales. A genetic and epidemiological study. Clin. Genet.
Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science (80-.).
Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A.C., Lyu, S., Li, P., Li, Y., et al. (2017). Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. Cancer Cell.
Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet.
Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat. Genet.
Zhang, S., Moy, W., Zhang, H., Leites, C., McGowan, H., Shi, J., Sanders, A.R., Pang, Z.P., Gejman, P. V., and Duan, J. (2018). Open chromatin dynamics reveals stage-specific transcriptional networks in hiPSC-based neurodevelopmental model. Stem Cell Res.

Chapter 2 - Shallow Sparsely-Connected Autoencoders for Gene Set Projection

Authors: Maxwell P. Gold, Alexander LeNail, and Ernest Fraenkel

This work was published in Pac Symp Biocomput 2019; 24:374-385. I was the primary contributor to this project, where I developed the idea, performed the analyses, and wrote the manuscript. This paper was also the product of many discussions with Alex LeNail and Ernest Fraenkel.

2.1 Abstract

When analyzing biological data, it can be helpful to consider gene sets, or predefined groups of biologically related genes. Methods exist for identifying gene sets that are differential between conditions, but large public datasets from consortium projects and single-cell RNA-sequencing have opened the door for gene set analysis using more sophisticated machine learning techniques, such as autoencoders and variational autoencoders. We present shallow sparsely-connected autoencoders (SSCAs) and variational autoencoders (SSCVAs) as tools for projecting gene-level data onto gene sets. We tested these approaches on single-cell RNA-sequencing data from blood cells and on RNA-sequencing data from breast cancer patients. Both SSCA and SSCVA can recover known biological features from these datasets and the SSCVA method often outperforms SSCA (and six existing gene set scoring algorithms) on classification and prediction tasks.

2.2 Introduction

RNA-sequencing (RNA-seq) experiments can quantify the RNA expression levels for ~20,000 human genes and this data may reveal differences between experimental conditions, such as cancerous tissue vs. healthy tissue. Typically, RNA-seq analysis begins with identifying genes with differential RNA levels across conditions and determining if such genes are over-represented in any predefined gene sets (i.e. groups of biologically related genes). This standard approach can be useful but is also quite simplistic; it ignores relationships among the genes and assumes all genes in a gene set are equally important to the group.

Consortium projects (such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013)) and the development of single-cell RNA-sequencing (scRNA-seq) (Tang et al., 2009) have yielded large public datasets for RNA-seq analysis; this has permitted the use of more complex machine learning techniques, such as autoencoders (Liou et al., 2008) and variational autoencoders (VAEs) (Kingma and Welling, 2013), for analyzing those data. These methods can project the high-dimensional gene space onto a lower-dimensional latent space, which may help with visualization, denoising, and/or interpretation (Wang et al., 2017; Xie et al., 2017; žurauskiene and Yau, 2016). Additionally, some neural networks and autoencoders have even been designed to

incorporate biological information by using sparsely-connected nodes that only receive inputs from biologically-related genes (Kang et al., 2017; Lin et al., 2017).

Many of these neural-network-based and autoencoder-based approaches have focused primarily on increasing accuracy, but recently, groups have used these methods for data interpretation. For example, Way and Greene (2018) used a VAE on TCGA data, wherein they projected RNA-seq data onto a reduced latent space, identified nodes that differentiate cancer subtypes, and used the learned model parameters to search for biological significance (Way and Greene, 2018). Chen *et al.* (2018) detailed a similar approach, whereby they used sparse connections to project genes onto gene sets and then had a fully-connected layer between the gene set nodes and latent nodes (Chen et al., 2018); a gene set was considered meaningful if it had a high input weight into a relevant latent superset node.

Here we describe a different approach for using autoencoders for gene set analysis. We present shallow sparsely-connected autoencoders (SSCAs) (Figure 1A) and shallow sparsely-connected variational autoencoders (SSCVAs) (Figure 1B) as tools for projecting gene-level data onto gene sets, wherein those gene set scores can be used for downstream analysis. These methods use a single-layer autoencoder or VAE with sparse connections (representing known biological relationships) in order to attain a value for each gene set. Chen *et al.* (2018) mentioned the SSCA model (Figure 1A) but did not thoroughly explore its utility for gene set projection (Chen et al., 2018). There are many statistical methods for gene set scoring (see Section 2.5), but these techniques often rely on assumptions that do not reflect the underlying biology (e.g. all genes are equally important to a gene set). That being said, the machine-learning approaches presented in this work allow for learning a specific nonlinear mapping function for each gene set; thus, each gene within a gene set can be weighted differently and a single gene can have distinct weights across gene sets.

Ideally, the gene set scores should be able to retain high-level information from the gene-level data and provide new insights regarding the relevant gene sets. To test whether the SSCA and SSCVA algorithms can extract such gene set scores, we ran both algorithms on scRNA-seq data from human blood cells and on RNA-seq data from patients with breast cancer; we used classification and prediction tasks to compare these new methods to six existing gene set scoring algorithms and assessed the biological interpretability of SSCA and SSCVA by performing differential analysis using the computed scores.

2.3 Methods

2.3.1 **Model Summary**

This work explores shallow sparsely-connected autoencoders (SSCAs) (Figure 1A) and shallow sparsely-connected variational autoencoders (SSCVAs) (Figure 1B). Autoencoders learn an encoder function that projects input data onto a lower dimensional space and a decoder function that aims to recover the input data from the

low-dimensional projections. The model is trained by minimizing the reconstruction loss (i.e. some measure of distance between the reconstructed output and the original input).

Variational autoencoders (VAEs), however, learn a continuous distribution (typically a multivariate gaussian) to represent the input data. The encoder learns projections onto both a mean vector and a standard deviation vector (which are used to represent a multivariate Gaussian) and the decoder takes samples from the encoded distribution and learns a function to project these samples onto the original space. For VAEs, the model is trained by minimizing both the aforementioned reconstruction loss and the KL divergence between the learned multivariate Gaussian and a chosen prior distribution (typically the unit Gaussian).

The shallow sparsely-connected autoencoders and VAEs discussed in this work are based on said algorithms, but with two notable restrictions: the encoding/decoding functions are only one layer deep and these layers are sparse (not fully-connected like standard autoencoders), with connections based on known biological relationships. For SSCA, each encoded node represents a gene set and only receives inputs from gene nodes included in the set. For SSCVA, each gene set is represented by a mean vector node and a standard deviation vector node, both of which only receive inputs from the relevant gene nodes. When analyzing the trained SSCVA models, we considered the score for each gene set to be the value of the mean vector node.

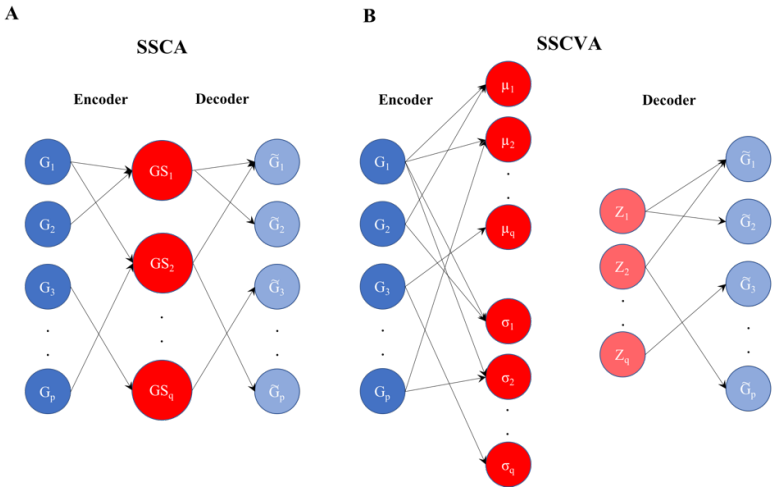


Fig 1. **Diagram for Shallow Sparsely-Connected Autoencoder (SSCA) and Variational Autoencoder (SSCVA)**. A) SSCA model. B) SSCVA model. For SSCA, the input genes (G_1 - G_p) are connected to gene set nodes (GS_1 - GS_q). Each gene set node only receives inputs from the genes within the gene set. Light blue denotes the reconstructed gene values ($\widetilde{G}_1$ - $\widetilde{G}_p$). SSCVA follows the same model, except there is μ node and σ node for each gene set. The z values are collected using the following scheme: $\bar{z} = \bar{\mu} + (\bar{\sigma} * \bar{\epsilon})$ where $\bar{\epsilon} \sim U(0,1)$. Those values are then used to project onto $\widetilde{G}_1$- $\widetilde{G}_p$.

2.3.2 **Model Coding**

We implemented the models in python using the TensorFlow package (Abadi et al., 2015) (version 1.8.0) and select functions from Keras (version 2.1.6) (Chollet, 2015). We employed hyperbolic tangent (tanh) activation for the encoder functions and sigmoid

activation for the decoder functions. For the encoders, we used batch normalization (which scales values to zero mean and unit variance) after linear activation and before tanh activation. Additionally, we trained both models using Adam optimization (Kingma and Ba, 2014). The SSCVA code is largely based on public code from Way and Greene (2018) (Way and Greene, 2018) and the sampling procedure follows the scheme where $\bar{z} = \bar{\mu} + (\bar{\sigma} * \bar{\epsilon})$ and $\bar{\epsilon} \sim U(0,1)$ (Figure 1B).

2.3.3 **Data and Gene Set Summary**

We used two publicly available data sets for this analysis: a single-cell RNA-seq dataset of 1078 blood cells (dendritic cells and monocytes) (Villani et al., 2017) and an RNA-seq dataset from patients with breast cancer from The Cancer Genome Atlas (TCGA) (Grossman et al., 2016; Weinstein et al., 2013). The scRNA-seq data matrix consists of preprocessed log TPM values for genes for 1078 high-quality cells (Villani et al., 2017). For training, the data was scaled to a range of 0-1 using min-max scaling. The breast cancer dataset includes 1093 patients with RNA-seq data ($\log_2$(FPKM + 1) transformed RSEM values) and matching clinical data (Grossman et al., 2016; Weinstein et al., 2013). A small number of patients have multiple RNA-seq runs and for these cases, the mean RSEM value for each gene across runs was assigned to the patient. After this step, the breast cancer data was processed in the same manner as the scRNA-seq dataset.

The gene sets used to create the sparse layers are from the Molecular Signatures Database (Subramanian et al., 2005). We used the transcription factor targets collection (C3.TFT) for scRNA-seq analysis and the cancer signatures collection (C6) for the breast cancer survival analysis. We then filtered the collections to include only gene sets with more than 15 genes and less than 500 genes, reducing the C3.TFT collection from 615 to 550 gene sets and the C6 collection from 189 to 187 gene sets. Using only the remaining genes, the input matrices were 1078 cells x 10992 genes for the scRNA-seq data and 1093 patients x 10650 genes for the breast cancer analysis.

2.3.4 **Hyperparameter Selection**

We considered the following variables for a parameter sweep: learning rate (0.0075, 0.001, 0.002), epochs (50, 100, 150), and L2 regularization (0, 0.05, 0.1). Additionally, we tested warmup (κ) (0.05, 1) for the SSCVA model, where κ controls how quickly the KL loss contributes to the total loss being minimized in the VAE (Sønderby et al., 2016). We kept the optimizer (Adam) and batch size (50) consistent for all trials. We used 90% of the samples for training and 10% for validation and chose the hyperparameters corresponding to the model with the lowest validation loss. For both the blood cell and the breast cancer data, the validation loss for SSCA was lowest for a learning rate of 0.002, 150 epochs, and no L2 regularization. For SSCVA in both analyses, the validation loss was minimized by a learning rate of 0.002, 150 epochs, L2 regularization of 0.1, and κ of 0.05. Hu and Greene (Hu and Greene, 2018) recently raised concerns about model comparison analysis when some models are heavily reliant on

hyperparameter tuning. Thus, in this work, the SSCA and SSCVA models chosen for comparison are the ones that minimize loss, without any regard for task performance.

2.3.5 **Other Projection Methods**

In addition to SSCA and SSCVA, we assessed the performance of six other methods for projecting gene data onto gene sets: Average Z-score (Z-Score) (Lee et al., 2008), Pathway Level Analysis of Gene Expression (PLAGE) (Tomfohr et al., 2005), Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013), single-sample Gene Set Enrichment Analysis (ssGSEA) (Barbie et al., 2009), FastProject (FP) (DeTomaso and Yosef, 2016), and simple averaging (Average). The Z-Score method normalizes each gene by z-score across samples and considers the gene set score to be the mean normalized value of all genes in a set. PLAGE uses the same z-score normalization and then performs singular value decomposition (SVD) for each gene set; the gene set scores are the first right singular vector obtained from the SVD. GSVA and ssGSEA are enrichment-based algorithms that utilize distinct methods to rank each gene per sample and then calculate a score for each gene set based on the difference in ranks for genes within the set compared to those outside of it. The averaging method is the arithmetic mean of the RNA-seq values of all the genes within a gene set. Lastly, FastProject is a tool built for scRNA-seq data; the algorithm normalizes the data using z-scores while also accounting for sparsity common in scRNA-seq data and then assigns the gene set score as the mean of the normalized values.

We used the GSVA package in R (version 1.26.0) (Sonja et al., 2014) to calculate GSVA, PLAGE, Z-Score, and ssGSEA scores and ran the FastProject program (DeTomaso and Yosef, 2016) to compute FP scores. Averaging and autoencoder training were performed in python (per the above procedure). To help with training, we used min-max scaled RNA-seq values as inputs for the SSCA and SSCVA methods. The other methods used the normalized RNA-seq values (log TPM for blood cells and RSEM for breast cancer). The only exception is that min-max scaled RNA-seq values were used for the averaging projection for the breast cancer survival prediction as raw values led to convergence issues.

2.3.6 **Dendritic Cell Type Classification**

We used the python package Scikit-learn (version 0.19.1) to train the logistic regression models and gaussian mixture models (GMMs) (Pedregosa et al., 2011). For the GMMs, we set k = 3 and initialized each model five times (using n_init = 5), with the best result being kept. To compare the predicted clusters to known cell types (provided by (Villani et al., 2017)), we calculated normalized mutual information using Scikit-learn (Pedregosa et al., 2011).

2.3.7 **Breast Cancer Prediction**

We analyzed five-year survival on the breast cancer dataset and only kept patients who survived greater than five years (i.e. TCGA “days_to_follow_up” > 1825 days) or who passed away within five years (i.e. TCGA “days_to_death” < 1825 days). This left 352 patients: 253 survivors and 99 who have passed away. For the survival analysis, we used the lifelines package in python (Davidson-Pilon et al., 2018) to train a Cox proportional hazards model (Cox PHM) with a step size of 0.3 to help with convergence. Using a 4:1 train/test split, we trained the Cox PHM to predict days of survival from the gene set scores and compared the predicted days of survival to the true values using the concordance index (CI). To assess the importance of gene sets in predicting days of survival, we ranked the gene sets in ascending order by their p-values using a Wald test. We generated boxplots using the python package Matplotlib (Hunter, 2007) and performed Mann-Whitney U tests using the python package SciPy (Oliphant, 2007).

2.4 Results

We analyzed scRNA-seq data from blood cells (Villani et al., 2017) and RNA-seq data from breast cancer patients (Weinstein et al., 2013) to assess the utility of shallow sparsely-connected autoencoders (SSCA) and variational autoencoders (SSCVA) for projecting gene data onto gene sets. We compared the two autoencoder-based methods to six existing methods for gene set projection (see Methods): GSVA, PLAGE, Z-Score, ssGSEA, FP, and Average.

2.4.1 **Blood scRNA-seq Analysis**

When analyzing scRNA-seq data, it can be difficult to assess the importance of specific transcription factors (TFs) because mRNA levels do not always correlate with protein abundance (Lundberg et al., 2010; Vogel and Marcotte, 2012), and TF activity is affected by other factors in the cell, such as chromatin accessibility. One potential solution is to use transcription factor target gene sets (i.e. genes whose expression is potentially affected by a given TF); if the genes regulated by a TF are differential between conditions, this could suggest that the TF is biologically relevant. Thus, in order to explore the scRNA-seq data set from human blood cells, we performed gene set analysis on 550 transcription factor target gene sets from the Molecular Signatures Database (Subramanian et al., 2005). We performed classification tasks using the gene set encodings to determine whether these projections retain high-level information about the dataset and then analyzed the differential features for biological significance.

2.4.2 **Supervised Classification of Cell Types**

The scRNA-seq data set contains over 1000 individual cells, each of which was assigned one of ten cell types by Villani *et al.* (2017) (Villani et al., 2017) (six dendritic

cell types (DC1-6) and four monocyte cell types (Mono1-4)). We first ran the eight projection methods using the transcription factor target gene sets and then used the resulting gene set scores to train a logistic regression to predict cell type. We used 80% of the samples for training and compared the methods on classification accuracy using test data. This procedure was repeated for multiple distinct cell type combinations (Figure 2).

The cell types used in a given run affected the peak model accuracy, which ranged from 84% (all ten cell types) to 100% (DC1-DC6-Mono1). The model trained using SSCVA gene set scores yielded the highest accuracy in all six trials and was the sole top performer in five of six trials (DC1-DC6-Mono1 being the exception, where many algorithms achieved 100% accuracy). We also compared the performance of SSCVA-based models to logistic regression models trained directly on the gene-level RNA-seq data; models trained on SSCVA gene set scores never outperformed the RNA-seq models (Figure 2) but were always within 2% accuracy. Average-based models often led to the second highest accuracy and SSCA-based models typically resulted in the lowest accuracy among the methods tested. These results suggest that for the blood cell dataset, the SSCVA encodings retain more gene-level information about cell type than the other projection methods.

Cell Types	Scaled RNA-Seq	Raw RNA-Seq	GSVA	PLAGE	Z-Score	ssGSEA	FP	Average	SSCVA	SSCA
DC1 - DC6 - Mono1	1	1	1	1	1	0.99	0.98	1	1	1
DC1 - DC3 - DC5	0.984	0.984	0.885	0.852	0.902	0.803	0.902	0.918	0.967	0.885
DC1 - DC2 - DC3	0.878	0.878	0.851	0.797	0.797	0.811	0.797	0.851	0.865	0.73
All Dendritic Cells (DC1-6)	0.919	0.899	0.799	0.812	0.832	0.758	0.866	0.839	0.899	0.758
All Monocytes (Mono1-4)	0.838	0.838	0.75	0.794	0.779	0.735	0.794	0.794	0.838	0.618
All Cells (DC1-6 & Mono1-4)	0.838	0.852	0.773	0.745	0.727	0.722	0.764	0.787	0.838	0.634

Fig 2. **Logistic Regression Test Data Accuracy**. Each row represents a trial with the specific cell types shown in the first column. Additional columns indicate the data type used for training for cell type prediction (i.e. gene-level RNA-seq data or gene set scores from one of eight algorithms). Values are the classification accuracy of cell types on test data. Yellow emphasizes the highest test accuracy in each row. Scaled RNA-seq (Min-max scaled gene TPM values from (Villani et al., 2017)). Raw RNA-seq (gene TPM values from (Villani et al., 2017)). See Methods for the full names of gene set projection algorithms.

A

Cell Types	Scaled RNA-Seq	Raw RNA-Seq	GSVA	PLAGE	Z-Score	ssGSEA	FP	Average	SSCVA	SSCA
DC1 - DC6 - Mono1	0.96	0.973	0.95	0.936	0.087	0.37	0.289	0.054	0.946	0.987
DC1 - DC3 - DC6	0.985	0.985	0.524	0.906	0.013	0.38	0.174	0.021	0.704	0.652
DC2 - DC6 - Mono3	0.976	0.686	0.482	0.974	0.133	0.418	0.179	0.107	0.66	0.625
DC2 - DC3 - DC4	0.631	0.598	0.389	0.48	0.027	0.069	0.039	0.049	0.474	0.562
DC1 - DC6 - Mono2	0.971	0.986	0.906	0.942	0.027	0.387	0.207	0.05	0.957	1

B

Cell Types	Scaled RNA-Seq	Raw RNA-Seq	GSVA	PLAGE	Z-Score	ssGSEA	FP	Average	SSCVA	SSCA
DC1 - DC6 - Mono1	1	1	0.959	0.959	0.045	0.326	0.145	0	0.959	0.919
DC1 - DC3 - DC6	1	1	0.494	0.77	0.014	0.647	0.093	0.008	0.719	0.204
DC2 - DC6 - Mono3	1	0.783	0.574	1	0.046	0.34	0.218	0.019	0.708	0.377
DC2 - DC3 - DC4	0.707	0.735	0.324	0.561	0.126	0.058	0.026	0.097	0.637	0.1
DC1 - DC6 - Mono2	1	1	0.807	0.957	0.077	0.226	0.224	0.06	1	0.838

Fig 3. **Gaussian Mixture Model Clustering Normalized Mutual Information (NMI) Values**. A) Training Data normalized mutual information (NMI). B) Test Data normalized mutual information (NMI). Each row represents a trial with the specific cell types shown in the first column. Additional columns indicate the data used for training (gene-level RNA-seq data or gene set scores from one of eight algorithms). Values are the normalized mutual information scores between output clusters and known cell types. Yellow emphasizes the highest NMI in each row. Scaled RNA-seq (Min-max scaled gene TPM values from (Villani et al., 2017)). Raw RNA-seq (gene TPM values from (Villani et al., 2017)). See Methods for the full names of gene set projection algorithms.

2.4.3 **Unsupervised Clustering of Cell Types**

We then examined whether unsupervised clustering of the gene set projections could separate samples by cell type. We trained a Gaussian mixture model on the gene set scores from each method for 80% of the relevant samples and this model was used to predict clusters for the training and test data. In order to evaluate the quality of clustering, we calculated the normalized mutual information (NMI) between the predicted clusters and the known cell types. This procedure was repeated for five distinct groups of three cell types and the results are summarized in Figure 3.

For the training data (Figure 3A), SSCA-based and PLAGE-based models performed best with SSCA-based models having the highest NMI in three cases and PLAGE-based models in two cases. SSCVA-based and GSVA-based models also led to comparatively high NMI scores, while Z-Score-based and Average-based models performed poorly in almost all cases. We observed different results for the test data (Figure 3B), however. The DC1-DC6-Mono1 task led to a tie between the models based on scores from GSVA, PLAGE and SSCVA; on the four remaining tasks, SSCVA-based models and PLAGE-based models each scored highest on two. It is noteworthy that the model trained using SSCVA encodings outperformed the SSCA-based model on the test data, a trend also observed in the logistic regression analysis.

2.4.4 **Top Features Detected for SSCVA and SSCA**

In addition to retaining high-level information about the samples, gene set projection methods should help identify biologically meaningful gene sets from the data. In order to assess whether these new methods can recover known biology, we performed differential analysis using the gene set scores. The first trial focused on the DC6 cells, which are also known as plasmacytoid dendritic cells (Villani et al., 2017). For each of the 550 gene sets, we calculated the median score for all DC6 samples and the median score for all other dendritic cell samples (DC1-5) and ranked the gene sets based on the absolute value of the difference between these medians. We then performed the same analysis comparing all the dendritic cell types (DC1-6) with monocytes (Mono1-4).

The top hits for these trials are shown in Figure 4. For the DC6 vs. DC1-5 experiment (Figure 4A), STAT5A target genes are the 5th ranked feature for SSCVA. STAT5 plays a substantial role in repressing the development of DC6 cells (Esashi et al., 2008) and thus it makes sense this gene set would distinguish DC6 cells from the other dendritic cells. For the dendritic cells vs. monocytes trial (Figure 4B), the top five hits from the SSCA algorithm include targets of AHR (aryl hydrocarbon receptor), which is noteworthy as AHR has been shown to promote the differentiation of monocytes into dendritic cells (Goudot et al., 2017). Additionally, CEBPB (also known as C/EBPβ) targets are the top differential feature for SSCVA and this result is reinforced by research showing that CEBPB is one of the key transcriptional regulators of monocyte cells (Huber et al., 2012). These few examples support the notion that SSCVA and SSCA may be able to utilize transcription factor target gene sets to help identify transcription factors with differential activity between conditions or cell types.

A

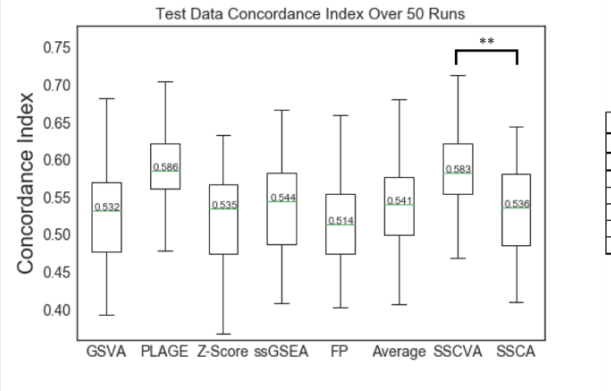
DC6 vs. Other Dendritic Cells (DC1 - 5)		
Rank	**SSCVA**	**SSCA**
1	RACCACAR_AML_Q6	YGTCCTTGR_UNKNOWN
2	ETS_Q4	PAX2_01
3	AAAYWAACM_HFH4_01	SRF_01
4	AREB6_04	EVI1_03
5	STAT5A_03	EVI1_06

B

Dendritic Cells vs. Monocytes		
Rank	**SSCVA**	**SSCA**
1	CEBPB_02	HTF_01
2	ELF1_Q6	YAATNRNNNYNATT_UNKNOWN
3	PU1_Q6	AHR_01
4	ETS2_B	PAX8_B
5	SP1_Q6_01	PAX3_B

Fig 4. Top Five Differential Features for Dendritic Cell Analysis. A) Top features comparing DC6 cells vs. the other five dendritic cell types (DC1 - 5). B) Top features comparing all dendritic cells (DC1 - 6) vs. all monocytes (Mono1 - 4).

A



B

Breast Cancer Survival Prediction				
	SSCVA		SSCA	
Rank	Gene Set	Avg. Rank	Gene Set	Avg. Rank
1	RB_DN.V1_DN	31.64	E2F1_UP.V1_UP	28.74
2	KRAS.50_UP.V1_UP	33.98	KRAS.LUNG.BREAST_UP.V1_DN	36.88
3	RAPA_EARLY_UP.V1_UP	38.58	CRX_DN.V1_UP	41.8
4	MYC_UP.V1_DN	48.58	KRAS.DF.V1_UP	46.5
5	GCNP_SHH_UP_LATE.V1_DN	49.58	E2F3_UP.V1_DN	49.98

Fig 5. Breast Cancer Survival Analysis. A) Box and Whisker Plot for Concordance Index Values. Each gene set projection algorithm was tested 50 times for survival prediction and the concordance index scores are plotted with the median CI value labeled. ** emphasizes the significant difference between SSCVA and SSCA at $p < 0.005$ (Mann-Whitney U test). SSCVA is also significantly different from GSVA, Z-Score, ssGSEA, FP and Average at $p < 0.005$. B) Top ranked features in predicting breast cancer survival (see Methods). Avg. Rank shows the mean rank out of 187 gene sets over the fifty runs.

2.4.5 **Breast Cancer Survival Analysis**

We also analyzed a dataset from The Cancer Genome Atlas (TCGA) that includes RNA-seq data and clinical survival data from 1093 breast cancer patients. In order to attain gene set scores, we first ran the RNA-seq data through the eight projection algorithms using 187 cancer signature gene sets; since the analysis was focused on predicting five-year survival, the dataset was then reduced to the 352 patients that have been followed for more than five years or have passed away. Once the final datasets were processed, we trained a Cox proportional hazards model (Cox PHM) to predict survival from the encodings for each method using 80% of the training data. The trained Cox PHM was then used to predict survival on the training and test data and success was measured by the concordance index between the actual and predicted days of survival. This was repeated fifty times with distinct training/test splits.

When analyzing the Cox PHM predictions on the test data, models for all eight gene set scoring methods showed a wide range of concordance index values across the fifty trials (Figure 5A). PLAGE-based and SSCVA-based models performed best (median concordance index ~0.58), while the other projection methods led to models with a median concordance index of ~0.54. There is no significant difference between the SSCVA and PLAGE results, but SSCVA concordance index values are significantly different than the other six models (p value < 0.005, Mann-Whitney U test).

Additionally, each Cox PHM outputs a list of features ranked by their effect on survival (see Methods). We collected this ranked list for each of the fifty models for the SSCA and SSCVA encodings (Figure 5B). For SSCVA, the top ranked feature across the fifty runs is RB_DN.V1_DN and the RB-loss signature (low RB1) is associated with poor disease outcome in breast cancer (Ertel et al., 2010). Additionally, the top ranked

feature for SSCA is E2F1_UP.V1_UP; this result is supported by previous research as well, as E2F1 transcript levels are related to breast cancer outcome (Hallett and Hassell, 2011).

2.5 Discussion

This work explores shallow sparsely-connected autoencoders (SSCAs) and variational autoencoders (SSCVAs) as methods for projecting RNA-seq data onto gene sets. When using test data, models trained on the SSCVA encodings often performed as well as the models trained on the gene-level RNA-seq data and frequently outperformed (or matched) the existing projection algorithms. SSCA-based models, however, performed well on training data, but poorly on test data. These results suggest that the SSCVA encoding space may be better suited to extrapolation than that of SSCA, but future work is necessary to confirm and interpret this trend.

Additionally, it is difficult to assess a method's ability to recover known biology without a ground truth, but we evaluated SSCA and SSCVA on whether differential analysis produced reasonable results. For the blood scRNA-seq data set, we found the top hits for SSCVA and SSCA included known transcriptional regulators of the groups being tested. Moreover, for the cancer analysis, the top gene sets for both SSCA and SSCVA are cancer signatures related to genes previously associated with breast cancer survival. These observations do not prove that SSCA and SSCVA can uncover insightful biology in all situations, but it is encouraging that the methods identify known features in the data sets tested.

Compared to the other methods discussed, the shallow sparsely-connected autoencoder framework provides greater flexibility for modeling biological phenomena. For instance, if a transcription factor acts as both an activator and a repressor, any given target gene may be up or downregulated. The averaging-based methods (Z-Score, FP, and Average) may miss this trend because the combination of high and low values can reduce the signal. Additionally, the averaging-based approaches and the enrichment-based approaches (ssGSEA and GSVA) both weight all genes equally within a gene set, despite the fact some genes may be more relevant to the gene set than others. PLAGE addresses this issue by learning a specific mapping for each gene set, but the algorithm is limited to finding a linear combination of gene values. SSCA and SSCVA, however, can learn specific nonlinear mappings for each gene set, which could be useful for modeling complex biological relationships. Moreover, the mapping functions learned by SSCA and SSCVA can potentially provide more information about the importance of genes within specific gene sets.

Further exploration is required to better understand the utility of these models for single-cell omics data sets. For instance, SSCVAs may be particularly useful for analysis of cellular differentiation. Variational autoencoders are designed to produce an encoding space where clusters are distinguishable, but close together, and this can result in smooth transitions between groups of samples; thus, the SSCVA scores can potentially be leveraged for identification and visualization of gene sets that transition in

importance throughout differentiation. Additionally, this framework could potentially be applied to other gene-associated omics types, such as methylation.

Unfortunately, a weakness of autoencoder-based methods is that the results may not be entirely consistent between runs; the other six methods tested yield the same result every time, but since autoencoders are initialized randomly each trial, the learned encoder function (and thus the gene set scores) may not be identical across runs. This observation has also been noted by Chen *et al.* (2018) (Chen et al., 2018) and we are currently exploring whether changes in activation functions, hyperparameters, and/or regularization can improve consistency, while maintaining classification accuracy.

Overall this work supports the use of SSCA and SSCVA for gene set analysis on large RNA-seq data sets. These methods still require more rigorous testing and evaluation, and future work on this project will be dedicated to improving consistency between runs and understanding situations and data types where SSCA and/or SSCVA may be particularly useful.

2.6 Acknowledgements

This work was supported by NIH grants R01NS089076 and 1U01CA18498. We would like to thank the PSB reviewers for their thoughtful comments and helpful suggestions and also want to acknowledge Ludwig Schmidt for informative conversations regarding the models.

2.7 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2015). {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Chen, H.-I., Chiu, Y.-C., Zhang, T., Zhang, S., Huang, Y., and Chen, Y. (2018). GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. ArXiv E-Prints.

Chollet, F. (2015). Keras.

Davidson-Pilon, C., Kalderstam, J., Kuhn, B., Fiore-Gartland, A., Moneda, L., Zivich, P., Parij, A., Stark, K., Anton, S., Besson, L., et al. (2018). CamDavidsonPilon/lifelines: v0.14.3.

DeTomaso, D., and Yosef, N. (2016). FastProject: A tool for low-dimensional analysis of single-cell RNA-seq data. BMC Bioinformatics *17*.

Ertel, A., Dean, J.L., Rui, H., Liu, C., Witkiewicz, A., Knudsen, K.E., and Knudsen, E.S. (2010). RB-pathway disruption in breast cancer. Cell Cycle *9*, 4153–4163.

Esashi, E., Wang, Y.H., Perng, O., Qin, X.F., Liu, Y.J., and Watowich, S.S. (2008). The Signal Transducer STAT5 Inhibits Plasmacytoid Dendritic Cell Development by Suppressing Transcription Factor IRF8. Immunity *28*, 509–520.

Goudot, C., Coillard, A., Villani, A.C., Gueguen, P., Cros, A., Sarkizova, S., Tang-Huau, T.L., Bohec, M., Baulande, S., Hacohen, N., et al. (2017). Aryl Hydrocarbon Receptor

Controls Monocyte Differentiation into Dendritic Cells versus Macrophages. Immunity *47*, 582-596.e6.
Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a Shared Vision for Cancer Genomic Data. N. Engl. J. Med.
Hallett, R.M., and Hassell, J.A. (2011). E2F1 and KIAA0191 expression predicts breast cancer patient survival. BMC Res Notes *4*, 95.
Hänzelmann, S., Castelo, R., Guinney, J., Kim, S.C., Seo, Y.J., Chung, W., Eum, H.H., Nam, D.-H., Kim, J., Joo, K.M., et al. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7.
Hu, Q., and Greene, C.S. (2018). Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. BioRxiv.
Huber, R., Pietsch, D., Panterodt, T., and Brand, K. (2012). Regulation of C/EBPβ and resulting functions in cells of the monocytic lineage. Cell. Signal. *24*, 1287–1296.
Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng.
Kang, T., Ding, W., Zhang, L., Ziemek, D., and Zarringhalam, K. (2017). A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. BMC Bioinformatics *18*.
Kingma, D., and Ba, J. (2014). Adam: a method for stochastic optimization. ArXiv Prepr. ArXiv1412.6980 1–13.
Kingma, D.P., and Welling, M. (2013). Auto-Encoding Variational Bayes PPT. Ppt.
Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. PLoS Comput. Biol. *4*.
Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-seq data. Nucleic Acids Res. *45*.
Liou, C.Y., Huang, J.C., and Yang, W.C. (2008). Modeling word perception using the Elman network. In Neurocomputing, pp. 3150–3157.
Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Älgenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol.
Oliphant, T.E. (2007). SciPy: Open source scientific tools for Python. Comput. Sci. Eng.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., and Winther, O. (2016). How to Train Deep Variational Autoencoders and Probabilistic Ladder Networks. Nips *48*.
Sonja, H., Castelo, R., and Guinney, J. (2014). GSVA : The Gene Set Variation Analysis package for microarray and RNA-seq data. Bioconductor.Org 1–20.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M. a, Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.
Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.
Tomfohr, J., Lu, J., and Kepler, T.B. (2005). Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics *6*.
Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science (80-.). *356*, eaah4573.
Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet.

Wang, Y., Solus, L., Dai Yang, K., and Uhler, C. (2017). Permutation-based Causal Inference Algorithms with Interventions. ArXiv Prepr. ArXiv 1705.10220.
Way, G.P., and Greene, C.S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pac. Symp. Biocomput. *23*, 80–91.
Weinstein, J.N., Collisson, E. a, Mills, G.B., Shaw, K.R.M., Ozenberger, B. a, Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.
Xie, R., Wen, J., Quitadamo, A., Cheng, J., and Shi, X. (2017). A deep auto-encoder model for gene expression prediction. BMC Genomics *18*.
žurauskiene, J., and Yau, C. (2016). pcaReduce: Hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics *17*.

Chapter 3- Aberrant Development Corrected in Adult-Onset Huntington's Disease iPSC-Derived Neuronal Cultures via WNT Signaling Modulation

Authors: Charlene Smith-Geater, Sarah J. Hernandez, Ryan G. Lim, Miriam Adam, Jie Wu, Jennifer T. Stocksdale, Brook T. Wassie, Maxwell P Gold, Keona Q. Wang, Ricardo Miramontes, Lexi Kopan, Iliana Orellana, Shona Joy, Paul J. Kemp, Nicholas D. Allen, Ernest Fraenkel, and Leslie M. Thompson.

This was published in Stem Cell Reports 2020 Mar 10; 14(3):406-419. My primary contributions to this project were processing and analyzing the epigenomics data and integrating it with the transcriptomics data.

3.1 Abstract

Aberrant neuronal development and the persistence of mitotic cellular populations have been implicated in a multitude of neurological disorders, including Huntington's disease (HD). However, the mechanism underlying this potential pathology remains unclear. We used a modified protocol to differentiate induced pluripotent stem cells (iPSCs) from HD patients and unaffected controls into neuronal cultures enriched for medium spiny neurons, the cell type most affected in HD. We performed single-cell and bulk transcriptomic and epigenomic analyses and demonstrated that a persistent cyclin $D1^+$ neural stem cell (NSC) population is observed selectively in adult-onset HD iPSCs during differentiation. Treatment with a WNT inhibitor abrogates this NSC population while preserving neurons. Taken together, our findings identify a mechanism that may promote aberrant neurodevelopment and adult neurogenesis in adult-onset HD striatal neurons with the potential for therapeutic compensation.

3.2 Introduction

Huntington's disease (HD) is an autosomal dominant CAG-repeat neurodegenerative disease (Huntington's Disease Collaborative Research Group, 1993). Mutation of the huntingtin (HTT) gene results in an expanded polyglutamine (Qn) repeat within the HTT (mHTT) protein. Individuals with ≥40Q develop HD, with juvenile-onset cases before age 20 years typically occurring above 60 repeats (Cronin et al., 2019). HD symptoms include uncontrollable movement, psychiatric disturbances, and cognitive impairment (Tabrizi et al., 2019) with progressive neurodegeneration and brain atrophy. The primary cell type that degenerates is medium spiny neurons (MSNs) of the striatum, and there remain gaps in our knowledge of early and initiating events in HD systems.

Due to the degenerative nature of HD, postmortem human tissues limit our understanding of early neuropathology of affected brain regions. Human induced pluripotent stem cells (iPSCs) have emerged as important models to study neurodegenerative disease mechanisms (Poon et al., 2017). HD iPSCs differentiated to neural lineages have elucidated deficits in axonal guidance, lipid metabolism, and

neuronal development and function (Conforti et al., 2018, HD iPSC Consortium, 2017, The HD iPSC Consortium, 2012), and dysregulation of the transforming growth factor β (Bowles et al., 2017, Ring et al., 2015), wingless-related integration site (WNT) (Lim et al., 2017), and p53 (Guo et al., 2013, Ring et al., 2015) signaling pathways.

Increased cell proliferation in HD patient brains proportional to disease grade in the subependymal layer that overlies the degenerating caudate nucleus (Curtis et al., 2003, Curtis et al., 2005) suggests increased neurogenesis in the adult brain, potentially as an adaptation to disease. Using HD iPSC-derived neuronal cultures and differentiated embryonic stem cells (ESCs), dysregulation of neurodevelopmental pathways implicates deficits in striatal maturation (Wiatr et al., 2018). These signaling deficits occur early and are associated with phenotypes such as increased vulnerability to growth factor withdrawal in persistent neural progenitors (Mattis et al., 2014, Ring et al., 2015), suggesting that while neurogenesis is initiated, neuronal maturation may be impaired, leaving developmentally arrested cells vulnerable to cellular stressors. The mechanisms involved remain undefined, and these studies have primarily centered on highly expanded juvenile-onset repeat lines, with little known about the more prevalent adult-onset repeat-induced mechanisms.

Here, we use a multi-omics approach to investigate mHTT-associated molecular phenotypes using a modified protocol to differentiate HD patient iPSCs into highly pure populations of neurons with striatal characteristics. RNA sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) analysis reveal significant dysregulation in cell-cycle-associated gene ontology (GO) terms. We show the persistence of a cyclin D1+ mitotic cellular population that emerges among neural progenitors in adult-onset lines and is maintained through terminal differentiation. Single-cell RNA-seq (scRNA-seq) confirms that cell-cycle-related signaling cues arise from the mitotic population of cells, identified as neural stem cells (NSCs), rather than the striatal neurons present in the same population. Furthermore, we show that the persistence of NSC populations is the result of aberrant WNT signaling and we can ameliorate this population through WNT inhibition. This perhaps identifies a pathway or mechanistic target for future therapeutic exploration.

3.3 Results

3.3.1 Modified Protocol for Differentiation of iPSCs to Pure Neurons for Epigenomic Study

Our goal was to define epigenetic signatures of neuronal cultures enriched for MSNs. iPSCs were differentiated to electrophysiologically active neurons (Arber et al., 2015, Telezhkin et al., 2016) with the addition of activin A to promote the formation of lateral ganglionic eminence (LGE)-like progenitors and the subsequent systematic programming of neural progenitors. Comparing the previous protocol (referred to as LI, due to the media containing LDN193189 and IWR1) (Telezhkin et al., 2016) to this protocol with addition of activin A (LIA = medium containing LDN193189, IWR1, and activin A), we observe an increase in the co-staining of Protein Phosphatase 1 Regulatory Inhibitor Subunit 1B (PPP1R1B, alias DARPP-32) and B-cell

lymphoma/leukemia 11B (BCL-11B or CTIP-2) when tested on 33Q control neuronal cultures without any noticeable change to overall neuronal maturation, as seen using phase-contrast microscopy. There is no change in the proliferation of these cell lines at the neural progenitor day 16 stage (Figure S1). Four control lines (CS25iCTR18n2—18Qn2, CS25iCTR18Qn6—18n6, CS14iCTR28n6—28Qn6 [28Q], CS83iCTR33n1—33Qn1 [33Q]), two adult-onset lines (CS04iHD46n10—46Qn10 [46Q] and CS03iHD53n3—53Qn3 [53Q]), and three juvenile-range repeat lines (CS02iHD66n4—66n4 [66Q], CS81iHD71n3—71n3 [71Q], CS09iHD109n1—109n1 [109Q]) were differentiated over 37 days to mature neuronal cultures in duplicate (18Qn2, 46Q) or triplicate (18Qn6, 28Q, 33Q, 53Q, 66Q, 71Q, 109Q) (Figure 1). Quality control measures were performed (Table S1). To control for sex effects, these lines include male (18Q, 46Q, 53Q) and female subjects (28Q, 33Q, 66Q, 71Q, 109Q). Mature neurons were defined by microtubule-associated protein 2 (MAP-2) expression and striatal neurons by DARPP-32 and BCL-11B co-staining (Figure S2); quantification can be found in Table S1. The 46Q iPSC line was the only sample not to reach 10% DARPP-32 expression in the total population following differentiation (Table S1). As this differentiation protocol provides a high percentage of mature neurons (79%–98% MAP-2+ for all cell lines), the following data represent a multi-omics analysis of mature neuronal cultures enriched for striatal characteristics derived from iPSCs within the context of HD.

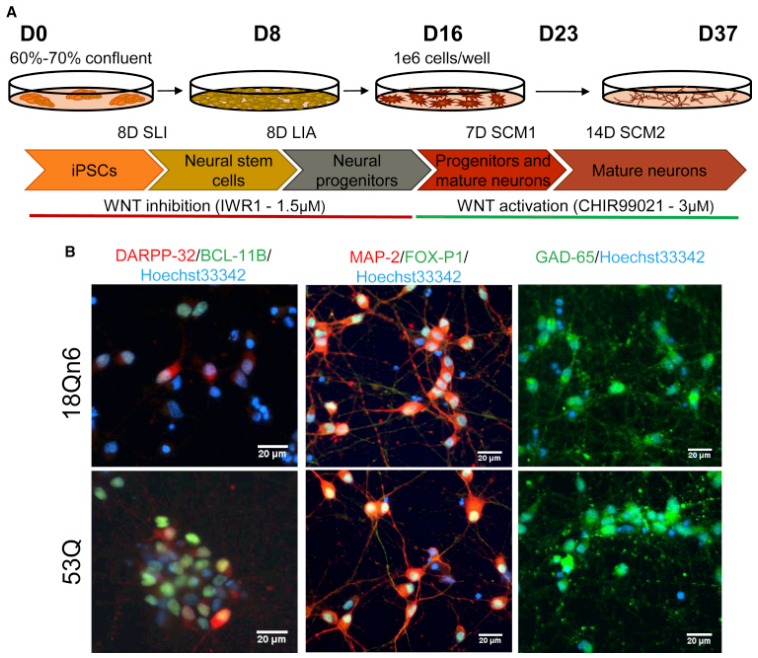


Figure 1: Differentiation Protocol for iPSC-Derived Striatal Neuronal Cultures

(A) Schematic of the differentiation protocol used for iPSC-derived striatal neurons enriched for medium spiny neurons (MSNs) adapted from Telezhkin et al. (2016). (B) Representative immunofluorescence images of cells at day 37 in control (top) and HD (bottom) cells showing that the cells are positive and co-localize for DARPP-32 and BCL-11B (left); cells are also positive for FOX-P1 and MAP-2 (middle) and are GABAergic, as they express the glutamate decarboxylase-2 enzyme (right). Scale bar, 20 μm.

3.3.2 **Transcriptomics Reveal an Upregulation of Cell-Cycle-Related Genes in HD**

Bulk, total RNA-seq was used to identify differentially expressed genes (DEGs) between control and HD mature neuronal cultures enriched for MSNs. Gene expression values and DEGs were used for principal component analysis (PCA), GO enrichment analysis (GOrilla) and Ingenuity Pathway Analysis. Curiously, PCA of control (red) and juvenile-repeat range (green) lines shows no clear separation (Figure S3). Only 105 DEGs were identified (Table S2, columns A–H) when comparing highly expanded repeat lines with controls, and there was no enrichment for specific GO terms, although RE1 Silencing Transcription Factor (REST) was downregulated as described in neural cultures (Buckley et al., 2010, HD iPSC Consortium, 2017). Finding limited DEGs was surprising given the documented transcriptional dysregulation in HD and in models expressing juvenile-range repeat lengths (Valor, 2015). Control (red) and adult-onset (blue) lines, however, show separation along principal component 1 (PC1) (Figure 2A). Given the statistical difference between the control and adult-onset lines and the biological relevance of exploring CAG-repeat lengths more prevalent in the HD population, we continued our analysis to define transcriptional differences between control and adult-onset lines. We identified 823 DEGs in adult-onset neuronal cultures enriched for MSNs compared with control lines (10% false discovery rate [FDR] cutoff), with 754 genes upregulated and 69 genes downregulated (Table S2, columns I–P). The top genes by adjusted p value are labeled in a volcano plot (Figure 2B). Interestingly, the second most statistically significantly upregulated gene was the pluripotency marker POU domain, class 5, transcription factor 1 (POU5F1, alias Oct4) whereas other markers of pluripotency were not dysregulated (NANOG homeobox gene [NANOG], Podocalyxin like [PODXL, alias TRA-1-81], Kruppel-like factor 4 [KLF4], MYC proto-oncogene [MYC]). GO enrichment revealed dysregulation of cell-cycle-associated biological processes (Figure 2C), DNA-binding-associated molecular functions (Figure 2D), and chromosomal gene-associated cellular components (Figure 2E). The top cellular function, predicted to be activated, is cell-cycle progression, with a p value of 7.57×10^{-39}.

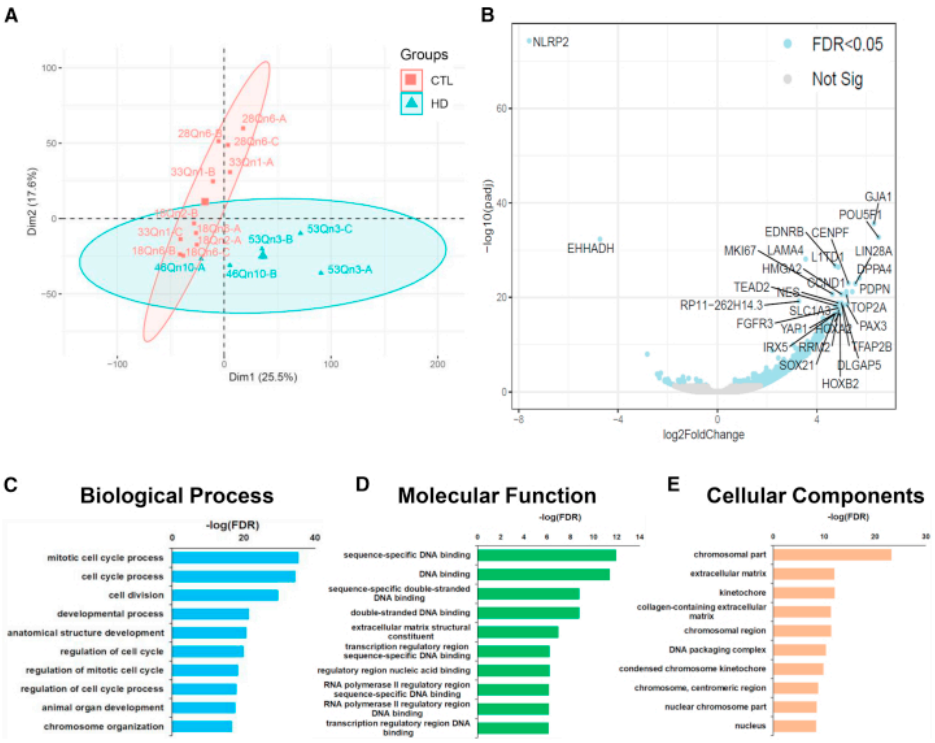


Figure 2: Cell-Cycle-Related Genes Are Upregulated in HD Adult-Onset Cell Lines by RNA-seq
(A) PCA of global gene expression in HD (46Q and 53Q) and control (18Q, 28Q, and 33Q) iPSC-derived striatal neurons. PC1, which captures the most gene expression variance between samples, shows separation of HD and control samples. (B) Volcano plot showing statistically significant DEGs (light blue) between HD (46Q and 53Q) and control (18Q, 28Q, and 33Q) samples. The most significant genes, by FDR, are labeled. Significant gene expression differences are skewed, showing many more upregulated DEGs in the HD samples (Table S2, columns A–H).(C–E) Top GO enrichment analysis terms for biological process (C), molecular function (D), and cellular components (E) showing enrichment for cell cycle, DNA binding, extracellular matrix, and chromosomal genes.

3.3.3 Upregulation of Cell-Cycle-Related Genes and Transcription Factors in HD Identified Using Epigenomic Analysis

We performed ChIP-seq on trimethylated, histone 3 lysine 4 (H3K4me3) residues to identify active promoters in control (18Qn2, 18Qn6, 28Q, 33Q) and HD (46Q, 53Q) neuronal cultures, whereby 99,765 K4me3 sites were found across the lines. Differential analysis of K4me3 occupancy between control and HD lines identified 2,153 sites (5% FDR cutoff); 1,133 of the sites had higher K4me3 levels in HD and 764 sites had higher levels in controls. PCA of differential peaks (Figure 3A) shows clear separation between control and HD K4me3-enriched signals. Differential peaks were annotated to 1,340 genes within a window of 10 kb from transcription start sites (TSS), with a majority of the genes (1,143 genes) having an upregulated K4me3 signal in HD. GO enrichment analysis of the genes annotated to the differential sites shows enrichment of cell-cycle-related terms (Figure 3B).

We overlapped the ChIP-seq and RNA-seq differential signals and found that almost half of the transcriptionally upregulated genes in HD (327 DEGs) had K4me3-upregulated sites in the area next to their TSS (Figure 3C). The subset of genes upregulated both transcriptionally and epigenetically show higher enrichment for cell-cycle GO terms (compare Figures 3D and 3B). We used Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) data to localize the transcription factor (TF)-binding sites within the K4me3 differential sites adjusted to the upregulated DEGs. TF analysis of these sites suggests enrichment for binding of several TFs related to cell-cycle regulation in HD lines (Figure 3E).

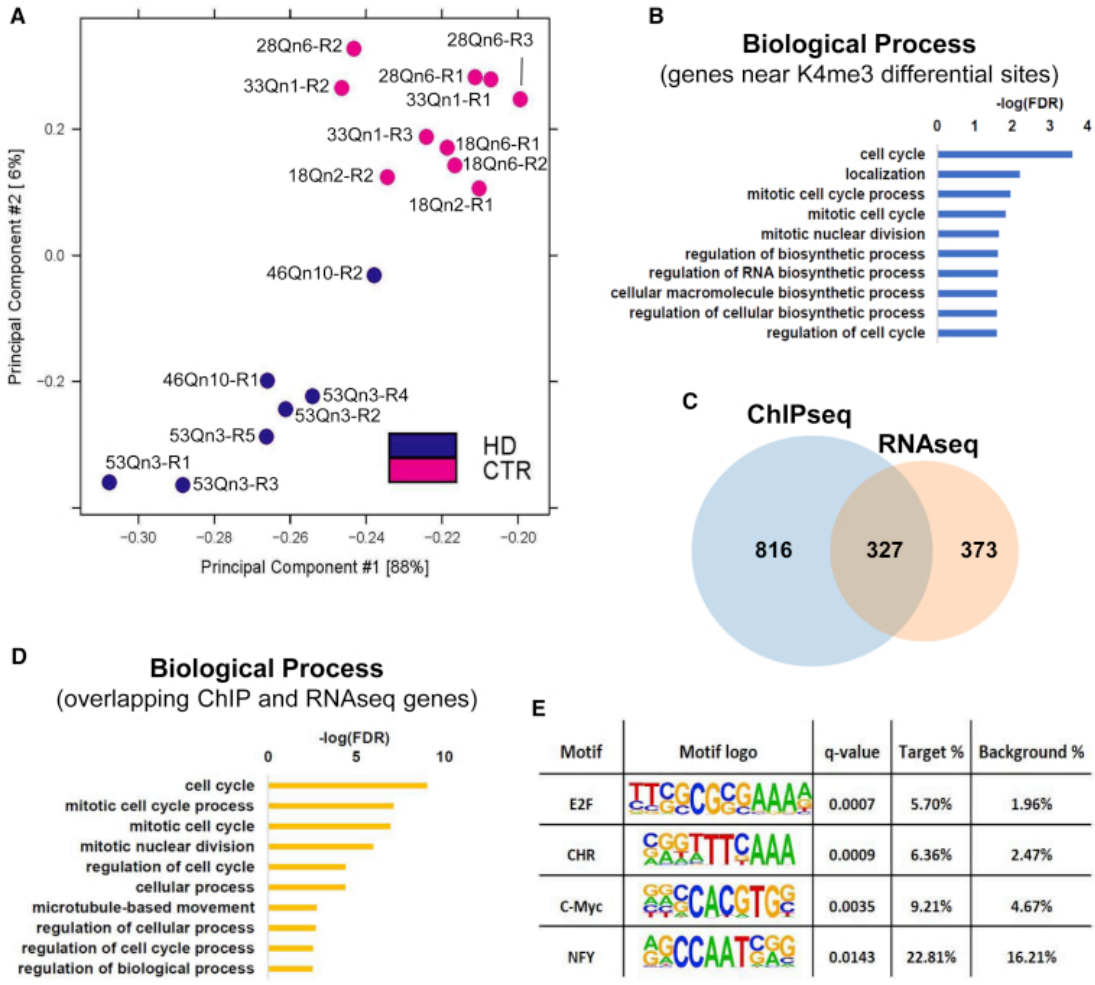


Figure 3: Epigenomics Upregulation of Cell-Cycle-Related Genes in HD
(A) PCA of K4me3 sites identified by ChIP-seq in HD (46Q and 53Q) and control (18Q, 28Q, and 33Q) iPSC-derived striatal neurons. (B) Top GO enrichment analysis terms for genes near K4me3 differential sites show enrichment for cell cycle. (C) Venn diagram of genes upregulated in HD in RNA-seq and ChIP-seq. Upregulated K4me3 sites were found near the TSS of almost half of the transcriptionally upregulated genes in HD. (D) Top GO enrichment analysis terms for the RNA and ChIP-seq overlapping genes show enrichment for cell-cycle processes. (E) Transcription factor (TF) analysis of the regulatory sites near genes upregulated in both RNA and ChIP-seq. Cell-cycle-related TF-binding sites are enriched in HD.

3.3.4 **Adult-Onset HD Cell Lines Have a Persistent Mitotically Active Population in Neuronal Cultures**

Additional adult-onset HD lines were added to the study to determine whether the altered mitotic signature was maintained across different subjects. Within each of the terminally differentiated adult-onset lines, we observe novel populations comprising cells with enlarged cell bodies that lack projections, which are quite distinct from neurons and are absent in control cell lines (Figure 4A). These morphologically distinct cell clusters are readily apparent at the end of culture, as identified by phase-contrast microscopy. These cells are visible at low frequency at day 18 and are observed at higher frequency by day 23 among differentiating neurons. These cell clusters are present in each of four adult-onset HD cell lines examined (CS13iHD43n13 [43Q], 46Q, CS87iHD50n7—50Qn7 [50Q], 53Q) by immunofluorescence and phase-contrast microscopy (Figures S4A and S4B). These aberrant cells are positive for the cell-cycle G1/S phase transition marker cyclin D1, the pluripotency marker Oct4, and the neuroectoderm marker nestin, and are negative for the neuronal marker MAP-2 (Figure 4B), an unexpected finding given that the differentiation protocol produces terminally differentiated, post-mitotic, mature neurons. This population is not the result of the LIA-adapted neural progenitor specification adopted in our protocol, as it is also observed using the previously published LI protocol (Telezhkin et al., 2016) (Figures S1D and S4A), which is important given the role that activin A can play in pluripotency (Beattie et al., 2005).

Flow cytometry was used to quantify the percentage of the population composed of cyclin D1+ cells, which revealed a significant increase ($p = 0.0003$) of cyclin D1+ cells (48.46% ± 4.54%) in terminally differentiated HD lines compared with controls (3.654% ± 1.206%) when grouped together (Figure 4C). Separately, HD lines (50Q, 53Q) have significantly more cyclin D1+ cells than controls (18Qn6 [18Q], CS71iCTR20n6—20Qn6 [20Q]) (Figure S4C). Almost all of the cyclin D1+ cells were also nestin positive (98.5% ± 0.55%) and a high proportion of them were Oct4 positive (73.9% ± 14.1%) (Figure 4D), implying that Oct4 and nestin are expressed in the same cells in a large proportion of the cyclin D1+ population. Of the total day-37 population, Oct4 was expressed at very low levels in control cells (1.03% ± 0.20%) and higher in adult-onset HD cells (22.7% ± 4.34%) (Figure S5). Importantly, there was a very low percentage (<0.054%: 20Q, 46Q, 50Q) of Oct4-positive cells in cultures at early stages (day 8) (Figure S5). Cyclin D1+ cells are proliferative throughout terminal differentiation, and increasingly make up more of the total population as the number of days in vitro progresses; however, this does not seem to have a detrimental effect on the remaining post-mitotic neurons.

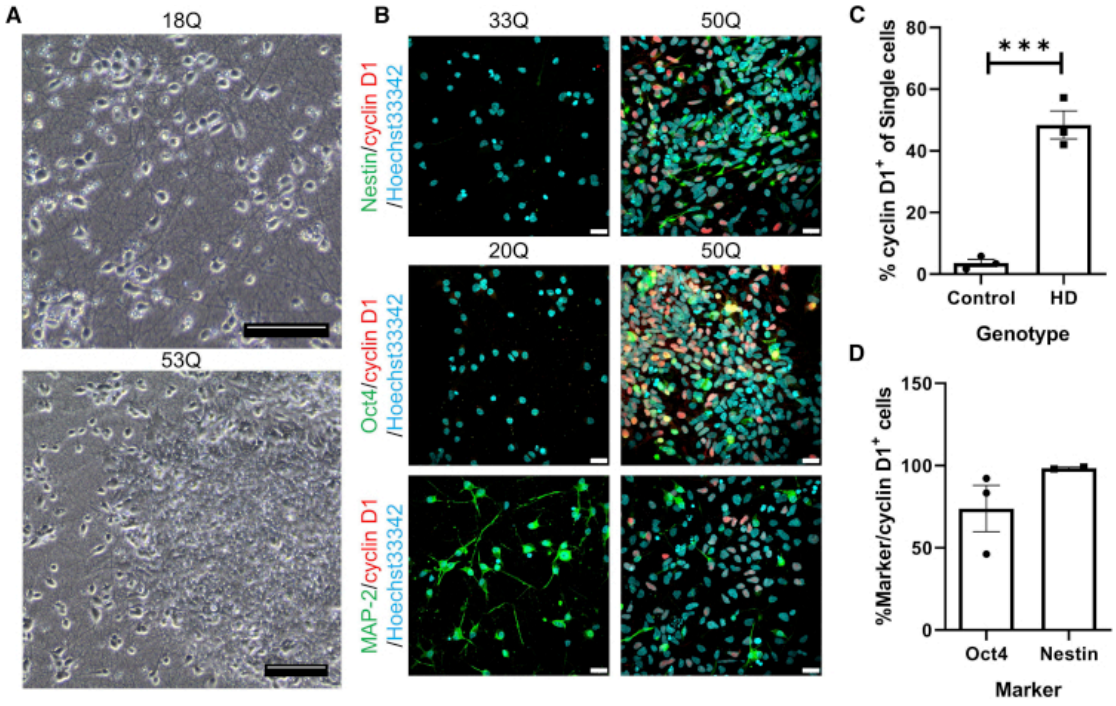


Figure 4: High Percentage of Adult-Onset HD Neuronal Populations Express Cyclin D1 Compared with Controls (A) Phase-contrast images show control (top) and HD (bottom) neurons, with the mitotic population persisting in the HD line occupying the majority of the field. Scale bar, 100 μm. (B) Assessment of the cyclin D1+ population by immunofluorescence shows co-staining of cyclin D1 with nestin (top) and Oct4 (middle), and no co-staining with MAP-2 (bottom). Scale bar, 20 μm. Images are representative of the four adult-onset lines tested (43Q, 46Q, 50Q, 53Q), quantification for which can be found in Figure S4A. (C) Analysis of cyclin D1 expression by flow cytometry shows a significant increase in percentage of cells expressing cyclin D1 compared with controls. Data were acquired across three experiments using three cell lines per condition for two experiments (18Q, 20Q, 33Q, 46Q, 50Q, 53Q) and one experiment with two lines per condition (18Q, 20Q, 50Q, 53Q). Data represent mean ± SEM. Unpaired one-tailed t test, $t = 9.539$, $df = 4$, $p = 0.0003$, $^{***}p < 0.001$. (D) Analysis of co-markers of HD cyclin D1+ cells by flow cytometry shows that the majority of cyclin D1+ cells are Oct4+ and nestin+. Data from 50Q and 53Q (nestin), n = 1 HD cell line; data from 46Q, 50Q, 53Q (Oct4), n = 2 HD cell lines. Data represent mean ± SEM.

3.3.5 Single-Cell RNA-seq Identifies the Cyclin D1+ Population Specific to Adult-Onset Lines as NSCs

To define the composition of the persistent cyclin D1+ cell population present in HD cultures, we performed scRNA-seq on neuronal populations differentiated from one representative control (18Qn6—18Q) and HD iPSC line (53Q). Single-cell viability for each sample was determined (86.9% for 18Q, 80.9% for 53Q). The estimated number of cells sampled was 5,070 (18Q) and 3,829 (53Q) with an average number of reads per cell of 31,675 (18Q) and 41,939 (53Q). After quality control of data filtering, the data were aggregated using Cell Ranger and gene-by-cell expression matrices generated

and used as input for Seurat and Scanpy. The top variable genes by PCA were used for exploratory analysis and visualization using t-distributed stochastic neighbor embedding (t-SNE) (Figure 5A). An unsupervised clustering approach identified seven distinct cell clusters, which were annotated using known cell-type markers (Figure 5B and data not shown). The NSC cluster (Figure 5A) is specific to the HD cell line, with the remaining clusters contributed to by both cell lines. None of the clusters express the astrocyte marker S100 calcium-binding protein β (S100β) or glial fibrillary acidic protein (GFAP), and most clusters express the neuronal marker MAP2, with the neural progenitor marker NES only expressed in the green and brown clusters. Only the NSC cluster in the HD line had high levels of CCND1. CCND1 identifies the aberrant HD population, and we find that other genes are also upregulated in this cluster that are either lowly or not expressed in the other six clusters (Table S2, columns Q–V), including the progenitor marker NES and SRY-box transcription factor 2 (SOX2), which identifies these cells as an NSC-type population.

While the signal from the NSCs likely obscured gene expression differences from the neuronal population in our bulk RNA-seq dataset, scRNA-seq allowed for the identification of gene expression differences in MAP2+, NES/SOX2− neurons between HD and control lines (Table S2, columns W–AB). In this population, genes known to be dysregulated in HD were observed, some of which are: activity-regulated cytoskeleton associated protein (ARC), Fos proto-oncogene, AP-1 transcription factor subunit (FOS), JunB proto-oncogene, AP-1 transcription factor subunit (JUNB), neuronal PAS domain protein 4 (NPAS4), and neurogranin (NRGN), suggesting the possibility that these populations contribute to previously defined HD molecular signatures.

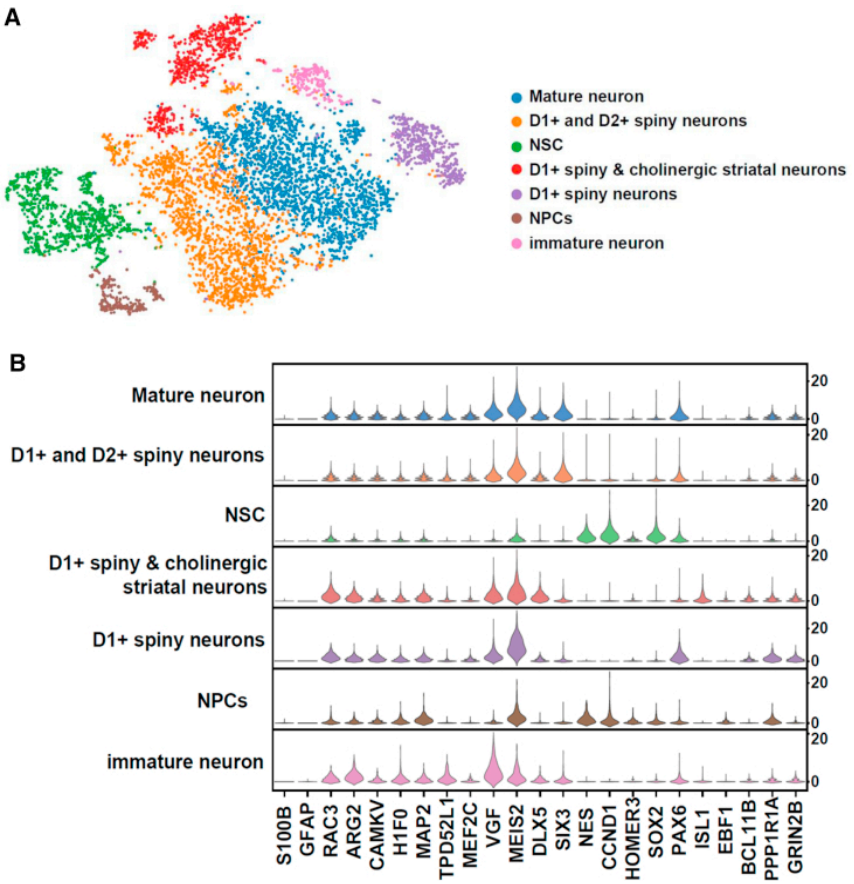


Figure 5: scRNA-seq Reveals a Unique Population of Cells Only in the Adult HD Sample(A) t-SNE plot of the aggregated scRNA-seq data from both the control (18Q) and HD (53Q) samples showing seven cell clusters. Clusters were annotated using known cell-type marker genes and are: mature neuron (blue; TPD52 like 1 [TPD52L1], myocyte enhancer factor 2C [MEF2C], paired box 6 [PAX6], and SIX homeobox 3 [SIX3]); D1 and D2+ spiny neurons (orange; Meis homeobox 2 [MEIS2], distal-less homeobox 5 [DLX5], hippocalcin [HPCA]); NSC cluster (green; nestin [NES], cyclin D1 [CCND1]); D1+ spiny and cholinergic striatal neurons (red; DLX5, ISL LIM homeobox 1 [ISL1], Rac family small GTPase 3 [RAC3], arginase 2 [ARG2]); D1+ spiny neurons (purple; MEIS2, protein phosphatase 1 regulatory inhibitor subunit 1A [PPP1R1A], CaM kinase-like vesicle associated [CAMKV], RAC3, glutamate ionotropic receptor NMDA type subunit 2B [GRIN2B]); neural progenitor cells (NPCs) (brown; CCND1, homer scaffold protein 3 [HOMER3], EBF transcription factor 1 [EBF1]); and immature neurons (pink; VGF nerve growth factor inducible [VGF], H1 histone family member 0 [H1F0]). (B) Stacked violin plot showing expression of glial, neural, and neuronal cell-type markers. CCND1 expression is mainly visible in the NSC cluster, which only exists in the HD (53Q) sample.

3.3.6 Inhibition of the WNT Signaling Pathway Abrogates NSC Populations in HD Neuronal Cultures

Two of the top three transcriptional regulators with the largest changes in RNA-seq expression are transcriptionally regulated by WNT signaling: Oct4 (POU5f1, 6.479-fold increase) and CCND1 (5.186-fold increase). Analysis of TF-binding sites demonstrates that two of the top four regulatory motifs near genes upregulated in both RNA-seq and

ChIP-seq are MYC and E2F, which are both involved with WNT signaling (Hughes and Brady, 2005, Shi et al., 2015, Yeo et al., 2011) (Figure 3E). Examination of the WNT/β-catenin signaling pathway reveals significant dysregulation of several key members, including components of the β-catenin destruction complex, APC regulator of WNT signaling pathway (APC), and glycogen synthase kinase 3β (GSK3β) downregulated in HD lines (Figure 6A). Additionally, scRNA-seq demonstrates that the dysregulation we observe in the WNT/β-catenin signaling pathway from the bulk analysis resides in the NSC population, which has increased expression of Wingless-Type MMTV Integration Site Family, Member 7A (WNT7A), increased expression of Transcription Factor 3 (TCF3), increased expression of frizzled transcripts (frizzled class receptor 2 [FZD2] and frizzled class receptor 7 [FZD7]), decreased expression of β-catenin destruction complex transcripts (APC and GSK3β), and increased expression of WNT transcriptional targets (CCND1) (Figure 6B). We were unable to detect POU5f1 expression in many of the HD cells (about 90 positive cells in HD lines and three positive cells in controls), presumably due to lower sequencing depth using the scRNA-seq platform.

Because of the regulatory role of WNT/β-catenin over the expression of mitotic and pluripotency genes, we hypothesized that downstream correction of the WNT/β-catenin signaling pathway might reduce the mitotic cues necessary for the persistence of the NSC population in HD lines. Therefore, we titrated several WNT inhibitors that act downstream of the β-catenin destruction complex and found that the addition of ICG-001 (hereafter ICG) eliminates the mitotic cell population within HD neuronal cultures. ICG is a selective, potent inhibitor of β-catenin-mediated transcription by preventing the interaction between CREB-binding protein and β-catenin (Emami et al., 2004). ICG addition at day 16 in the differentiation protocol significantly reduced cyclin D1 protein expression in HD lines (50Q, 53Q) by a factor of 10 (Figures 6C and 6D). A highly expanded repeat line (109Q) was assessed to determine the potential effect of ICG based on polyQ length; however, no additional effect was observed (Figure S6). Representative phase-contrast images of a representative adult-onset line (53Q; Figure 6E) demonstrate that ICG treatment abrogates the morphologically distinct NSC populations within adult-onset HD neuronal cultures. Additionally, ICG does not alter MAP-2 expression (Figure S6A), suggesting that the presence of neurons is not altered by treatment. While expression of synaptophysin and Disks Large MAGUK Scaffold Protein 4 (alias PSD95) remain unchanged in HD lines (Figures S6B and S6C), expression of both is reduced in control lines by ICG treatment (Figure S6B). Notably, for these differentiations the 50Q line overwhelmingly had the highest percentage of mitotic cells and therefore, correspondingly, DMSO-treated 50Q cells have the lowest levels of MAP-2, PSD95, and synaptophysin overall.

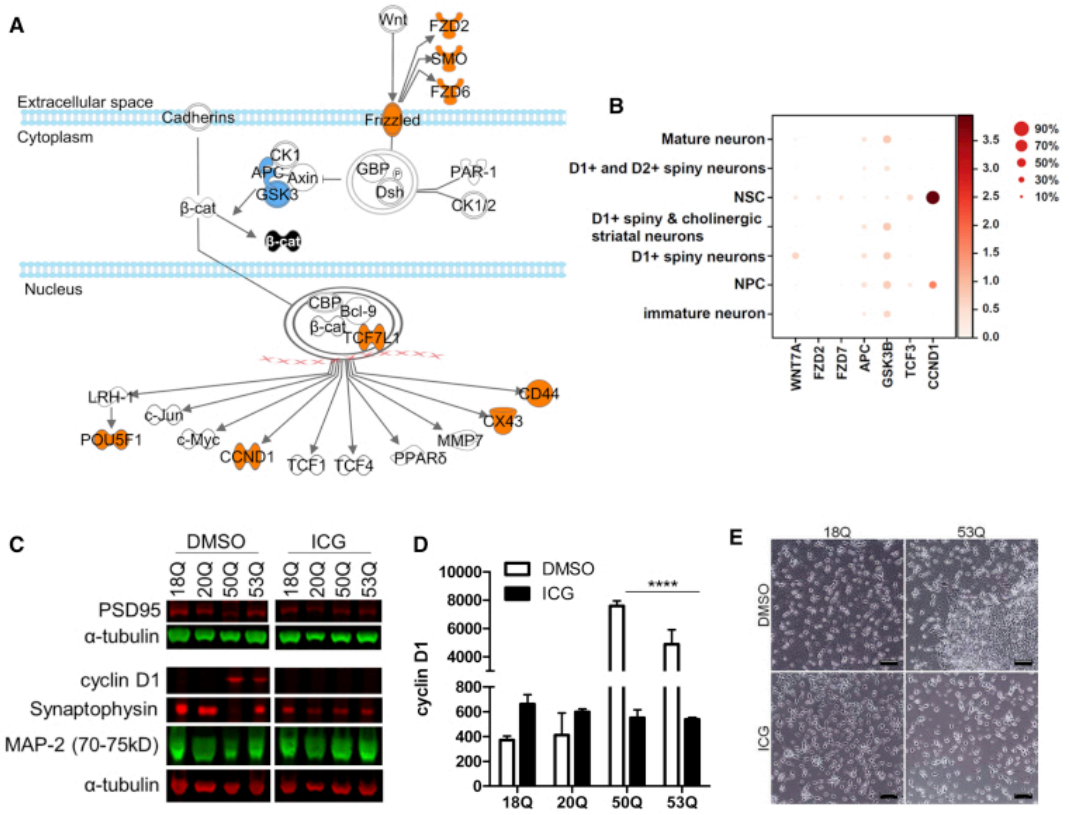


Figure 6: WNT/β-Catenin Signaling Is Significantly Dysregulated and WNT Inhibition Rescues Aberrant Cyclin D1 Overexpression while Abrogating the Mitotic Population of Cells in Adult-Onset Neurons (A) Dysregulation of the WNT/β-catenin signaling pathway from bulk RNA-seq was analyzed using Ingenuity Pathway Analysis. While Frizzled receptors are upregulated, members of the β-catenin destruction complex, specifically APC and GSK3β, are downregulated, allowing for upregulation of WNT/β-catenin transcriptional targets, such as CCND1 and Oct4 (POU5F1). Orange molecules are upregulated, blue are downregulated, and black are targeted for destruction. A full legend can be found in Figure S5. 46Q and 53Q lines were compared with controls with 18Q (2 clones), 28Q, and 33Q in triplicate using a 0.1 FDR. (B) Dot plots show expression of WNT/β-catenin pathway genes and CCND1. The NSC cluster shows higher expression of all WNT-related genes except for genes of the destruction complex, APC and GSK3β, and WNT7A, which shows higher expression in the NSC cluster over all other clusters besides the cortical cluster. (C) The WNT inhibitor ICG or DMSO was added to the NPC population at day 16 of differentiation. Cells were harvested at day 37 and western blots were performed for cyclin D1, MAP-2, synaptophysin, and PSD95 using LI-COR. (D) Quantitation demonstrates that treatment with ICG completely rescues aberrant overexpression of cyclin D1 in adult-onset striatal neurons (50Q and 53Q). For analysis, lines were grouped by genotype. White bars represent DMSO-treated cells and black bars represent ICG-treated cells. Analysis was done using a two-way ANOVA with Bonferroni correction. Multiplicity-adjusted p values were calculated for each comparison at a 95% confidence interval. Data represent mean ± SEM. ****$p < 0.0001$. Data are from two independent experiments using four cell lines grouped by control (18Q, 20Q) and HD (50Q, 53Q). Further statistical details can be found in Tables S3 and S4. (E) Representative phase images show the mitotic population of cells at day 37 in adult-onset line (53Q) treated with DMSO (top-right panel), which is completely abrogated by ICG addition (bottom-right panel). The 53Q line is a representative adult-onset line. Scale bar, 100 μm.

3.4 Discussion

Multi-omics-level analyses were performed on HD neuronal cultures enriched for MSNs using a protocol that induces increased expression of LGE markers including distal-less homeobox 1 and 2 (DLX1/2) and GS homeobox 2 (GSX2) mRNA. This response is likely via upregulation of BCL-11B, which is important for MSN development (Arlotta et al., 2008). Introduction of activin A increased co-localization of BCL-11B and DARPP-32+ neurons, which is more indicative of MSNs than DARPP-32 positivity alone (Carri et al., 2013). Total RNA-seq and scRNA-seq showed that the primary transcriptional differences were reflected in a persistent population of NSCs within terminally differentiated, post-mitotic neuronal cultures from patients with adult-onset repeats. These findings are consistent with mitotic aberrations found in HD iPSCs differentiated toward a striatal lineage (Mattis et al., 2014). Here, we uncover a key mechanism underlying this mitotic persistence.

mHTT has an impact on neurogenesis in HD knockin (Molero et al., 2009), transgenic R6/1 (Lazic et al., 2004), and R6/2 (Phillips et al., 2005) mouse models, and patient postmortem tissue (Curtis et al., 2003, Curtis et al., 2005). Neurodevelopmental changes associated with mHTT expression are observed in other stem cell-derived models: HTT-null NSCs derived from mouse ESCs (Lo Sardo et al., 2012), HD patient-derived iPSC striatal cells (Conforti et al., 2018), and an isogenic allelic series of human ESCs differentiated toward a cortical neuron fate form disorganized neural rosettes (Ruzo et al., 2018), suggesting HD-related aberrations during development of the neural tube. While technical limitations of scRNA-seq limit analysis of Oct4 transcript abundance specifically within the NSC population, our observation of Oct4 persistence by bulk RNA-seq is corroborated by persistent Oct4 expression found in neuronal differentiations (Conforti et al., 2018). Expression of mHTT impairs Oct4 protein downregulation even in highly expanded repeat lines (60Q, 109Q, 180Q) during differentiation when control cultures transition from pluripotency to neuroectoderm formation. Aberrant maintenance of Oct4 expression in ESC-derived neuronal cultures has also been noted with expanded CAG repeats in other contexts, for example within the hypoxanthine phosphoribosyltransferase (Hprt) gene (Lorincz et al., 2004). Temporal dysregulation of other developmental markers has also been observed; juvenile HD iPSCs differentiated toward a striatal fate maintain nestin+ NPCs susceptible to death by brain-derived neurotrophic factor withdrawal, suggesting that mitotically persistent populations of cells are vulnerable to factors mimicking the HD environment (Mattis et al., 2014). These nestin+ NPCs were reduced by targeting the canonical Notch signaling pathway and lowering HTT by antisense oligonucleotide treatment (Mathkar et al., 2019). Additionally, transcriptomic analysis of HD iPSC isogenic lines showed that differences related to neuronal development and dorsal striatum formation occur at the NSC stage (Ring et al., 2015), further implicating the NSC stage in manifestation of developmental aberrations due to mHTT expression. In HD patient iPSC-derived neural cultures, one-third of all gene expression changes were related to pathways involved in neuronal development and maturation (HD iPSC Consortium, 2017).

Several studies have implicated mitotic persistence in altered neurogenesis in HD (Agus et al., 2019, Ooi et al., 2019). ESC-isogenic HD lines differentiated into forebrain neuronal cultures showed altered transcriptional signatures for cell-cycle progression. Other studies implicated cell-cycle re-entry as a mechanism of mitotic dysregulation in HD (Ditsworth et al., 2017, Liu et al., 2015, Pelegrí et al., 2008) and other neurodegenerative diseases (Höglinger et al., 2007, Seward et al., 2013). In contrast, our data indicate a failure to exit the cell cycle as the primary mechanism of persistent mitosis. It is possible that rare cells persist randomly and give rise to a proliferating population during differentiation as opposed to a specific mHTT-mediated effect; however, this is unlikely given that this persistent mitotic population is only observed in the adult-onset lines, not control lines, and is present in all adult-onset lines tested (n = 4). Furthermore, scRNA-seq data showed no clusters having significant differentially expressed levels of CCND1 between HD and control cells that also showed high levels of mature neuronal marker expression, which would be expected if mature neurons were re-entering the cell cycle.

The work presented offers a mechanism for mitotic persistence, aberrant WNT signaling leading to propagation of mitotic cell populations. This prolonged mitotic state may underlie the transient excess neurogenesis observed in animal models of HD and patient adult striatum (Curtis et al., 2003, Ernst et al., 2014, Tattersfield et al., 2004). Altered WNT signaling is implicated in HD pathogenesis. β-Catenin pathway modulation rescued the toxic effects of mHTT expression and increased the life span in HD Drosophila, attenuated neurodegeneration of primary striatal neurons (Godin et al., 2010), and WNT dysregulation may promote aberrant angiogenesis-related signaling observed in HD iPSC-brain microvascular endothelial cells (Lim et al., 2017), suggesting that this pathway may be critical in understanding primary mechanisms affected as a consequence of chronic mHTT expression.

While HTT interacts with the β-catenin destruction complex in vitro and in vivo, the expanded polyQ region in mHTT prevents binding, leading to toxic β-catenin stabilization and eventual neurodegeneration (Godin et al., 2010). One possibility for the presence of the NSC population only in adult-onset and not in juvenile-range repeat HD neuronal cultures is a threshold activation of the WNT pathway controlled by polyQ length. In neurons, highly expanded polyQ lengths may cause structural changes different from those observed in adult-onset cases that further dysregulate protein-protein interactions and downstream signaling (Caron et al., 2013) such that the highly expanded polyQ region no longer prevents β-catenin from binding the destruction complex and allows for β-catenin degradation, thus preventing transcription of WNT targets such as CCND1 and Oct4. This could allow juvenile lines to follow the same developmental paradigm as control lines without the presence of aberrant NSCs in neuronal populations.

Many neuronal induction protocols exploit inhibition of the WNT pathway due to its role in determining neural crest cell fate (Menendez et al., 2011); two striatal differentiation protocols add the WNT inhibitor DKK1 after initial neural induction (Aubry et al., 2008,

Delli Carri et al., 2013). Consistent with the hypothesis that WNT may be differentially activated in the adult-onset versus juvenile-repeat striatal-enriched neuronal cultures, RNA-seq showed that β-catenin destruction complex members were downregulated in adult-onset lines compared with similar levels of both APC and GSK3β in control and juvenile lines (data not shown). With the addition of CHIR99021, the WNT agonist, in our protocol, the adult-onset lines may no longer be able to compensate for the overactivation of WNT, thereby uncovering a previously undefined mechanism.

We expected to uncover myriad cell-intrinsic differences between control and juvenile HD neuronal cultures and were surprised to find few transcriptional differences. This may be informative for the neurodegenerative disease field in the context of pure populations of iPSC-derived cell types to investigate cell-intrinsic effects. Furthermore, these findings perhaps highlight the phenotypic differences observed between adult- and juvenile-repeat HD, which present quite differently, with manifestation of dyskinesia and global neurodegeneration observed in juvenile cases not present with adult onset (Ruocco et al., 2006). We do not know whether the lack of transcriptional differences between control and juvenile groups is a consequence of (1) juvenile-repeat cell cultures bypassing some point in the differentiation process, (2) cells within the juvenile-repeat lines that remain as NSCs being highly susceptible to cell death, (3) the length of the polyQ repeat expansion acting as flexible domain that modulates the spatial proximity of N17 and polyproline-flanking domains, thereby affecting downstream signaling (Caron et al., 2013), or (4) the need for signals from other cell types for full manifestation of disease. However, regarding point (4), scRNA-seq data from MAP2+, NES/SOX2− neurons identifies genes known to be dysregulated in HD neurons, suggesting that there are HD signatures in adult-onset iPSC-derived neurons that overlap with previously published data (HD iPSC Consortium, 2017). The protocol used here reflects very early cell-intrinsic effects. Thus far, we have observed the mitotic population of cells in all adult-onset lines differentiated with this protocol, but only rarely in juvenile-repeat lines and never in control lines.

Overall, our data show that there is a persistent, mitotically active cell population unique to adult-onset repeat HD cell lines that we can prevent in vitro by targeting the WNT pathway with the application of the WNT inhibitor ICG. While ICG is used as a tool to demonstrate the mechanism behind the persistence of NSCs in the context of HD, we speculate that specific components of the WNT signaling pathway inhibited by ICG could be therapeutic targets for HD with the potential to compensate for neurodevelopmental defects.

3.5 Methods

3.5.1 **iPSC Differentiation**

iPSC lines were generated as described by HD iPSC Consortium (2017). iPSCs were cultured on hESC-qualified Matrigel (Corning) in mTESR1 (STEMCELL Technologies) and passaged with Versene (Gibco). At 60%–70% confluency, medium was changed to start neural differentiation as described by Telezhkin et al. (2016) However, at day 8

cells were replated in activin A-containing medium (see Supplemental Experimental Procedures).

3.5.2 **Immunofluorescence**

Cells grown on glass coverslips were fixed with 4% paraformaldehyde (PFA; Electron Microscopy Sciences) in PBS for 10 min. Antibody incubations and imaging are described in Supplemental Experimental Procedures.

3.5.3 **RNA-seq**

Total RNA was isolated using the Qiagen RNeasy Kit and QIAshredders for cell lysis. One microgram of RNA with RNA integrity number values >9 were used for library preparation using the strand-specific Illumina TruSeq Total RNA protocol. Libraries were sequenced on the HiSeq 2500 platform using 100 cycles to obtain paired-end 100 reads at >50 million reads per sample.

3.5.4 **scRNA-seq**

Samples for scRNA-seq were harvested using Dispase (Gibco) for 30 min at 37°C, pipetted off the plate with advanced DMEM/F12 (1:1) with 2% B27, and collected by centrifugation. Cells in 0.1% BSA PBS were filtered through a 70-µm cell strainer and counted. Cells were resuspended at 700 cells/µL in 0.04% BSA in PBS at >80% viability.

3.5.5 **ChIP-Seq**

Frozen cell pellets (minimum of two replicates per line) were resuspended in lysis buffer and incubated on ice for 20 min. Chromatin was digested and samples processed as described in Supplemental Experimental Procedures. Libraries were prepared using NEBNExt Ultra II DNA Library Prep Kit for Illumina (NEB) and sequenced on an Illumina HiSeq 2000 for single-end 50-bp reads.

3.5.6 **ATAC-Seq**

Cryostored cell samples were processed in duplicates as previously described (Milani et al., 2016). The ATAC-seq libraries were sequenced on an Illumina HiSeq 2000 for single-end 50-bp reads.

3.5.7 **Data Processing**

For RNA-seq, fastq files were trimmed using a base quality score threshold of >20 and aligned to the hg19 genome with TopHat 2. Reads passing quality control were used for

quantification using HTSeq and analyzed with the R package DESeq2 to identify DEGs. Genes passing an FDR of 10% were used for GO enrichment analysis using GOrilla (Eden et al., 2007, Eden et al., 2009) (http://cbl-gorilla.cs.technion.ac.il/).

For scRNA-seq, fastq files were quality controlled and aligned to the reference transcriptome to obtain a gene count matrix. Before downstream tertiary analysis, quality control of cells was performed by filtering out low-quality cells and low-abundance genes. After quality control and expression matrix formation, normalization was performed using UMI (unique molecular identifiers), and log transformation was used to control variance. Dimension reduction was used to visualize and explore major features of the data using PCA and t-SNE, and differential expression statistical analyses were performed using Seurat and Scanpy. Cell types were assigned based on known cell-type markers.

For ChIP-seq, fastq files were aligned to the hg38 genome assembly using Bowtie2. The ENCODE blacklist regions and PCR duplicates were removed (using SAMtools and Picard). Peak calling was done using MACS2 (--broad flag, q-value < 0.1). Differential analysis of the peaks was performed with Diffbind (Ross-Innes et al., 2012) (http://bioconductor.org/packages/DiffBind/), using EdgeR with FDR <0.05. Differential peaks were annotated to genes within a window of 10 kb from the TSS. GO enrichment analysis was done using Panther (Mi et al., 2012, Thomas et al., 2003) (http://pantherdb.org/webservices/go/overrep.jsp).

For ATAC-seq, fastq files were aligned to hg38 genome assembly using Bowtie2. ENCODE blacklist regions and PCR duplicates were removed (using SAMtools and Picard). Peak calling was done using MACS2 (--nomodel --shift −100 --extsize 200 settings and q-value < 0.05). The consensus peak file for all samples was generated using Diffbind.

3.5.8 **Motif Analysis**

HOMER (v.4.9.1) was used to identify TF motifs that are enriched in the differential ChIP-seq peaks. ATAC-seq data were used to identify TF-binding locations; thus, for motif analysis we considered consensus ATAC-seq peaks that overlap with the aforementioned differential ChIP-seq peaks. To identify enriched motifs, we used the findMotifs function in HOMER with default parameters; we looked for motif enrichment in the consensus ATAC-seq peaks within ChIP-seq differential peaks compared with a background of all consensus ATAC-seq peaks that are within 10 kb of the TSS of any gene.

3.5.9 **Flow Cytometry**

Neuronal cultures were harvested at day 37 and fixed in suspension with 4% PFA. Aliquots of cells were made for FMO (Fluorescence Minus One) and unstained controls pooled from all samples incubated for 2 h at room temperature with primary antibody,

washed three times, and incubated with secondary antibody for 30 min. Cells were washed and resuspended in FACS (fluorescence-activated cell sorting) buffer, pipetted through a cell strainer (Falcon), and run on a BD LSRII flow-cytometry machine. Gating was performed using BD FACSDiva software. Further analysis was performed using De Novo FCS Express 6.

3.5.10 **Western Blotting**

Cells for western blots were scraped and harvested in cold PBS-Mg2+/-Ca2+ and flash frozen until processing. Analysis was performed using the LI-COR Odyssey CLx imaging system (see Supplemental Experimental Procedures).

3.5.11 **Statistical Analysis**

Error bars represent the mean ± SEM where stated. One-tailed t tests were used for comparison of two groups of data. For dataset comparisons of three or more, one-way ANOVA was used; for comparing datasets with two variables among multiple samples, two-way ANOVA was used as stated. Statistical analyses were performed using GraphPad Prism v.8 using $p < 0.05$ as significant. Further information from statistical tests not in figure legends can be found in Tables S3 and S4. For specific assay statistical information, see Supplemental Experimental Procedures.

3.6 Author Contributions

L.M.T., E.F., C.S.-G., and S.J.H. conceived and designed the experiments. C.S.-G., S.J.H., M.A., J.T.S., K.Q.W., L.K., and I.O. performed the experiments and carried out data analysis. S.J., C.S.-G., N.D.A., and P.J.K. developed the differentiation paradigm. R.G.L., J.W., B.T.W., M.P.G., and R.M. performed omics data analysis. S.J.H., C.S.-.G., M.A., R.G.L., E.F., and L.M.T. wrote the manuscript. L.M.T. and E.F. supervised the project and acquired funding.

3.7 Acknowledgments

We thank the HD patients and their families for their essential contributions to this research and Dr. David Housman for helpful discussions of the data. Primary support for this work was from NIH NS089076 (L.M.T., E.F.) and the CHDI Foundation (P.J.K., N.D.A.). Additional support was provided by NIH (U54 NS091046 NeuroLINCS center, L.M.T., E.F.; NIH NS078370, L.M.T., E.F.), California Institute for Regenerative Medicine (CIRM) (C.S.-G.), the Hereditary Disease Foundation (C.S.-G.), the UCI Institute for Clinical and Translational Science (L.M.T.), HDSA Berman/Topper Career Development Fellowship (S.J.H.), HD CARE (L.M.T.), NIH T32 grant TG32 AG00096 (S.J.H.), and the UK Medical Research Council (P.J.K. and N.D.A.). This work was made possible, in part, through access to the Genomic High Throughput Facility Shared Resource of the Cancer Center Support grant (CA-62203) at the University of California,

Irvine. Support also included computing resources from National Science Foundation grant DB1-0821391 and sequencing support from National Institutes of Health grant P30-ES002109. The authors declare no competing interests.

3.8 References

Agus F., Crespo D., Myers R.H., Labadorf A. The caudate nucleus undergoes dramatic and unique transcriptional changes in human prodromal Huntington's disease brain. BMC Med. Genomics. 2019;12:137.
Arber C., Precious S.V., Cambray S., Risner-Janiczek J.R., Kelly C., Noakes Z., Fjodorova M., Heuer A., Ungless M.A., Rodriguez T.A. Activin A directs striatal projection neuron differentiation of human pluripotent stem cells. Development. 2015;142:1375–1386.
Arlotta P., Molyneaux B.J., Jabaudon D., Yoshida Y., Macklis J.D. Ctip2 controls the differentiation of medium spiny neurons striatum. J. Neurosci. 2008;28:622–632.
Aubry L., Bugi A., Lefort N., Rousseau F., Peschanski M., Perrier A.L. Striatal progenitors derived from human ES cells mature into DARPP32 neurons in vitro and in quinolinic acid-lesioned rats. Proc. Natl. Acad. Sci. U S A. 2008;105:16707–16712.
Beattie G.M., Lopez A.D., Bucay N., Hinton A., Firpo M.T., King C.C., Hayek A. Activin A maintains pluripotency of human embryonic stem cells in the absence of feeder layers. Stem Cells. 2005;23:489–495.
Bowles K.R., Stone T., Holmans P., Allen N.D., Dunnett S.B., Jones L. SMAD transcription factors are altered in cell models of HD and regulate HTT expression. Cell. Signal. 2017;31:1–14.
Buckley N.J., Johnson R., Zuccato C., Bithell A., Cattaneo E. The role of REST in transcriptional and epigenetic dysregulation in Huntington's disease. Neurobiol. Dis. 2010;39:28–39.
Caron N.S., Desmond C.R., Xia J., Truant R. Polyglutamine domain flexibility mediates the proximity between flanking sequences in huntingtin. Proc. Natl. Acad. Sci. U S A. 2013;110:14610–14615.
Carri A.D., Onorati M., Castiglioni V., Faedo A., Camnasio S., Toselli M., Biella G., Cattaneo E. Human pluripotent stem cell differentiation into authentic striatal projection neurons. Stem Cell Rev. Rep. 2013;9:461–474.
Conforti P., Besusso D., Bocchi V.D., Faedo A., Cesana E., Rossetti G., Ranzani V., Svendsen C.N., Thompson L.M., Toselli M. Faulty neuronal determination and cell polarization are reverted by modulating HD early phenotypes. Proc. Natl. Acad. Sci. U S A. 2018;115:E762–E771.
Cronin T., Rosser A., Massey T. Clinical presentation and features of Juvenile-onset Huntington's disease: a systematic review. J. Huntingtons Dis. 2019;8:171–179.
Curtis M.A., Penney E.B., Pearson A.G., Van Roon-mom W.M.C., Butterworth N.J., Dragunow M., Connor B., Faull R.L.M. Increased cell proliferation and neurogenesis in the adult human Huntington's disease brain. Proc. Natl. Acad. Sci. U S A. 2003;100:9023–9027.
Curtis M.A., Penney E.B., Pearson J., Dragunow M., Connor B., Faull R.L.M. The distribution of progenitor cells in the subependymal layer of the lateral ventricle in the normal and Huntington's disease human brain. Neuroscience. 2005;132:777–788.

Delli Carri A., Onorati M., Lelos M.J., Castiglioni V., Faedo A., Menon R., Camnasio S., Vuono R., Spaiardi P., Talpo F. Developmentally coordinated extrinsic signals drive human pluripotent stem cell differentiation toward authentic DARPP-32+ medium-sized spiny neurons. Dev. Stem Cells. 2013;140:301–312.
Ditsworth D., Maldonado M., McAlonis-Downes M., Sun S., Seelman A., Drenner K., Arnold E., Ling S.-C., Pizzo D., Ravits J. Mutant TDP-43 within motor neurons drives disease onset but not progression in amyotrophic lateral sclerosis. Acta Neuropathol. 2017;133:907–922.
Eden E., Lipson D., Yogev S., Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput. Biol. 2007;3:e39.
Eden E., Navon R., Steinfeld I., Lipson D., Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.
Emami K.H., Nguyen C., Ma H., Kim D.H., Jeong K.W., Eguchi M., Moon R.T., Teo J.-L., Oh S.W., Kim H.Y. A small molecule inhibitor of β-catenin/cyclic AMP response element-binding protein transcription. Proc. Natl. Acad. Sci. U S A. 2004;101:12682–12687.
Ernst A., Alkass K., Bernard S., Salehpour M., Perl S., Tisdale J., Possnert G., Druid H., Frisen J. Neurogenesis in the striatum of the adult human brain. Cell. 2014;156:1072–1083.
Godin J.D., Poizat G., Hickey M.A., Maschat F., Humbert S. Mutant huntingtin-impaired degradation of β-catenin causes neurotoxicity in Huntington's disease. EMBO J. 2010;29:2433–2445.
Guo X., Disatnik M.-H., Monbureau M., Shamloo M., Mochly-Rosen D., Qi X. Inhibition of mitochondrial fragmentation diminishes Huntington's disease-associated neurodegeneration. J. Clin. Invest. 2013;123:5371–5388.
HD iPSC Consortium Developmental alterations in Huntington's disease neural cells and pharmacological rescue in cells and mice. Nat. Neurosci. 2017;20:648–660.
Höglinger G.U., Breunig J.J., Depboylu C., Rouaux C., Michel P.P., Alvarez-Fischer D., Boutillier A.-L., DeGregori J., Oertel W.H., Rakic P. The pRb/E2F cell-cycle pathway mediates cell death in Parkinson's disease. Proc. Natl. Acad. Sci. U S A. 2007;104:3585–3590.
Hughes T.A., Brady H.J.M. Cross-talk between pRb/E2F and Wnt/β-catenin pathways: E2F1 induces axin2 leading to repression of Wnt signalling and to increased cell death. Exp. Cell Res. 2005;303:32–46.
Huntington's Disease Collaborative Research Group A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72:971–983.
Lazic S.E., Grote H., Armstrong R.J.E., Blakemore C., Hannan A.J., van Dellen A., Barker R.A. Decreased hippocampal cell proliferation in R6/1 Huntington's mice. Neuroreport. 2004;15:811–813.
Lim R.G., Quan C., Reyes-Ortiz A.M., Housman D.E., Agalliu D., Thompson L.M., Lim R.G., Quan C., Reyes-ortiz A.M., Lutz S.E. Huntington's disease iPSC-derived brain microvascular endothelial cells reveal WNT- article Huntington's disease iPSC-derived brain microvascular endothelial cells reveal WNT-mediated angiogenic and blood-brain barrier deficits. Cell Rep. 2017;19:1365–1377.

Liu K.Y., Shyu Y.C., Barbaro B.A., Lin Y.T., Chern Y., Thompson L.M., Shen C.K.J., Marsh J.L. Disruption of the nuclear membrane by perinuclear inclusions of mutant huntingtin causes cell-cycle re-entry and striatal cell death in mouse and cell models of Huntington's disease. Hum. Mol. Genet. 2015;24:1602–1616.
Lorincz M.T., Detloff P.J., Albin R.L., O'Shea K.S. Embryonic stem cells expressing expanded CAG repeats undergo aberrant neuronal differentiation and have persistent Oct-4 and REST/NRSF expression. Mol. Cell. Neurosci. 2004;26:135–143.
Lo Sardo V., Zuccato C., Gaudenzi G., Vitali B., Ramos C., Tartari M., Myre M.A., Walker J.A., Pistocchi A., Conti L. An evolutionary recent neuroepithelial cell adhesion function of huntingtin implicates ADAM10-Ncadherin. Nat. Neurosci. 2012;15:713–721.
Mathkar P.P., Suresh D., Dunn J., Tom C.M., Mattis V.B. Characterization of neurodevelopmental abnormalities in iPSC-derived striatal cultures from patients with Huntington's disease. J. Huntingtons Dis. 2019;8:257–269.
Mattis V.B., Tom C., Akimov S., Saeedian J., Østergaard M.E., Southwell A.L., Doty C.N., Ornelas L., Sahabian A., Lenaeus L. HD iPSC-derived neural progenitors accumulate in culture and are susceptible to BDNF withdrawal due to glutamate toxicity. Hum. Mol. Genet. 2014;24:3257–3271.
Menendez L., Yatskievych T.A., Antin P.B., Dalton S. Wnt signaling and a Smad pathway blockade direct the differentiation of human pluripotent stem cells to multipotent neural crest cells. Proc. Natl. Acad. Sci. U S A. 2011;108:19240–19245.
Mi H., Muruganujan A., Thomas P.D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2012;41:D377–D386.
Milani P., Escalante-Chong R., Shelley B.C., Patel-Murray N.L., Xin X., Adam M., Mandefro B., Sareen D., Svendsen C.N., Fraenkel E. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. Sci. Rep. 2016;6:25474.
Molero A.E., Gokhan S., Gonzalez S., Feig J.L., Alexandre L.C., Mehler M.F. Impairment of developmental stem cell-mediated striatal neurogenesis and pluripotency genes in a knock-in model of Huntington's disease. Proc. Natl. Acad. Sci. U S A. 2009;106:21900–21905.
Ooi J., Langley S.R., Xu X., Utami K.H., Sim B., Huang Y., Harmston N.P., Tay Y.L., Ziaei A., Zeng R. Unbiased profiling of isogenic huntington disease hPSC-derived CNS and peripheral cells reveals strong cell-type specificity of CAG length effects. Cell Rep. 2019;26:2494–2508.e7.
Pelegrí C., Duran-Vilaregut J., del Valle J., Crespo-Biel N., Ferrer I., Pallàs M., Camins A., Vilaplana J. Cell cycle activation in striatal neurons from Huntington's disease patients and rats treated with 3-nitropropionic acid. Int. J. Dev. Neurosci. 2008;26:665–671.
Phillips W., Morton A.J., Barker R.A. Abnormalities of neurogenesis in the R6/2 mouse model of Huntington's disease are attributable to the in vivo microenvironment. J. Neurosci. 2005;25:11564–11576.
Poon A., Zhang Y., Chandrasekaran A., Phanthong P., Schmid B., Nielsen T.T., Freude K.K. Modeling neurodegenerative diseases with patient-derived induced pluripotent cells: possibilities and challenges. Nat. Biotechnol. 2017;39:190–198.

Ring K.L., An M.C., Zhang N., O'Brien R.N., Ramos E.M., Gao F., Atwood R., Bailus B.J., Melov S., Mooney S.D. Genomic analysis reveals disruption of striatal neuronal development and therapeutic targets in human Huntington's disease neural stem cells. Stem Cell Reports. 2015;5:1023–1038.
Ross-Innes C.S., Stark R., Teschendorff A.E., Holmes K.A., Ali H.R., Dunning M.J., Brown G.D., Gojis O., Ellis I.O., Green A.R. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481:389–393.
Ruocco H.H., Lopes-Cendes I., Laurito T.L., Li L.M., Cendes F. Clinical presentation of juvenile Huntington disease. Arq. Neuropsiquiatr. 2006;64:5–9.
Ruzo A., Croft G.F., Metzger J.J., Galgoczi S., Gerber L.J., Pellegrini C., Wang H., Fenner M., Tse S., Marks A. Chromosomal instability during neurogenesis in Huntington's disease. Development. 2018;145
Seward M.E., Swanson E., Norambuena A., Reimann A., Cochran J.N., Li R., Roberson E.D., Bloom G.S. Amyloid-β signals through tau to drive ectopic neuronal cell cycle re-entry in Alzheimer's disease. J. Cell Sci. 2013;126:1278–1286.
Shi Y., Shu B., Yang R., Xu Y., Xing B., Liu J., Chen L., Qi S., Liu X., Wang P. Wnt and Notch signaling pathway involved in wound healing by targeting c-Myc and Hes1 separately. Stem Cell Res. Ther. 2015;6:120.
Tabrizi S.J., Ghosh R., Leavitt B.R. Huntingtin lowering strategies for disease modification in Huntington's disease. Neuron. 2019;101:801–819.
Tattersfield A.S., Croon R.J., Liu Y.W., Kells A.P., Faull R.L.M., Connor B. Neurogenesis in the striatum of the quinolinic acid lesion model of Huntington's disease. Neuroscience. 2004;127:319–332.
Telezhkin V., Schnell C., Yarova P., Yung S., Cope E., Hughes A., Thompson B.A., Sanders P., Geater C., Hancock J.M. Forced cell cycle exit and modulation of GABAA, CREB, and GSK3β signaling promote functional maturation of induced pluripotent stem cell-derived neurons. Am. J. Physiol. Cell Physiol. 2016;310:C520–C541.
The HD iPSC Consortium Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-expansion-associated phenotypes. Cell Stem Cell. 2012;11:264–278.
Thomas P.D., Campbell M.J., Kejariwal A., Mi H., Karlak B., Daverman R., Diemer K., Muruganujan A., Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129–2141.
Valor L.M. Transcription, epigenetics and ameliorative strategies in Huntington's disease: a genome-wide perspective. Mol. Neurobiol. 2015;51:406–423.
Wiatr K., Szlachcic W.J., Trzeciak M., Figlerowicz M., Figiel M. Huntington disease as a neurodevelopmental disorder and early signs of the disease in stem cells. Mol. Neurobiol. 2018;55:3351–3371.
Yeo H.C., Beh T.T., Quek J.J.L., Koh G., Chan K.K.K., Lee D.-Y. Integrated transcriptome and binding sites analysis implicates E2F in the regulation of self-renewal in human pluripotent stem cells. PLoS One. 2011;6:e27231.

Chapter 4- Omics Integration for Understanding Subtype-Specific Drug Response in Glioblastoma Cell Lines

Authors: Maxwell P. Gold and Obada T. Alhalabi

This section includes parts of a manuscript focused on understanding subtype-specific response to dasatinib in glioblastoma cell lines. The data was generated and processed by Obada Alhalabi and collaborators in the Goidts lab at the German Cancer Research Center (DKFZ). My primary contributions were analyzing the proteomics and phosphoproteomics data and developing the SamNet 2.0 algorithm to integrate the shRNA screen hits with the proteomics data. I have included my sections of the manuscript and provided brief summaries of other sections for context.

4.1 Abstract

Glioblastoma is a highly aggressive brain tumor with a five year survival rate of 5% (Stupp et al., 2010). There are three molecular subtypes of IDH-wt glioblastomas (classical, mesenchymal, and proneural) and single-cell RNA-sequencing (scRNA-seq) studies revealed that the subtypes contain many of the same types of cells, but at distinct proportions (Patel et al., 2014, Neftel et al., 2019). It was recently established the tyrosine kinase inhibitor dasatinib shows efficacy in mesenchymal-like cells, but not proneural-like cells. We used two experimental approaches to better understand this subtype-specific response, with the goal of identifying targets for a combination therapy that would be more effective against proneural-like cells. First, we collected protein and phosphoprotein data from proneural and mesenchymal cell lines treated with dasatinib and compared the results to data from untreated controls. We also collected shRNA screen data from treated and untreated cells to identify genes that affect dasatinib efficacy. These experiments led to many potential targets for combination therapies, so we developed a significantly improved version of the SamNet multi-commodity flow algorithm to integrate the data and help prioritize hits for validation. Together, these analyses highlighted cell cycle genes, and specifically WEE1, as a strong potential combination therapy target to affect proneural-like cells. This result was validated by showing that WEE1 inhibition has synergistic effects with dasatinib in proneural cell lines, but not mesenchymal ones.

4.2 Introduction

Glioblastoma (GBM) is the most frequent malignancy of the nervous system (Louis et al., 2016). Even after aggressive therapies like surgical resection and chemo-radiotherapy, the patient prognosis is dire, with a median survival of 12 to 15 months (Stupp et al., 2005). Genomic, transcriptomic, methylomic, and proteomic studies have revealed molecular heterogeneity between GBM samples (Brennan et al., 2013, Patel et al., 2014, Sturm et al., 2012; Wang et al., 2018, Verhaak et al., 2010) and the research community recognizes three consensus subtypes of GBM: mesenchymal (MES),

proneural (PN) and classical (CL). Additionally, single-cell and lineage-tracing studies highlight GBM heterogeneity by showing spatial and temporal diversity of subtypes within the same patient (Klughammer et al., 2018; Neftel et al., 2019; Patel et al., 2014). This diversity of cells has been implicated in GBM resistance to chemo-radiotherapy because even if treatments are effective against one cell type, other resistant populations can come dominate the tumor (Bao et al., 2006; Campos et al., 2014; Chen et al., 2012).

In a previous study, we showed that the MES subtype was more sensitive to tyrosine-kinase inhibitor dasatinib than the PN subtype in subtype-specific models of GBM stem cells (GSCs) (Alhalabi et al., 2021). Even though dasatinib may eliminate MES GSCs, it would be better to have a combination therapy capable of eradicating both MES and PN GSCs. In this work, we used two experimental approaches to study dasatinib response and identify potential combination therapies to improve efficacy against PN cells. First, we collected protein and phosphoprotein expression data from treated and untreated cells for both MES and PN models to better understand the molecular response to dasatinib. In the second experiment, we used a pooled shRNA screen, which is commonly used to study tumor vulnerability in other cancers (D'Alesio et al., 2016, Khorashad et al., 2015, Schramm et al., 2019). For these experiments, shRNAs are used to reduce the expression of specific genes and determine the effect on cell viability. In our case, we performed the experiment on dasatinib-treated and untreated cells from both subtypes to identify genes that affect dasatinib efficacy.

We also sought to computationally integrate the data from both assays to prioritize hits and get a more comprehensive understanding of subtype-specific dasatinib response. Network flow algorithms like ResponseNet (Yeger-Lotem et al., 2009) and SamNet (Gosline et al., 2012) have been previously used to integrate data from multiple biological assays. These methods work by creating a directed network with hits from one assay (e.g. proteomics) at the top and from another assay (e.g. shRNA screen) at the bottom. These hits are represented by nodes in a network and are connected through a protein-protein interaction network where edges represent physical interactions. Flow travels from the top of the network, through some select proteins from the PPI, and then down to the bottom, resulting in a subnetwork of particularly important proteins that connect the two datasets. These methods have led to biological insights in Parksinon’s Disease and non-small cell lung cancer (NSCLC) but are limited because they do not adjust for high-degree hub nodes and cannot handle data from more than two assays.

In this work, we developed the SamNet 2.0 algorithm to address these flaws and integrate the proteomics response data with the shRNA screen hits. Analysis of the individual assays and integrated network highlighted cell cycle genes, and specifically WEE1, as an attractive target for combination therapy in PN cells. This hypothesis was validated in cell line models, suggesting that WEE1 inhibition + Dasatinib could be an effective combination for eliminating MES and PN cells in GBM tumors.

4.3 Results

4.3.1 **Mesenchymal and Proneural cells have distinct proteomic response to dasatinib**

We compared the proteomic response to dasatinib between the mesenchymal and proneural cell lines by analyzing the 1617 features that were detected in all samples and that map uniquely to a single protein. Dasatinib leads to the significant dysregulation of 570 proteins in mesenchymal cells (418 down, 152 up) and 319 proteins in proneural cells (244 down and 75 up) (one-sample t-test adj-p < 0.01). There are 245 proteins significantly downregulated in mesenchymal cell lines, but not in proneural cell lines; using *TopGO* (Alexa and Rahnenführer, 2007), we found that these proteins are enriched in many gene sets related to translation (e.g. translational initiation, p = 6.7E-3) and the cell cycle (e.g. G1/S transition of mitotic cell cycle, p = 6.9E-5). The 71 genes significantly downregulated only in the proneural cells are enriched for categories related to the nervous system (e.g. Neuron Death, p = 0.015).

We then directly compared the subtypes, identifying 197 proteins with a significantly different response to dasatinib (adj-p <0.01). The 80 genes whose mean response to dasatinib is lower in the proneural cell lines are enriched for gene sets related to metabolism, such as carbohydrate catabolic process (p=0.002). The 117 proteins lower in mesenchymal cells are enriched for gene sets related to DNA replication initiation (p = 2.5E-5) and complex assembly involved with cell cycle DNA replication (p = 1.3E-7). This is consistent with the previous analysis and suggests that proteins necessary for the cell cycle and DNA replication are being downregulated more substantially in mesenchymal cells than proneural cells.

4.3.2 **Subtype specific phosphoproteomic response to dasatinib**

We performed a similar analysis for the phosphoproteomic data on 1153 unique phosphopeptides. When comparing the mesenchymal and proneural cells, we found 201 phosphosites that differ significantly in their response to dasatinib (adj-p < 0.05). We performed hypergeometric tests using the PhosphositePlus database (Hornbeck et al., 2015) to determine if these differential phosphosites are enriched for targets of particular kinases; Protein Kinase A targets (adj-p = 0.039) and Protein Kinase C targets (adj-p = 0.039) are both significantly overrepresented in the differential phosphosites, suggesting these kinases may have subtype-specific responses to dasatinib. Figure 1A highlights the differential phosphosites that are targets of PKA based on PhosphositePlus or Scansite (Obenauer et al., 2003), another database of computationally-derived kinase-phosphate targets. 15 of the 201 differential phosphosites are phosphorylated by PKA and 10 of those 15 show a lower response in the mesenchymal cells.

We also performed PTM-SEA (Krug et al., 2019) , a modification of Gene Set Enrichment Analysis (Subramanian et al., 2005), which utilizes all the collected phosphoproteomics data to identify relevant sets of phosphosites. We first implemented PTM-SEA for the proneural and mesenchymal cells separately; for the mesenchymal

cell lines, Casein Kinase 2A1 targets are significantly dysregulated (adj-p = 0.030), while there are no significant results for the proneural cells. We then re-ran PTM-SEA using the difference in response to dasatinib between the mesenchymal and proneural cell lines and CDK7 targets are the top hit. This phosphosite set does not reach statistical significance (adj-p= 0.059), but the trend supports the proteomic evidence that cell cycle processes may distinguish the mesenchymal and proneural response to dasatinib.

Given the PTM-SEA and proteomics results, we further investigated the phosphosites on genes related to the cell cycle (Figure 1B). CDK1-T161, the activating phosphosite on CDK1 (Li et al., 2016), and RB1-T373, the site on RB1 whose phosphorylation releases E2F and promotes the cell cycle(Gorges et al., 2008), are both significantly downregulated by dasatinib in the mesenchymal group but not the proneural cells . Additionally, multiple phosphosites on MCM2 are significantly downregulated compared to baseline and lower in mesenchymal cells than proneural ones (Stead et al., 2012). Even though these data suggest downregulation of the cell cycle in mesenchymal cells, we also observe significant upregulation of the activating phosphosite (T170) on key cell cycle protein CDK7 in mesenchymal cells (Martinez et al., 1997). Overall, these data indicate that key cell cycle phosphosites are affected more substantially in the dasatinib-sensitive mesenchymal cells than the dasatinib-resistant proneural ones.

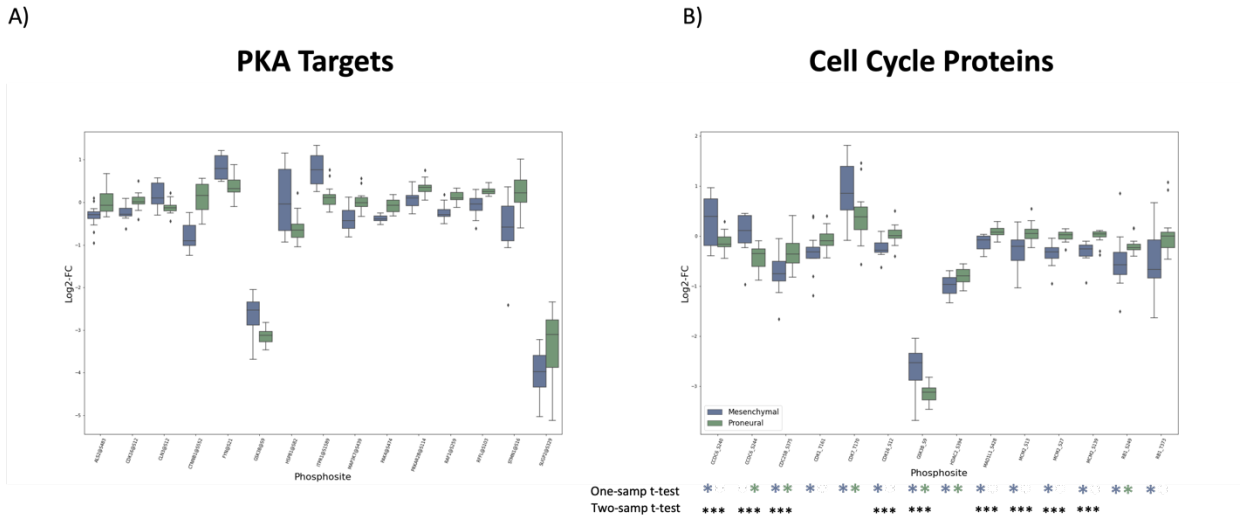
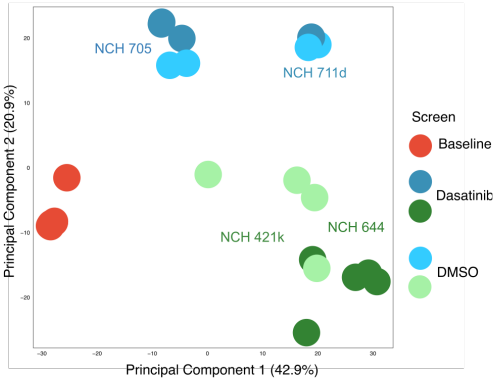


Figure 1: Phosphosite analysis of PKA and cell cycle proteins.
A) Log2 fold-change in response to dasatinib for targets of Protein Kinase A. Blue indicates mesenchymal cell lines and green labels proneural cell lines.
B) Log2 fold-change in response to dasatinib for phosphosites on genes related to the cell cycle. Blue stars below chart indicate whether the log2 fold change is significantly different from 0 for either subtype. The three black stars indicate that the response is significantly different between mesenchymal and proneural cell lines.

4.3.3 **Brief Summary of shRNA Screen Results from DKFZ**

After analyzing the molecular response to dasatinib in each subtype, we employed a pooled shRNA screen to understand what genes affect dasatinib response. This method uses a barcoded library of 27500 shRNAs against 5000 genes and relies on the transduction of shRNAs cloned into a vector that includes TagRFP for monitoring. The transduced cells were randomized to Dasatinib (or DMSO) treatment at an the IC_{20} concentration. Additionally, a set of all cells were harvested at 18 hours after transduction and used as an untreated baseline. Principal Component Analyses (PCAs) of the screen data revealed separation between baseline, PN, and MES GSCs (Figure 2A). Since the MES cells are already sensitive to dasatinib, we focused our analysis on PN cells by comparing the viability in the dasatinib condition to the DMSO-treated condition. Figure 2B summarizes the strongest hits from the PN cells.

A)



B)

Gene	p-Value	MES logCPM foldchange			PN logCPM foldchange			MES vs PN
		Dasatinib to Baseline	DMSO to Baseline	Dasatinib to DMSO	Dasatinib to Baseline	DMSO to Baseline	Dasatinib to DMSO	
WEE1	5.9E-05	-0.707	-0.617	-0.09	-1.595	-0.367	-1.228	-1.138
CREM	0.00256	-0.774	-0.94	0.166	-1.744	-0.634	-1.11	-1.276
RHOA	0.00455	-0.484	-0.438	-0.045	-1.961	-1.08	-0.881	-0.836
ITGA3	0.00627	-0.166	0.022	-0.188	-1.357	-0.471	-0.885	-0.697
MBD1	0.0066	0.065	-0.069	0.133	-0.913	-0.13	-0.783	-0.916
WNT8A	0.00678	-1.26	-1.229	-0.031	-2.714	-1.404	-1.31	-1.279
SNCB	0.008	-0.417	-0.54	0.123	-1.999	-1.178	-0.821	-0.944
BPNT1	0.00879	-0.398	-0.498	0.1	-1.148	-0.36	-0.789	-0.889

Figure 2: shRNA screen summary.
A) shRNA screen Principal Component Analysis. PCA of shRNA screen results shows distinct clustering between baseline, DMSO, and Dasatinib conditions. Blue circles are for MES cell lines, while green are for PN cell lines.
B) Top screen hits for PN cells. Chart shows the logCPM values comparing dasatinib, DMSO, and baseline conditions. These genes are the top p-values when comparing the Dasatinib and DMSO conditions in the PN subtype (see Methods).

4.3.4 **SamNet 2.0 For Integrating Omics Data Using Multi-Commodity Flow Algorithms**

The omics and shRNA data provided new information about how dasatinib affects proneural and mesenchymal glioblastoma cell lines. The proteomics and phosphoproteomics revealed features whose expression were significantly altered by dasatinib, while the shRNA screen highlighted genes that can be targeted to potentially enhance the effectiveness of dasatinib. We sought to integrate these data to prioritize shRNA screen hits for validation and to better understand how the proteomic effects of

dasatinib could lead to shRNA screen hits. To do this, we developed and implemented SamNet 2.0, a novel algorithm for data integration.

SamNet 2.0 builds on established network flow methods, such as ResponseNet (Yeger-Lotem et al., 2009) and SamNet (Gosline et al., 2012), which have been used to integrate distinct biological data types. ResponseNet (Figure 3A) works by creating a directed graph that connects hits from one assay (e.g. proteomics) to hits from another experiment (e.g. shRNA screen), through an established biological network like a protein-protein interaction (PPI) network. Each edge in the directed network is assigned a specific capacity and penalty based on the experimental data or prior knowledge; these values are chosen to encourage flow through strong experimental hits and high-confidence protein-protein interactions. After the setup, a minimum-cost maximum-flow problem is solved, where flow travels through a small number of nodes revealing an optimal network of important features that connect the two experiments. By using the PPI as prior knowledge, this final network can include key proteins that were not measured by either experiment.

SamNet (Figure 3B) is a multi-commodity flow algorithm that was designed to run ResponseNet for multiple conditions (e.g. proneural cells and mesenchymal cells) in the same optimization. Each condition has specific hits at the top and bottom of the network, but the flow through the PPI must be shared among all conditions; Gosline *et al.* (2012) found that this setup reveals more subtype-specific features than implementing ResponseNet individually for each condition. For this work, we designed and implemented SamNet 2.0 (Figure 3C), which significantly improves upon SamNet in two ways. First, SamNet 2.0 penalizes high-degree hub nodes to prevent highly connected proteins, like ubiquitin, from dominating the solution. Second, SamNet 2.0 incorporates baseline expression data so that the optimal network is more likely to contain proteins that are present in a condition of interest (see Methods).

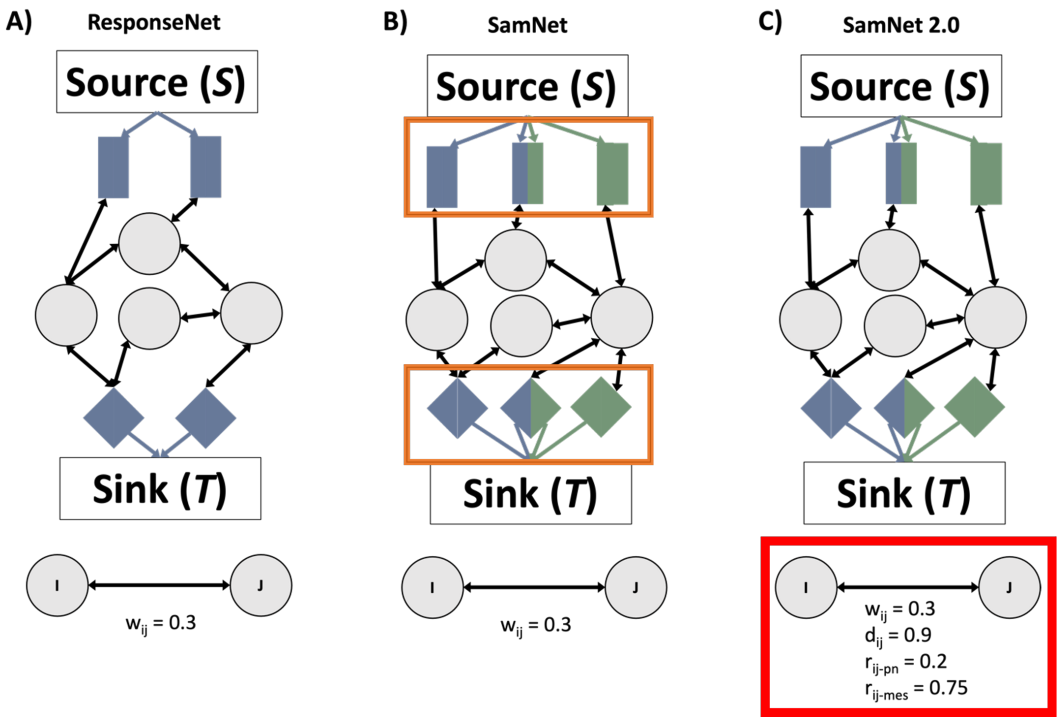


Figure 3: Summary of network flow algorithms.
A) ResponseNet Setup. A source node (*S*) is connected through directed edges to hits from one assay (e.g. differential proteins) from one condition (e.g. mesenchymal glioblastoma cells). These hits (represented by rectangles) are connected to non-hit proteins (represented by gray circles) through edges that represent physical interactions. The network then ends with hits from another assay (e.g. shRNA screen hits) represented by diamonds which have outgoing edges into a sink node (*T*). Each edge in the network has an upper limit of flow and a single penalty (w_{ij}). The w_{ij} penalty is assigned such that stronger hits and more confident edges have lower penalties (see Methods)
B) SamNet Setup: SamNet is setup the same as ResponseNet, except now there are multiple conditions represented by different colors (highlighted by the yellow box). The hits have condition-specific edges connecting the nodes to the source (*S*) or the sink (*T*), but all the internal PPI edges are shared among all conditions. Each edge still only has one penalty w_{ij}, which follows the same scheme as ResponseNet.
C) SamNet 2.0 Setup: SamNet 2.0 has the same setup and condition-specific edges as SamNet, but now the internal edges have additional penalties to account for node degree and expression (highlighted by the red box). The w_{ij} penalty is the same as for ResponseNet and SamNet. The D_{ij} penalty is based on the degree of the nodes it connects, whereby edges connecting nodes with higher degrees have larger penalties. The r_{ijk} penalty is reflects the baseline expression of nodes and there is a specific penalty for each condition. Thus, an edge that connects genes I and J, which are highly expressed in proneural cells, will have a low proneural penalty. If genes I and J have low expression in the mesenchymal cells, the mesenchymal specific penalty ($r_{ij\text{-}mes}$) will be high.

4.3.5 SamNet 2.0 reveals cell cycle genes as candidate for combination therapy in proneural tumors

We implemented SamNet 2.0 to integrate the proteomics and shRNA screen data collected from mesenchymal and proneural cell lines. To identify subtype-specific pathways, we treated the mesenchymal and proneural cells as separate commodities. We set up the directed graph (Figure 3C) by connecting the source node (*S*) to 145 proteins significantly affected by dasatinib (adj-p < 0.01 and abs($\log_2$FC) > 1). This

includes 31 mesenchymal-specific proteins, 23 proneural-specific proteins and 91 proteins represented in both subtypes; even though some hits are present in both conditions, they were assigned subtype-specific capacities and penalties based on the response to dasatinib in the given condition (see Methods). We connected the terminal sink node (*T*) to the 44 shRNA hits (16 mesenchymal-specific, 28 proneural-specific) that cause a significantly different effect on survival in the presence of dasatinib compared to DMSO ($p < 0.05$, and $abs(\log_2 FC) > 1$ compared to baseline). The proteomics and shRNA hits were then connected through a protein-protein interaction network. This graph setup was designed to understand how proteins affected by dasatinib may relate to the shRNA hits that have dasatinib-specific effects; following the flow from the dasatinib-affected proteins, through the PPI and to the shRNA hits, can help yield hypotheses for how proteins and pathways affected by dasatinib may cause an shRNA hit to enhance the effectiveness of dasatinib.

We used Gurobi (Gurobi Optimization, 2020) to solve the optimization outlined above (see Methods Equation 3). The algorithm has multiple parameters, so we performed a grid search to identify high quality networks (29 out of 288 possible networks). We then selected one optimal network that has a substantial number of proteomic and shRNA hits, limited flow through high-degree hub nodes, like ubiquitin, and genes with high subtype-specific RNA expression (see Methods).

We performed gene ontology analysis with *TopGO* (Alexa and Rahnenführer, 2007) to identify pathways enriched in this optimal network. For the enrichment, we only considered proteins that were in the optimal network and were not one of the proteomic or shRNA hits; this was done to identify gene sets related to the integration and not simply the omics experiments. The proteins that only contain flow for the mesenchymal condition are enriched for pathways related to DNA damage, such as “nucleotide excision repair, DNA Damage Recognition” (p=3.0E-6) and “apoptotic DNA fragmentation” (p = 1.0E-4). Many of these proteins help connect the differential proteins to the shRNA hits of DDB2 and DFFB (Figure 4A), suggesting that affecting DNA damage repair could potentially make dasatinib more effective in mesenchymal cells.

For the proneural-specific proteins, there are multiple top pathways related to the nervous system, such as oligodendrocyte development (p=0.0005) and axonogenesis (p=0.002). Additionally, many significant ontologies are related to the cell cycle, such as “G1/S transition of mitotic cell cycle” (p=0.003) and “regulation of cell cycle” (p=0.02). We then assessed all 29 high quality networks to determine whether the enrichment of cell cycle genes was robust; we consistently observed enrichment for cell cycle related pathways, such as “regulation of cell cycle”, which was significantly enriched ($p<0.05$) for proneural-specific proteins in 27/29 networks. In the optimal network, many of the cell cycle nodes connect the differential proteins to the shRNA hits of WEE1 and KIF11, which both play key roles in cell cycle (McGowan and Russell, 1995; Wojcik et al., 2013) (Figure 4B). We also observe many interactions related to cell cycle regulation, such as 14-3-3 proteins (like YWHAB) promoting WEE1 stability (Rothblum-Oviatt et al., 2001) and PIN1 inactivating WEE1 (Okamoto and Sagata, 2007).

A)

B)

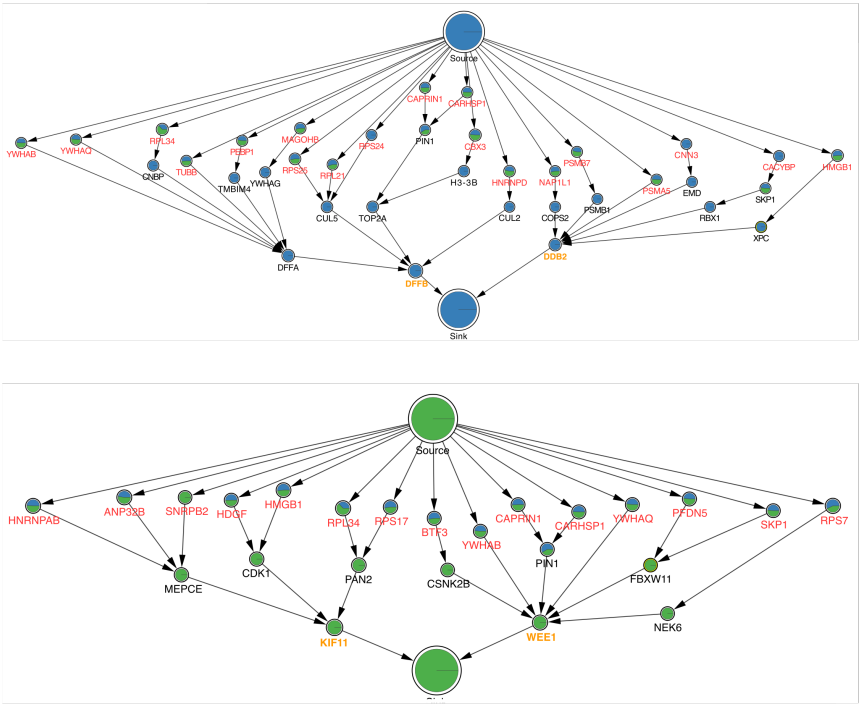


Figure 4: SamNet Networks for Proneural and Mesenchymal cases
A) Optimal Network Highlights DNA Damage Proteins for Mesenchymal Subtype: This is a subnetwork of the optimal network (gamma-20, epsilon-25, rho-3, capping) that highlights the enrichment of proteins with mesenchymal flow that are related to DNA Damage processes. Each node is represented by a pie chart, where the blue section is proportional to the amount of mesenchymal flow and the green section represents the proportion of proneural flow. The size of the pie chart is related to the amount of total flow that travels through the node. Label colors represent the type of node: red is proteomic hit, yellow is an shRNA hit, and black is a non-hit protein from the PPI.
B) Optimal network Highlights Cell Cycle Proteins for Proneural Subtype: the nodes and labels are the same as part B. This highlights the enrichment of cell cycle related proteins that connect differential proteins to proneural shRNA hits.

4.3.6 **WEE1 inhibition produces synergistic effect with dasatinib in PN cell lines**

Proteomic, shRNA screen, and integrative analyses highlighted the relevance of cell cycle genes in dasatinib response. Specifically, WEE1 shows up as a strong target for a combination therapy targeting PN cells. To validate this hypothesis, we performed a cell viability assay using dasatinib combined with Mk1775, a small molecular inhibitor of WEE1, on PN and MES cells. We found that PN cells showed a higher synergistic effect (δ=49 for NCH644, δ =19 for NCH421k, Figure 5A) compared to the MES cells (δ =5 for NCH705, δ =10 for NCH711d, Figure 5B). We are currently performing validation experiments to assess whether these differences in viability are from lower proliferation, higher cell death, or both.

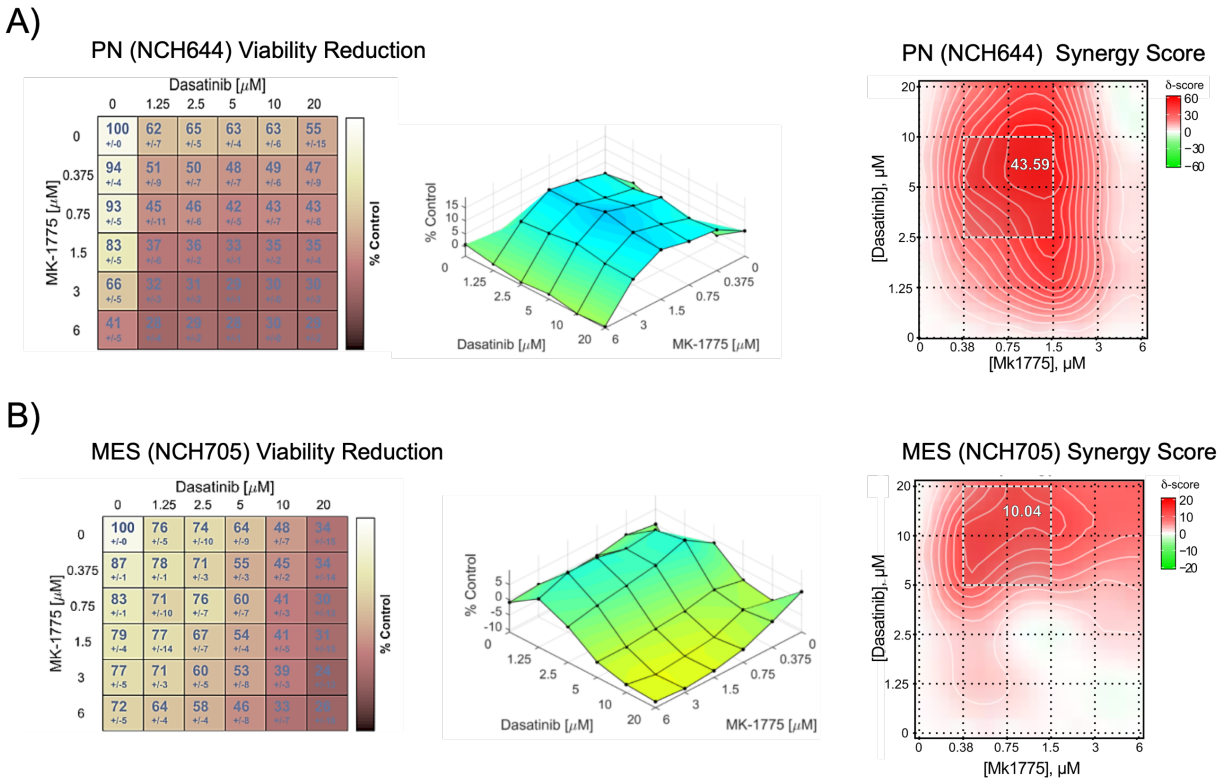


Figure 5: Wee1 Inhibitor Shows Subtype-Specific Efficacy
A) Dasatinib and Mk1775 effects in PN cell line NCH644: Left panel shows percent of cell viability compared to untreated control for each combination of dasatinib and Mk1775. Middle panel shows 3D plot of data in left panel. Right plot shows synergy scores calculated by SynergyFinder for each condition. The synergy scores are much higher in PN cell line NCH644 than MES cell line NCH705.
B) Dasatinib and Mk1775 effects in MES cell line NCH705: Same panels as A, except for cell line NCH705. Dasatinib shows clear effects on viability that are not dramatically affected by Mk1775.

4.4 Conclusion

This work presents progress in our understanding the subtype-specific responses to dasatinib in glioblastoma. Our data suggests that the combination of dasatinib and WEE1 inhibition could be an effective combination therapy for treating MES-like and PN-like cells. Further validation should be performed in mouse models to test this hypothesis. This work also details significant advancements in the utilization of network flow algorithms for omics data integration. These improvements to the SamNet algorithm produce more realistic outputs by reducing flow through hub nodes and allowing for the incorporation of baseline expression data.

This work is the product of a collaboration with the Goidts lab the German Cancer Research Center (DKFZ). Obada Alhalabi has led the project and Mona Gottmann did

significant work for the shRNA screen and validations. My primary contributions were analyzing the data and developing the improved SamNet algorithm. This project is still ongoing as Obada and Mona are working to better understand the effects of the dasatinib/Mk1775 combination.

4.5 Methods

4.5.1 **Differential Analysis**

The proteomics and phosphoproteomics data were filtered down to features present in all samples. Additionally, for the proteomics data, features were only included if they mapped to a single protein. Ensembl BioMart (Kinsella et al., 2011) was used to account for gene synonyms output by MaxQuant. In the case where a feature mapped to two proteins, but the first protein product is a part of the second protein product as well (e.g. ARPC4 and the ARPC4-TTLL3 readthrough protein), this was considered as uniquely mapping to the first protein. For the phosphoproteomics data, phosphosites were named according to their amino acid motif (+/- 7 AA from the phosphosite); multiple instances containing the same AA motif were collapsed to a single value by taking the mean.

Limma (Ritchie et al., 2015) was used for all differential analysis on the log2-normalized quantification values for the proteins and phosphosites. The Benjamini-Hochberg method (Benjamini and Hochberg, 1995) was used to adjust for multiple hypotheses.

4.5.2 **Gene Ontology Enrichment**

The *topGO* package (Alexa and Rahnenführer, 2007) was implemented for gene ontology enrichment using the fisher statistic, elimination algorithm, and hgu95av2.db mapping. For the proteomics data, differential proteins were compared to a background of all identified proteins. For SamNet 2.0 analysis, the significant proteins for a given subtype were ones that only contained flow for that subtype and were not included in the proteomics or shRNA hits. This filtering was done to identify gene sets related to the integration and not simply the omics assays. These proteins were compared to a background of all proteins in the PPI, excluding the proteomics and shRNA hits.

4.5.3 **Figure Creation**

Boxplots in Figure 1 were generated using the *seaborn* python package (version 0.9.0) (Waskom, 2021). Networks were generated by the *networkx* python package (v.2.4) (Hagberg et al., 2008) and network images were created using cytoscape (v3.7.2) (Shannon et al., 2003).

4.5.4 **Phosphosite Set Analysis**

PTM-SEA (Krug et al., 2019) was run on the results from the limma analysis of the phosphosite data, using the amino acid motifs (+/- 7 amino acids from the phosphosite) and the flanking sequence human dataset (v.1.9.0) from PTMSigDB (Krug et al., 2019). There were three values tested: 1) the change in response in mesenchymal cells (i.e. limma t-value from one-sample t-test of only mesenchymal samples), 2) the change in response in proneural cells (i.e. limma t-value from one-sample t-test of only proneural samples), and 3) the difference in response between mesenchymal and proneural cells (limma t-value from t-test between mesenchymal vs. proneural). The t-statistic was used instead of a p-value because PTM-SEA considers directionality of trends. PTM-SEA was run with the standard parameters, except the minimum set overlap was reduced from 10 to 5 because of the limited number of phosphosite-sets being considered with the standard implementation.

Hypergeometric tests were performed using the same flanking sequence PTM-SigDB database to identify potential upstream kinases. The input was phosphosites with significantly differential response to dasatinib (adj-p < 0.05) between the mesenchymal and proneural cells. These phosphosites were compared against the kinase-target sets in PhosphoSitePlus (Hornbeck et al., 2015); like PTM-SEA, phosphosite-sets were only considered if they contained 5 phosphosites from the whole dataset of 1153 detected phosphosites.

When compiling the Scansite phosphosites for PKA (Figure 1), we submitted motifs (+/- 7 AA) to Scansite (Obenauer et al., 2003) and only considered upstream kinase matches that meet “High” stringency.

4.5.5 **Multi-Commodity Flow Optimization with SamNet 2.0**

SamNet 2.0 is a significant improvement compared to the original ResponseNet (Yeger-Lotem et al., 2009)and SamNet (Gosline et al., 2012) algorithms. As stated above, these methods are based on setting up a directed graph where a source node (*S*) connects to hits from one biological experiment (e.g. differential proteins) and a sink node (*T*) receives inputs from hits from another type of experiment (e.g. shRNA screen hits). The hits from these experiments are then connected through an established biological network, such as a protein-protein interaction network, where bidirectional edges (allowing flow either way) connect proteins known to physically interact.

In this graph, every edge has a capacity (upper limit of flow) and a penalty weight. Edges that connect experimental hits to the source or sink node have a capacity and penalty related to the strength of the hit; stronger hits get a larger capacity and lower penalty. For the biological network, every edge has a constant capacity (e.g. 1) and a penalty based on the amount of evidence supporting the biological interaction, where high-confidence edges receive a lower penalty. With the graph established, linear

programming is used to solve the optimization in equation 1. The Gamma (γ) term is used to control the flow through the network, where higher values increase flow through network, counteracting the w_{ij} penalty term.

SamNet was originally developed to run ResponseNet for multiple related biological conditions using the same optimization. Each condition (e.g. mesenchymal cell lines vs. proneural cell lines) is treated as commodity and has subtype-specific edges from the source and sink that connect to condition-specific hits. Still, all commodities must share the internal edges of the PPI network, which helps reveal subtype-specific features. SamNet is solved using equation 2, which is the same formulation as ResponseNet except that it accounts for all *k* conditions.

Two substantial improvements to SamNet were implemented for this work. The first change addressed the fact that a disproportionate amount of flow in SamNet solutions goes through hub nodes, or nodes with an exceptionally high degree in the protein-protein interaction network. We implemented our parameter sweep again with no hub penalty (Epsilon = 0) and found that for these networks, an average of 33% of the network flow goes through hub protein (top 0.1% of nodes by degree). To address this, a penalty term (d_{ij}) was added to the optimization (Equation 3) where each edge in the PPI is penalized based on the mean degree of the two nodes it connects. d_{ij} is higher for edges that connect high degree nodes and this penalty is the same across all commodities. The strength of this penalty compared to other features is controlled by the Epsilon (ε) parameter.

The second improvement integrates expression data into the optimization. The original SamNet cannot account for baseline expression and thus may yield networks that including proteins not present in each biological condition. This was addressed by adding another penalty term (r_{ijk}) to the optimization. Each edge is assigned an r_{ijk} value based on the mean expression of the nodes it connects in the condition of interest; edges connecting nodes with higher expression have a lower penalty, favoring their inclusion in the optimal network. The strength of this term is controlled by the Rho (ρ) parameter (equation 3).

Equation 1: $\min_f \sum_{i \in V,\, j \in V} w_{ij} \times f_{ij} - \sum_{i \in source} \gamma \times f_{Si}$

Equation 2: $\min_f \sum_{k \in C} \sum_{i \in V,\, j \in V} w_{ij} \times f_{ij} - \sum_{k \in C} \sum_{i \in source} \gamma \times f_{Si}$

Equation 3: $\min_f \sum_{k \in C} \sum_{i \in V,\, j \in V} w_{ij} \times f_{ij} + \sum_{k \in C} \sum_{i \in V,\, j \in V} \varepsilon \times d_{ij} \times f_{ij} + \sum_{k \in C} \sum_{i \in V,\, j \in V} \rho \times r_{ijk} \times f_{ij} - \sum_{k \in C} \sum_{i \in source} \gamma \times f_{Si}$

4.5.6 **Implementation of SamNet 2.0**

SamNet 2.0 was implemented to gain a better understanding of the subtype-specific effects of dasatinib on glioblastoma cell lines. The SamNet 2.0 graph was designed such that proteomic hits received inputs from the source and shRNA hits had outgoing edges into the terminal sink node. Proteomics hits for each subtype were proteins

significantly affected by dasatinib (i.e. one-sample t-test was significantly different from 0), with an absolute mean log_2FC greater than 1. The shRNA hits were selected on two criteria: 1) knockdown of this gene caused significantly different ($p<0.01$) survival effects between the dasatinib-treated cells and the DMSO treated cells 2) the log_2FC in survival between the dasatinib-treated cells and the untreated cells is more extreme than 1 in the same direction as the DMSO comparison.

The edge capacities and penalties were assigned based on the ResponseNet and SamNet procedures (Gosline et al., 2012; Yeger-Lotem et al., 2009). For the proteomic hits, the capacities of the incoming edges were calculated based on the magnitude of the protein's fold change in response to dasatinib and scaled so that the capacities sum to one for the hits in each commodity: $\sum_{k \in C} \frac{F_{ik}}{\sum_{i \in hits} F_{ik}}$ where F_{ik} is the magnitude of the mean dasatinib response fold change for protein i in subtype k. Each edge was then assigned a penalty (w_{Sik}) of $-log_2(capacity_{ik})$; thus, stronger hits with higher absolute fold changes had higher capacities and smaller penalties. The same scheme was implemented for the shRNA hits and their edges into the sink node (*T*). In this case, the capacities and penalties were based on the magnitude of the fold change in survival between the baseline control cells and the dasatinib treated cells.

The proteomic and shRNA hits were connected through the IREF Index v14 (Razick et al., 2008) protein-protein interaction network, which has been utilized by other network approaches, like OmicsIntegrator (Tuncbag et al., 2016). Protein names for the interactome were mapped to gene names using using Biomart (Kinsella et al., 2011) to be consistent with the proteomic hits (see differential analysis methods). All internal edges of the PPI were bidirectional and assigned a constant capacity of 1. The penalty (w_{ij}) was $-log_2(ev_{ij})$ where ev_{ij} is the evidence score for the interaction between nodes i and j provided by IREF Index.

For SamNet 2.0, each edge of the PPI also was assigned a degree penalty (d_{ij}) and a subtype-specific expression penalty (r_{ijk}). First, the degree for each node in the PPI was calculated and then transformed to be between 0.01 and 0.99 using a modified min-max scaling:
$1 - (0.01 + (1 - 2 \times 0.01) \times \frac{x_i - \min(x)}{\max(x) - \min(x)})$; nodes with higher degrees received values closer to 0.01. Then for each edge, d_{ij} was calculated as $-log_2(mdeg_{ij})$ where $mdeg_{ij}$ is the mean of the transformed degrees for nodes i and j. This procedure was chosen to amplify the penalty for flow through outlier hub nodes; some other normalizations (such as rank normalization) would assign outlier hub nodes, like ubiquitin, very similar penalties to non-outlier nodes with a high degree.

The subtype-specific expression penalty (r_{ijk}) was calculated using microarray mRNA data collected from the cell types. For an edge between nodes i and j, the penalty for flow for subtype k (r_{ijk}) was $-log_2(mexp_{ijk})$ where $mexp_{ijk}$ is the mean of the rank-normalized expression values for genes i and j for condition k. Thus, edges connecting genes with lower expression ranks received a higher penalty. In cases where expression data was not present for one of the two genes, $mexp_{ijk}$ was assigned the rank-normalized expression of the present gene. In the limited cases where no

expression data was collected for either gene, we assumed the genes ranked in the middle of expression and set $mexp_{ijk} = 0.5$.

4.5.7 **Parameter Selection for SamNet 2.0**

After the implementation of SamNet 2.0, a grid search was performed over the three parameters: Gamma (15,20,25,30,35,40), Epsilon (25,50,75,100), and Rho (0,3,5,7,10,20,30). Additionally, each run was completed with and without capping, a procedure that sets an upper limit to the edge confidence from the IREF PPI. Lan *et al* (Lan et al., 2011) found that ResponseNet can sometimes lead to very long strings of high-confidence protein interactions because the w_{ij} edge penalty is so low, which can be solved by capping the edge confidence scores. In this SamNet 2.0 implementation, runs with capping set a maximum edge confidence of 0.8.

High-quality networks were selected based on the following criteria:

1. The network includes a high percentage (>70%) of both proteomic hits and shRNA hits
2. Less than 5% of the network flow goes through hub nodes (defined as top 0.1% of nodes based on degree in the interactome)
3. Less than 1% of the edges in the network are low confidence (confidence score <0.3)
4. The RNA expression levels of the included proteins are significantly higher (Mann-Whitney U test adj-p <0.01) than the same parameter set with Rho set to 0.

These criteria resulted in 29 high quality networks out of 288 possible networks (all parameter sets besides Rho = 0 which were used for comparison in criterion 4). The optimal network selected has capping and Gamma of 20, Epsilon of 25, and Rho of 3. This solution was chosen based on principles from ResponseNet (Yeger-Lotem et al., 2009) which favored lower gamma values and a small number of low-confidence edges. This optimal network was tied for the lowest gamma in any high-quality network (i.e. 20) and was the only high-quality network that included zero low-confidence edges.

4.5.8 **Virus production**

HEK 293T cells were seeded into 15 cm plates. After 24 hours, transfection with the packaging plasmid mix and plasmids of the decipher shRNA library module 1(figure 7) was carried out according to Cellecta's® user manual: Packaging, Titering, and Transduction of Lentiviral Constructs. At 24 hours post-transfection, the medium was changed to fresh D-MEM medium supplemented with 10% FCS, DNase I (1 U/ml), MgCl2 (5 mM), and 20mM HEPES, pH 7.4. 48 hours post-transfection, the lentiviral supernatant was collected and filtered through a Nalgene 0.2 µm PES filter and concentrated with Cellecta's LentiFuge™ Viral Concentration Reagent according to the user manual provided by Cellecta. The only deviation was that after incubation with

LentiFuge™, samples were centrifuged for 15000 rpm for 75 mins (rather than 1 hour). After concentration and centrifuging, virus pellets were resuspended in PBS and aliquoted.

4.5.9 **Lentiviral Titer Calculation**

NCH 705 cells were transduced with the lentivirus and medium changed after 24 hours. After 72 hours TagRFP-positivity was measured on the measured at 72 hours with the LSRFortessa™ flow cytometer (BD biosciences). MOI (Multiplicity of infection) was determined using the conversion table provided by CELLECTA®. Lentiviral titer calculation was carried out using the following formula: TU/ml = (*Number of cells at Transduction*) × [MOI / (ml of *Lentiviral Stock* used at Transduction)].

4.5.10 **Viability screen**

GSCs (NCH 711d, NCH 705, NCH 421k and NCH 644) were transduced with an aim MOI (multiplicity of infection) of 0.5 with the lentiviral constructs containing module 1 of the decipher library from Cellecta® with approximately 27500 pooled barcoded shRNAs against more than 5000 genes. Half of the transduced GSCs were harvested at 18 hours post-transduction as a baseline sample. 24 hours after transduction the medium was changed. 72 hours after transduction flow cytometry was used to verify correct transduction percentage and confirm desired MOI. Transduced GSCS were then split into two arms: one was treated with the respective IC_{20} concentration of dasatinib, the other with the corresponding DMSO dilution. GCSs were further cultured until reaching 8-10 doublings by splitting every 4 days with dasatinib/DMSO reapplied to the respective and the TagRFP-positivity monitored. After reaching the desired doubling number final pellets of the dasatinib and DMSO arms were harvested for sequencing.

4.5.11 **gDNA Extraction**

Pellets were resuspended in 5ml QIAGEN buffer P1 (RNaseA added) and incubated with 250 µl of 10% SDS for 5 minutes at room temperature. Samples were then sonicated for 5 cycles each with 40% and 10s pulse to shatter the DNA into 10-100kb sized fragments and incubated with 10µl of proteinase K for 15 minutes at room temperature. Next, 5ml Phenol:Chloroform:Isoamylalcohol solution was added and thoroughly mixed. Samples were centrifuged at 8000 rpm for 60min, 20°C in a SORVALL RC 5B Plus rotor. The upper phase was then extracted and 500 µl 3M Sodium Acetate along with 4ml isopropanol added and incubated for 3 hours. After that, samples were spun for 30min, 20°C, at 8000rpm and the supernatant discarded. With 70% ethanol added and centrifuged again for 10 minutes at 20°C and 8000rpm. The supernatant was then decanted with the pellet air-dried for 30 minutes. Dried pellets were dissolved in nuclease-free H20 to a concentration of about 2µg/µl. Samples were then incubated overnight at 4°C. and then incubated for 15min at 80'C. DNA purity was assessed by measuring A260/280 and A260/230 ratios on the Nanodrop®.

4.5.12 **Amplification and high-throughput sequencing of shRNA barcodes**

Amplification of shRNA barcodes was performed according to Cellecta's guidelines ("Cellecta DECIPHER™ Pooled Lentiviral shRNA Libraries User Manual v9a") with some modification. For each sample, the entire amount of genomic DNA (200 – 1250 µg) extracted from each screening sample was used in the first round of PCR in order to ensure full representation of the barcodes from all cells harvested. Using the Titanium® Taq PCR kit (Takara-Clontech), 50 µl PCR reactions with 25 µg genomic DNA each and 0.3 µM FwdHTS and RevHTS1 primers (FwdHTS: 5'-TTCTCTGGCAAAAGACGGCATA-3' and RevHTS1: 5'-TAGCCAACGCATCGCACAAGCCA-3') were set up and run for 16 cycles with 65 °C annealing temperature. First round PCR amplified samples were pooled and 4 µl were used for the second round PCR, in order to reduce the amount of genomic DNA carryover and to attach adapters for subsequent high-troughput sequencing. In brief, 4 x 50 µl second round PCR reactions with 1µl first round PCR product each and 0.5 M FwdGex and RevGex2 primers were set up and run for 12, 14 or 16 cycles (65 °C annealing temperature), with the optimal number of cycles determined beforehand in order to obtain equal band intensities between samples. Second round PCR products were pooled, purified using the QIAquick® PCR purification kit (Qiagen) according to the manufacturer's instructions and quantified using a Qubit® 2.0 Fluorometer (Life Technologies) and Qubit® dsDNA HS assay kit (Thermo Fisher). Samples were adjusted to 5 nM, denatured with 0.1 M NaOH for 5 minutes at room temperature and subjected to single-read 18 bp sequencing on a MiSeq System (Illumina) using the MiSeq Reagent kit v2 (Illumina) with 11 pM sample and 500 nM HPLC-purified custom sequencing primer GexSeqSext. All primers used are specified in Supplementary Table S1.

4.5.13 **Quality control of high-throughput sequencing data**

Sequenced barcodes were deconvoluted using the Barcode Deconvoluter software (Cellecta) with a tolerance of 1 error and correction of N symbols. For complete dropouts in endpoint samples, a minimum value of 1 readcount was added. In order to adjust for varying total readcounts, all samples were normalized to a constant number of total reads (Supplementary Table S2). For each screen, fold change (FC; endpoint/baseline) values were calculated for all individual shRNAs. Library internal controls, 1) Luciferase (*LUC*)-targeting control shRNAs and 2) shRNAs with identical mRNA target sequence but different barcodes, served as technical quality controls for each screen.

After running the reaction, PCR products were purified using the Monarch® PCR & DNA Cleanup Kit (#T1030S) and eluted in 30 µL Elution Buffer. Samples were then run on a 3.5% agarose-gel (after Added 5 µL 6x LD to each sample) at 100V for 70 minutes in TAE buffer and made visible using Gelgreen®. The band corresponding to the 282 bp

amplicon cut out and subsequently purified the band using Monarch® DNA Gel Extraction Kit (#T1020S). This was then eluted to a final volume of 11 µL using elution buffer. Concentration was measured using Qubit™ dsDNA HS Assay Kit (Catalog number: Q32854). To quantify the libraries produced for Illumina sequencing, qPCR was performed using the KAPA Library Quantification Kit for Illumina platforms (#KK4844). Samples were normalized to 10 nM and multiplexed at equimolar concentrations.

4.5.14 **Sequencing**

Sequencing was carried out using the NextSeq500 (The NextSeq 500/550 High Output Kit v2.5 was used). For dilution and denaturation, the NextSeq System Denature and Dilute – Libraries Guide was followed. In brief, the library was denatured with 0.2 N NaOH for 5 minutes at room temperature and then 200 mM Tris-HCl, pH 7 was added (equivalent to the RSB of the kit). After that, libraries were diluted to a final concentration of 20pM using the HT1 solution of the kit. The PhiX library (sequencing control) was also denatured and diluted using the same reagents. The library was combined with PhiX and diluted with HT1 to obtain 1.3 mL of a 1.3 pM library concentration containing 15% PhiX. The Seq-12 NGS and Index-12 NGS primers were used at 0.3 µM according to the manufacturer's instructions. Read lengths for sequencing and indexing were set to 26 and 6 nt respectively.

4.5.15 **Deconvolution**

Sequencing reads were demultiplexed and aligned using the Cellecta Alignment software according to the user manual available under https://manuals.cellecta.com/ngs-prep-kit-for-sgrna-shrna-dna-barcode-libraries/v1/de/topic/demultiplex-and-align-sequencing-reads

4.5.16 **Data Analysis**

shRNA counts were normalized with results expressed as counts per Million (CPM). For baseline samples, a minimal number of 2 CPMs was used as a cut-off pro shRNA. Endpoint missing values of shRNAs were treated as zero counts. Next, trimmed mean of M values (TMM) normalization was applied.
To analyse shRNAs for enrichment or depletion, an RNAseq analysis with edgeR (with a negative binomial model) was used. Significance testing was carried out on both an shRNA- and gene-level for the endpoint compared to baseline. Due to the absence of a clear consensus on a gene-wise analysis based on all shRNA targeting a particular gene. Each gene was tested based on the shRNAs in a gene-set test approach. Genes with a consistent signal across their corresponding shRNAs should therefore be ranked high. In addition to comparing PN and MES separately for enriched and depleted

shRNAs under dasatinib and DMSO, samples were tested for shRNAs/genes that show a significant difference at endpoint between PN and MES, all compared to the baseline.

4.5.17 **Microarray gene expression profiling classification of GSCs**

For the classification of GSCs using the Wang et al. signatures (Wang et al., 2018) were used please refer to Alhalabi et al (Alhalabi et al., 2021) for specific details on the methodology and the data reposited.

4.5.18 **SILAC (stable isotope labeling of amino acids in cell culture) with mass spectrometry**

NCH 711d and NCH 705 (MES), NCH 421k and NCH 644 (PN) were cultured with SILAC heavy and light medium (DMEM:F12 with Arg-10 and Lys-6) for 6 doublings. After that, cell pellets were isolated, and protein extracted according to the phosphoprotein isolation protocol (see below) for incorporation percentage check. After confirming 95% incorporation in 'heavy' samples, GSCs containing 'heavy' amino acids were treated with 1µM Dasatinib and 'light' GSCs with DMSO vehicle control for 72 hours. After that, 1M cells were isolated for the day-4 time point and cell death determined using Propidium Iodide (PI) with flow cytometry. The remaining GCSs were further cultured under the same conditions and further reseeded and retreated with the same dasatinib concentration on day 8 and cultured for further four days, totaling 12 days of inhibition. GSC pellets were then harvested, and proteins isolated for proteomics and phosphoproteomics. Cell pellets of samples were resuspended in five volumes of SILAC lysis buffer (100 mM Tris-HCl pH 8.5, 7 M Urea, 1 % Triton, 10 U/ml DNase I, 1 mM magnesium chloride, 1% benzonase (add 1 µl to 1000 µl buffer), 1 mM sodium orthovanadate, phosphoSTOP phosphatases inhibitors and complete mini EDTA free protease inhibitors), sonicated at 40% output 3 times for 10s at 4°C with 30s breaks between cycles (0.5s-on and 0.5s-off setting on ice-water bath). Residual cell debris was removed by ultracentrifugation a 140 000g for 1h at 4°C then incubated at room temperature for 2 hours. After that, protein concentration was measured, and samples frozen. Frozen samples were submitted for further processing at the DKFZ Proteomics core facility. Protein samples were applied to 1D-SDS-PAGE and fractionated. Gel pieces were cut out and cysteins reduced by adding DTT and carbamidomethylated using iodoacetamide followed by overnight digestion with Trypsin. Resulting peptides were loaded on a cartridge trap column, packed with Acclaim PepMap300 C18, 5µm, 300Å wide pore (Thermo Scientific) and separated in a 180 min gradient from 3% to 40% ACN on a nanoEase MZ Peptide analytical column (300Å, 1.7 µm, 75 µm x 200 mm, Waters). Eluting peptides were detected by an online coupled Q-Exactive-HF-X mass spectrometer.

4.5.19 **Proteomics and phosphproteomics data pre-analysis**

Data analysis was carried out by MaxQuant (version 1.6.0.16). In total 79386 peptides and 8000 proteins were identified by MSMS based on an false discovery rate (FDR) cut-off of 0.01 on peptide level and 0.01 on protein level. Match between runs

option was enabled to transfer peptide identifications across Raw files based on accurate retention time and m/z. Quantification was done using a SILAC duplex approach with Arg10;Lys8 as labeled amino acids. A minimum number of quantified peptides was required for protein quantification, requantify option was activated to enable quantification of proteins with very high ratios. The output provided is a heavy/light ratio i. e. Dasatinib treated/control. Peptide and phosphopeptide ratios were automatically normalized to the total read count of the median value of peptide ratio (H/L) by the MaxQuant software.

4.6 References

Alexa, A., and Rahnenführer, J. (2007). Gene set enrichment analysis with topGO. Bioconductor Improv.

Alhalabi, O.T., Fletcher, M.N.C., Hielscher, T., Kessler, T., Lokumcu, T., Baumgartner, U., Wittmann, E., Schlue, S., Rahman, M.G., Hai, L., et al. (2022). A novel patient stratification strategy to enhance the therapeutic efficacy of dasatinib in glioblastoma. Neuro. Oncol. *24*, 39–51.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B.

Gorges, L.L., Lents, N.H., and Baldassare, J.J. (2008). The extreme COOH terminus of the retinoblastoma tumor suppressor protein pRb is required for phosphorylation on Thr-373 and activation of E2F. Am. J. Physiol. - Cell Physiol.

Gosline, S.J.C., Spencer, S.J., Ursu, O., and Fraenkel, E. (2012). SAMNet: A network-based approach to integrate multi-dimensional high throughput datasets. Integr. Biol. (United Kingdom).

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. In 7th Python in Science Conference (SciPy 2008), p.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. Nucleic Acids Res.

Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. Database.

Krug, K., Mertins, P., Zhang, B., Hornbeck, P., Raju, R., Ahmad, R., Szucs, M., Mundt, F., Forestier, D., Jane-Valbuena, J., et al. (2019). A Curated Resource for Phosphosite-specific Signature Analysis. Mol. Cell. Proteomics.

Lan, A., Smoly, I.Y., Rapaport, G., Lindquist, S., Fraenkel, E., and Yeger-Lotem, E. (2011). ResponseNet: Revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Res.

Li, L., Han, L., Zhang, J., Liu, X., Ma, R., Hou, X., Ge, J., and Wang, Q. (2016). Epsin2 promotes polarity establishment and meiotic division through activating Cdc42 in mouse oocyte. Oncotarget *7*.

Martinez, A.M., Afshar, M., Martin, F., Cavadore, J.C., Labbé, J.C., and Dorée, M. (1997). Dual phosphorylation of the T-loop in cdk7: Its role in controlling cyclin H binding and CAK activity. EMBO J.

McGowan, C.H., and Russell, P. (1995). Cell cycle regulation of human WEE1. EMBO

J.
Obenauer, J.C., Cantley, L.C., and Yaffe, M.B. (2003). Scansite 2.0: Proteome-wide prediction of cell signalling interactions using short sequence motifs. Nucleic Acids Res.
Okamoto, K., and Sagata, N. (2007). Mechanism for inactivation of the mitotic inhibitory kinase Wee1 at M phase. Proc. Natl. Acad. Sci. U. S. A.
Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res.
Rothblum-Oviatt, C.J., Ryan, C.E., and Piwnica-Worms, H. (2001). 14-3-3 Binding regulates catalytic activity of human Wee1 kinase. Cell Growth Differ.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res.
Stead, B.E., Brandl, C.J., Sandre, M.K., and Davey, M.J. (2012). Mcm2 phosphorylation and the response to replicative stress. BMC Genet.
Stupp, R., Tonn, J.C., Brada, M., and Pentheroudakis, G. (2010). High-grade malignant glioma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann. Oncol.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M. a, Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.
Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput. Biol. *12*.
Waskom, M. (2021). seaborn: statistical data visualization. J. Open Source Softw.
Wojcik, E.J., Buckley, R.S., Richard, J., Liu, L., Huckaba, T.M., and Kim, S. (2013). Kinesin-5: Cross-bridging mechanism to targeted clinical therapy. Gene.
Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat. Genet.

Chapter 5- Proteomics, post-translational modifications, and integrative analyses reveal heterogeneity of molecular mechanisms within medulloblastoma subgroups

Authors: Tenley C. Archer*, Tobias Ehrenberger*, Filip Mundt*, Maxwell P. Gold, Karsten Krug , Clarence K. Mah, Elizabeth L. Mahoney, Colin J. Daniel, Alexander LeNail, Divya Ramamoorthy, Philipp Mertins, D. R. Mani, Hailei Zhang, Michael A. Gillette, Karl Clauser, Michael Noble, Lauren C. Tang, Jessica Pierre-François, Jacob Silterra, James Jensen, Pablo Tamayo, Andrey Korshunov, Stefan M. Pfister, Marcel Kool, Paul A. Northcott, Rosalie C. Sears, Jonathan O. Lipton, Steven A. Carr, Jill P. Mesirov, Scott L. Pomeroy, Ernest Fraenkel

This was published in Cancer Cell. 2018 Sep 10; 34(3): 396-410. My primary contributions were analyzing the SHH-activated medulloblastomas and writing the first draft for a significant portion of manuscript.

5.1 Abstract

5.1.1 **Summary**

There is a pressing need to identify therapeutic targets in tumors with low mutation rates such as the malignant pediatric brain tumor medulloblastoma. To address this challenge, we quantitatively profiled global proteomes and phospho-proteomes of 45 medulloblastoma samples. Integrated analyses revealed that tumors with similar RNA expression vary extensively at the post-transcriptional and post-translational levels. We identified distinct pathways associated with two subsets of SHH tumors, and found post-translational modifications of MYC that are associated with poor outcomes in Group 3 tumors. We found kinases associated with subtypes and showed that inhibiting PRKDC sensitizes MYC-driven cells to radiation. Our study shows that proteomics enables a more comprehensive, functional readout, providing a foundation for therapeutic strategies.

5.1.2 **Significance**

Genomic and epigenomic analyses have revolutionized cancer diagnostics. Nevertheless, it has been difficult to identify therapeutic targets for tumors that lack recurrent genomic lesions. Here we used global, mass spectrometry-based measurements of protein levels and post-translational modifications to identify functional pathways associated with subtypes of medulloblastoma. Strong proteomic signals revealed altered pathways that were not detected transcriptionally. One molecular subgroup of tumors showed marked discordance of RNA and protein levels, suggesting global changes in translation and/or proteostasis. We demonstrate the utility of an integrative approach for discovery of candidate biomarkers or drug targets and provide a multi-omic dataset that will serve as a resource for the community. This study has the potential to impact clinical trial design.

5.2 Introduction

Medulloblastoma is one of the most common pediatric brain tumors. Survival rates are high, but current therapies can leave lasting side effects including problems with speech, cognition and behavior, and increased risks of secondary cancers (Archer, Mahoney and Pomeroy, 2017). Identifying the pathways driving medulloblastoma could guide development of less toxic and more effective targeted therapies. Previous analyses have demonstrated that, at the molecular level, medulloblastoma is extremely heterogeneous and comprises at least four major consensus subgroups: wingless (WNT), sonic hedgehog (SHH), Group 3, and Group 4 (Cho *et al.*, 2011; Tamayo *et al.*, 2011; Kool *et al.*, 2012; Pugh *et al.*, 2012; Robinson *et al.*, 2012; Taylor *et al.*, 2012; Hovestadt *et al.*, 2014). Almost all WNT tumors carry activating mutations in the β-catenin gene (*CTNNB1*) (Pugh *et al.*, 2012). Unfortunately, the ubiquity of WNT signaling in non-cancer cells makes this pathway a challenging one for targeted cancer therapeutics (Kahn, 2014). Some SHH tumors respond to SMO inhibitors (Kool *et al.*, 2014; Robinson *et al.*, 2015). Group 3 and Group 4 are the least understood subgroups and constitute more than half of all medulloblastoma cases. These tumors have few consistent genetic abnormalities amenable to currently available targeted therapies (Northcott *et al.*, 2012).

Deeper analysis of the transcriptome and epigenome has revealed subtypes within the four consensus subgroups as defined by Taylor *et al.* (2012). Based on transcriptome evidence, we reported two subtypes of Group 3 tumors, including one with a dominant MYC-driven signature and exceedingly poor prognosis (Pfister *et al.*, 2009; Cho *et al.*, 2011; Schwalbe *et al.*, 2017). More recently, Northcott *et al.* (2017) defined eight subtypes of Group 3 and Group 4 tumors based on DNA methylation, including a category of MYC-driven samples called subtype II. Similarly, Cavalli *et al.* (2017) proposed a division of Group 3 and Group 4 into a total of six different subtypes. Recent evidence suggests that the MYC-driven Group 3 tumors may be susceptible to CDK inhibitors or to combinations of PI3K and HDAC inhibitors (Hanaford *et al.*, 2016; Pei *et al.*, 2016).

Here, we analyzed the abundance of proteins and their post-translational modifications (PTMs) in medulloblastomas, and we integrated the data with previously reported DNA methylation, whole genome sequencing, and RNA expression. Since proteins are ultimately the functional effectors of biological activity in cancer cells, we hypothesized that global proteomic analysis may be an especially sensitive method for identifying potential therapeutic targets in medulloblastoma. Recent advances in mass spectrometry have allowed for proteomic and phospho-proteomic profiling of other cancers, and integration of these data with other biological data developed a more complete understanding of specific cancers and their genetic drivers (Zhang *et al.*, 2014, 2016; Edwards *et al.*, 2015; Lawrence *et al.*, 2015; Mertins *et al.*, 2016; Huang *et al.*, 2017).

We show that proteomics provides functional insights into the heterogeneity underlying medulloblastoma. We find a split in SHH samples that is caused by post-transcriptional

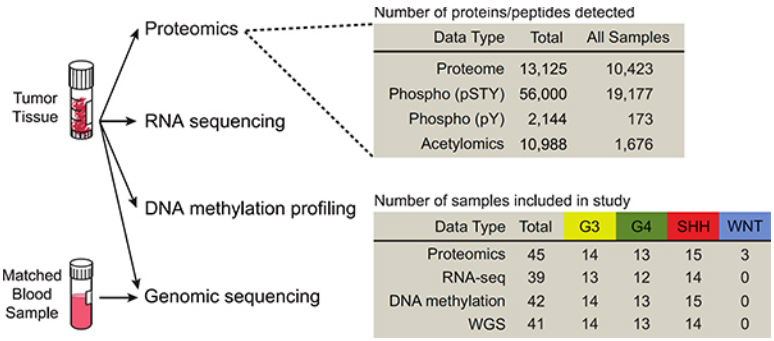
regulation changes, which impact protein abundances but not RNA levels. There are no genomic features that fully explain these proteomic subsets. However, one group exclusively harbors chromosome 3q gains and mutations in *PTCH1*, *PRKAR1A*, and the *TERT* promoter. We also identify a subset of Group 3 samples that are MYC-driven, have a poor prognosis, and have PTMs to MYC that can affect its stability. Additionally, integrative network and kinome analyses along with functional assays in cell lines identify targetable pathways. Functional validation experiments confirm several of these predictions in models of medulloblastoma, including providing evidence that inhibition of PRKDC may have therapeutic benefit in MYC-driven Group 3 tumors.

5.3 Results

5.3.1 **Global proteomics reveals medulloblastoma subgroups**

To obtain a comprehensive view of medulloblastoma, we selected 45 primary tumors from all consensus subgroups in a cohort that had been previously characterized (Cho *et al.*, 2011; Kool *et al.*, 2012; Hovestadt *et al.*, 2014; Northcott *et al.*, 2017). Of these, 42 had been analyzed for DNA methylation, 41 had preexisting whole genome sequencing and 39 had RNA-seq data (Figure 1, Table S1)(Northcott *et al.*, 2017). We collected proteomic, phospho-proteomic, and protein acetylation data for all 45 samples using isobaric labeling with TMT10 mass-tag reagents (Rauniyar and Yates 3rd, 2014) followed by high-resolution liquid chromatography-tandem mass spectrometry. Phosphorylation events were enriched using metal-affinity enrichment (detecting phosphorylated serine, threonine, and tyrosine peptides: pSTY). In addition, we used antibodies to enrich for phosphorylated tyrosine peptides (pY) and acetylated lysine residues (acK). Over 13,000 proteins, more than 50,000 phosphosites, and almost 11,000 acetylated sites were quantified in total. Unless otherwise noted, only the complete data sets, in which proteomic features had been measured across all 45 samples, were used for analyses (Figure 1).

We applied uniform data normalization methods to each type of data and clustered samples using each data type separately (Figure 2A). Consensus clustering, principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE) carried out separately on each data type predominantly revealed the known subgroups: Group 3, Group 4, and SHH (Figures 2A, S1-S3). The proteomic data also identified very stable subsets of two known subgroups (Figures S1-S3). Here, we refer to Group 3 clusters as Group 3a (G3a) and Group 3b (G3b), and SHH clusters as SHHa and SHHb (Figure 2A, Table S1). Including WNT samples in any of the clustering approaches on the proteomic data revealed the same five subsets with an additional sixth group for WNT samples (Figure S3). Performing pairwise statistical comparisons, we found 4,365 proteins and 2,642 phosphopeptides that differed significantly between these subsets of patients (FDR < 0.01, ANOVA; Table S2).



Number of proteins/peptides detected

Data Type	Total	All Samples
Proteome	13,125	10,423
Phospho (pSTY)	56,000	19,177
Phospho (pY)	2,144	173
Acetylomics	10,988	1,676

Number of samples included in study

Data Type	Total	G3	G4	SHH	WNT
Proteomics	45	14	13	15	3
RNA-seq	39	13	12	14	0
DNA methylation	42	14	13	15	0
WGS	41	14	13	14	0

Figure 1:Summary of data types included in this study, depth of proteomic data types, and cohort composition. The extent of the data from proteomics, including post-translational modifications, is summarized at the top. pSTY: phosphorylation on serine, threonine, or tyrosine detected after immobilized metal affinity chromatography; pY: phosphorylated tyrosine detected after antibody purification; Total: the total number of features identified; All Samples: the number of features measured across all samples, i.e. without any missing values. The number of samples covered by each data type and their split by subgroups are summarized at the bottom. G3: Group 3; G4: Group 4; WGS: whole genome sequencing. Proteomics includes proteome, pSTY, pY, and acetylomics data sets. See also Table S1.

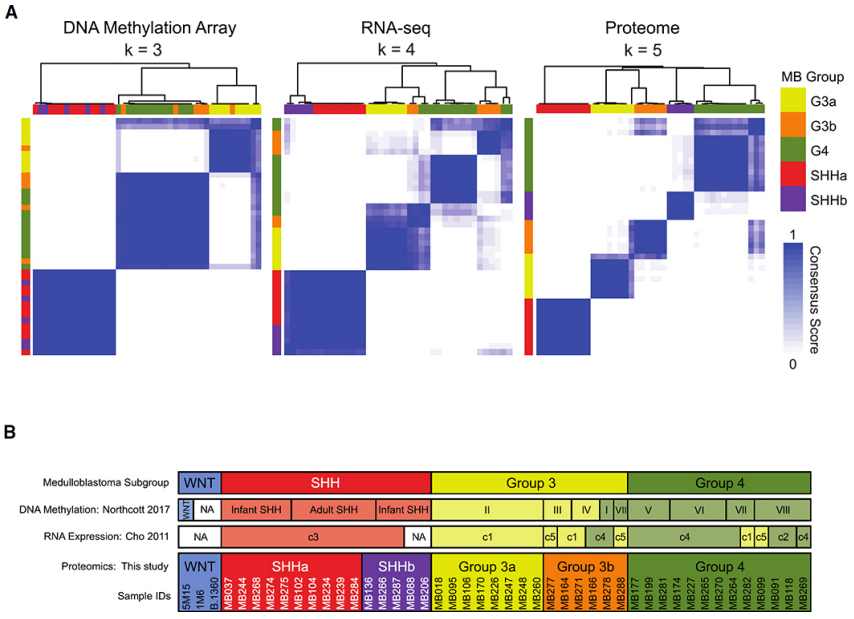


Figure 2: Comparison of clustering results. (A) The optimal clustering of DNA methylation data, RNAseq, and proteome, as determined using Pearson correlation as distance metric. k, number of clusters. Consensus scores are indicated using a color scale from white (samples never cluster together) to blue (samples always cluster together). (B) Comparison of the assignment of samples using the four "consensus subgroups" (Taylor et al., 2012), the DNA Methylation-based subtype-calls assigned in Northcott et al. (2017), which included most samples used in our study, and RNA expression assignments based on application of the classifier described in Cho et al. (2011). NA: no assignment available. See also Figures S1-S3 and Tables S1 and S2.

5.3.2 **Proteome suggests post-transcriptional heterogeneity within SHH medulloblastoma**

The SHHa and SHHb subsets that emerged (Figures 2A, S1-S3) are distinct from the age-dependent subtypes found in DNA methylation data (Northcott *et al.*, 2011; Kool *et al.*, 2014; Cavalli *et al.*, 2017; Schwalbe *et al.*, 2017). All but one of the adult samples were found in SHHa, while the pediatric SHH samples were split evenly between SHHa and SHHb (Figures 2B, 3, S3, Table S1). We identified 510 proteins that differed between the SHHa and SHHb using ANOVA (FDR < 0.005). Proteins higher in SHHa were associated with mRNA processing, splicing, and transcription, as well as the MYC pathway, chromatin remodeling, and DNA repair (Figure 3A, Table S3). In contrast, proteins with higher levels in SHHb were linked to neuronal and neurotransmitter-like activity, including CD47, an anti-phagocytic cell surface ligand (Figure S4). Many proteins in the glutamatergic synaptic pathway were elevated in SHHb, including glutamate, calcium, and MAPK/ERK signaling (Figures 3A, 3B, Table S4). SHHb samples consistently clustered closer to Group 4 samples than to SHHa samples (Figures 2A, 3A, S1-S3). Despite the differences in neuronal-like gene sets between SHHa and SHHb, there were no differences in histology (Figure 3C).

As no transcripts (RNA-seq) differed significantly (FDR < 0.05, ANOVA) between SHHa and SHHb (Table S2), we asked whether the subsets differed in post-transcriptional regulation. We compared the Spearman correlations for approximately 8,700 mRNA-protein pairs in every patient sample (Figure 3D). All clusters of samples except SHHb have a median Spearman correlation near 0.5, consistent with studies in other systems (Zhang *et al.*, 2014; Mertins *et al.*, 2016). However, the median correlation for SHHb was 0.38 ($p = 0.012$, compared to SHHa; Mann-Whitney U test), suggesting that SHHa and SHHb differ in translation and/or proteostasis.

We next asked if genetic lesions might account for the observed differences between SHHa and SHHb. Gains of chromosome 3q are characteristic of SHH medulloblastoma (Cho *et al.*, 2011; Kool *et al.*, 2012; Taylor *et al.*, 2012), and in our cohort they occur in several SHHa samples, and do not occur in SHHb samples (Figure 3C). The neural progenitor regulator SRY-box 2 gene (*SOX2)* lies within this region, and its protein levels and phosphorylation events (T7, S18, and T26) were all increased in SHHa samples compared to SHHb (Figures 3C, S4)(Archer, Jin and Casey, 2011). Mutations in *PTCH1*, *PRKAR1A*, and in the *TERT* promoter (Figure 3B) occurred only in SHHa. In contrast, other SHH pathway activating alterations were found in both SHHa and SHHb, which suggests that there are relatively few differences in the genetic lesions of these SHH subsets.

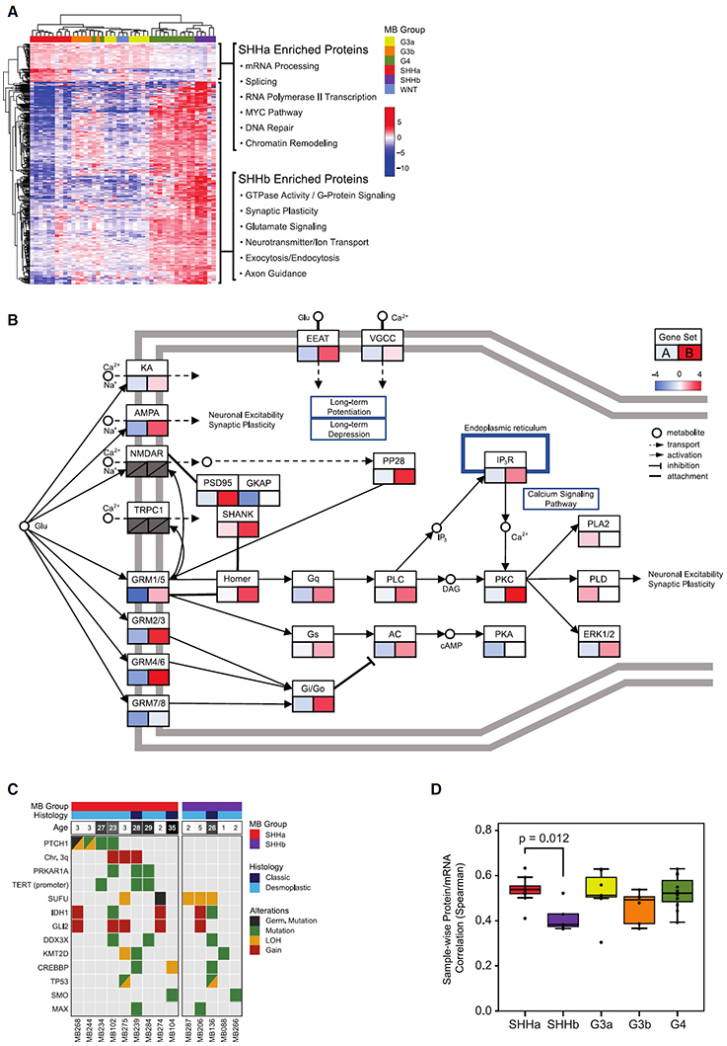


Figure 3: Molecular and functional differences between SHHa and SHHb.
(A) Hierarchical clustering of 510 proteomic features that differ significantly between SHHa and SHHb (FDR < 0.005; ANOVA). Summaries of gene sets that significantly differ (FDR < 0.01; hypergeometric overlap tests) between SHHa and SHHb are shown alongside the heatmap. (B) A schematic of the pathway for the glutamatergic synapse. Each box represents a gene or gene set taken directly from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway. The left and right boxes summarize SHHa and SHHb protein levels, respectively. Grey boxes indicate a lack of proteomic data for the gene product. Small white circles represent metabolites, solid directed arrows activation, and lines ending with a bar inhibition. Dotted lines indicate transport of metabolites. Solid lines and connected boxes represent physical interactions in the cell. Figure adapted from KEGG pathway hsa04724 (Kanehisa et al., 2017). (C) Clinical annotations of and recurrent genomic alterations in SHHa and SHHb samples. (D) Box-and-whisker plots of Spearman correlations between the mRNA and protein levels for approximately 8,700 genes for each tumor sample. Boxes range from first to third quartile in each group, with line indicating the group-median. The whiskers extend to the lowest and highest data points within 1.5 times the interquartile range from the box (p value shown from Mann Whitney U test). See also Figure S4 and Tables S3 and S4.

5.3.3 Post-translational modifications of MYC in Group 3 tumors are predictive of patient outcome

Clustering of the proteomic data identified two subsets of Group 3 samples, which we refer to as G3a and G3b (Figures 2, S1-S3). First, we sought to understand how these clusters related to the subgroups that had been previously identified using much larger cohorts. All G3a samples were assigned to subtype II that was defined by Northcott *et al.* (2017), while the G3b samples were assigned to the Northcott *et al.* (2017) subtypes I, III, IV, and VII (Figure 2B, Table S1). The transcription-based classifier of Cho *et al.* (2011) assigned the G3a samples to the MYC-activated c1 subtype. Two G3b samples and one Group 4 sample also were classified as c1 (Figure 2B, Table S1). The remaining G3b samples were assigned to c4 and c5. These results suggest that the proteomic features associated with G3a likely represent the MYC-activated form of medulloblastoma and that the proteomic data for G3b samples represent the known Group 3/4 continuum. Indeed, several proteomic signatures in G3a associated with MYC activation including significant differences in ribosomal proteins and proteins related to ribosome assembly, mitochondrial ribosomal proteins, and proteins involved in transcription (Figure S5A)(Morrish and Hockenbery, 2014; Staal *et al.*, 2015).

To identify the sources of MYC activation in G3a, we investigated the MYC events in all our data (Figure 4A). While *MYC* amplification is a “hallmark” of MYC-activated medulloblastoma, it does not occur in every tumor of this type (Cho *et al.*, 2011; Northcott *et al.*, 2017). Here, only two G3a tumors have a *MYC* amplification. However, all G3a samples have increased post-translational modifications of MYC at multiple sites (Figures 4A, S5B). Acetylation of lysine 148 (K148) was pronounced, as was increased phosphorylation of serine 71 (S71) and a serine at either position 62 or 64 (the ambiguity of these nearby sites cannot be resolved in the spectra). Our data also revealed peptides that are simultaneously phosphorylated at both serine (S62) and threonine 58 (T58). Peptides phosphorylated on T58 and S62 are particularly informative for MYC activity, as these sites regulate MYC half-life and transcriptional activity (Arnold *et al.*, 2009; Wang *et al.*, 2011). Considering known MYC regulators, our data showed significantly increased protein levels of the B55a subunit of PP2A (encoded by *PPP2R2A*, FDR = 0.0085, ANOVA), and the deubiquitinating enzymes USP28 (significant with FDR = 0.018, ANOVA) and USP36 (not significant with FDR = 0.084, ANOVA) in G3a compared to G3b (Figure S5C). By contrast, we did not observe differing levels of PP2A-B56 or the phosphatase inhibitors SET or CIP2A (Figure S5C).

To examine the localization of active forms of MYC, we stained formalin-fixed and paraffin-embedded (FFPE) slides of the tumors for MYC phosphorylated at S62 or T58 (Figure 4B). Cell mean immunofluorescence density for phosphorylation of residues S62 and T58 MYC (pS62 and pT58) was significantly higher in G3a compared to G3b (Figure 4C), while Group 4 tumors showed lower levels of signal. Localization of pS62 MYC was primarily in the nucleus with exclusion from the nucleolus (Figure 4B). In contrast, pT58 MYC localized mainly in the cytoplasm for G3a and G3b tumors. Elevated pS62 and pT58 MYC suggest a breakdown in the canonical pathway of MYC degradation (Farrell and Sears, 2014), consistent with the increased expression of

some kinases upstream of S62 phosphorylation as well as increased expression of deubiquitinating enzymes USP36 and USP28.

We sought to understand whether the observed MYC modifications had clinical implications. Indeed, G3a and G3b patients differed dramatically in rates of five-year progression free (PFS) and overall survival (OS), but these differences were not statistically significant, possibly due to the small sample size (Figures 4D). To extend our analysis to a larger cohort, we built a binary classifier based on single sample gene set enrichment analysis (Barbie *et al.*, 2009) to distinguish G3a and G3b using only transcriptional data. We then applied it to the c1 and c5 samples from Cho *et al.* (2011) to give them G3a/G3b labels (see Methods, Figure S5D, Table S1). Surprisingly, while we had expected c1 samples from the original Cho cohort to be assigned to G3a and c5 to G3b, the classifier assigned approximately one quarter of the samples to the other subtype. The combined cohort using proteomic labels performed better at predicting PFS and OS (Figure 4D, S5E). The difference in PFS using c1 and c5 labels (Figure 2B and Table S1) on the combined cohort did not reach statistical significance ($p = 0.098$), while the difference using G3a vs. G3b classification did ($p = 0.002$). The newly defined sets of patients had the same median age (5 years), ruling out that variable as a potential trivial explanation for the improved predictor. Taken together, these data demonstrate that the proteomic distinctions provide a strong signal of clinical relevance.

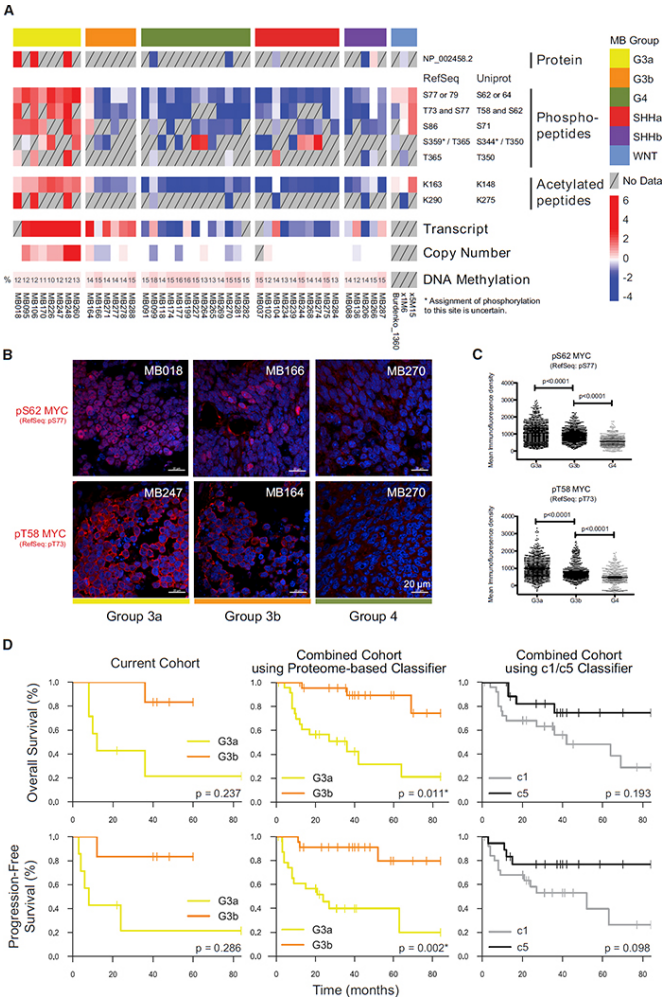


Figure 4: G3a medulloblastomas have post-translational modifications of MYC protein that strongly correlate with MYC downstream activity and patient outcome. (A) Heatmap of multi-omic data for MYC, showing protein levels, post-translationally modified peptides, mRNA expression, DNA copy number, and DNA methylation. Scale bar is of log2 values for all data sets except DNA methylation, which ranges from 0 to 1 on a linear scale. Methylation values, converted to percentages, are superimposed on the heatmap. The phosphorylation locations are provided both based on RefSeq (as used in our data sets) and Uniprot (frequently used in the literature). S77 or 79: phosphorylated at either S77 or S79 but not at T73; S359* / T365: phosphorylated at T365 as well as a proximal site that could not be uniquely determined. (B) Representative confocal microscopy images showing pS62 and pT58 MYC in G3a, G3b, and Group 4 tumors. (C) Scatter bar charts showing mean immunofluorescence densities quantified across 50–100 nuclei in 3–4 fields of view captured at 40x for MYC pS62 and pT58. Error bars represent standard error of means (SEM) and significance reported is from two-tailed unpaired t-tests. (D) Kaplan-Meier plots of overall survival and progression-free survival of Group 3 medulloblastoma from the current cohort (left) and the current cohort combined with Group 3 samples from Cho et al. (2011). The survival analysis on the combined cohort was performed using the classifier defined by the proteomic clusters G3a and G3b ("Proteome-based Classifier" in the center) and the original classifier from Cho et al. (2011) (c1/c5, right, gray scale). In the combined cohort with proteomebased group assignments (G3a/G3b), both outcome predictions reached statistical significance (p values as shown; Log-rank test). See also Figure S5.

5.3.4 **Medulloblastoma subgroups differ in activity of kinases**

We developed two approaches to investigate kinase signaling in medulloblastoma (see kinome analysis in the Methods). Potential kinases of the phosphopeptides in each subset of tumors were first identified by leveraging kinase specificity data from PhosphoSitePlus (Hornbeck *et al.*, 2015) and Scansite (Obenauer, Cantley and Yaffe, 2003). Filtering this list based on the protein levels of the kinases in our data, we identified the 38 kinases listed in Figure 5 (Table S5). PRKDC, GSK3B, and CDK5 were significant in both our analyses. PRKDC is important for repair of DNA double-stranded breaks through non-homologous end-joining (Ma *et al.*, 2004) and downstream targets of PRKDC were consistently elevated in Group 3 and WNT samples (Figure 5, Scansite). GSK3B is a promising therapeutic target. It is overexpressed in many cancers (McCubrey *et al.*, 2014) and its substrates have increased phosphorylation in the Group 4 and SHHb samples. CDK5 has been associated with oncogenesis and resistance to cancer therapies (Pozo and Bibb, 2016) and its targets had increased phosphorylation in Group 4. CDK5 phosphorylates MYC at S62, much like CDK1 and CDK7, which were fittingly associated with G3a. AURKB was also linked to this disease group. Diaz *et al.* (2015) previously treated MYC overexpressing medulloblastoma cell lines and orthotopic xenografts with an AURKB inhibitor, which led to growth impairment and induction of apoptosis in the cell lines, and inhibited intracranial growth and prolonged survival in mice. Many kinases were predicted to be active in SHHa, and this was the only subgroup with consistent phosphorylation of sites targeted by ATM (PhosphoSitePlus) and PIK3R1 (Scansite). G3b was unusual, as only one kinase, EEF2K, was associated with this group in our analyses. Overall, these analyses highlight potential upstream kinases that may be driving some of the molecular differences between medulloblastoma subgroups and that represent possible novel targets to explore for future therapeutic gains.

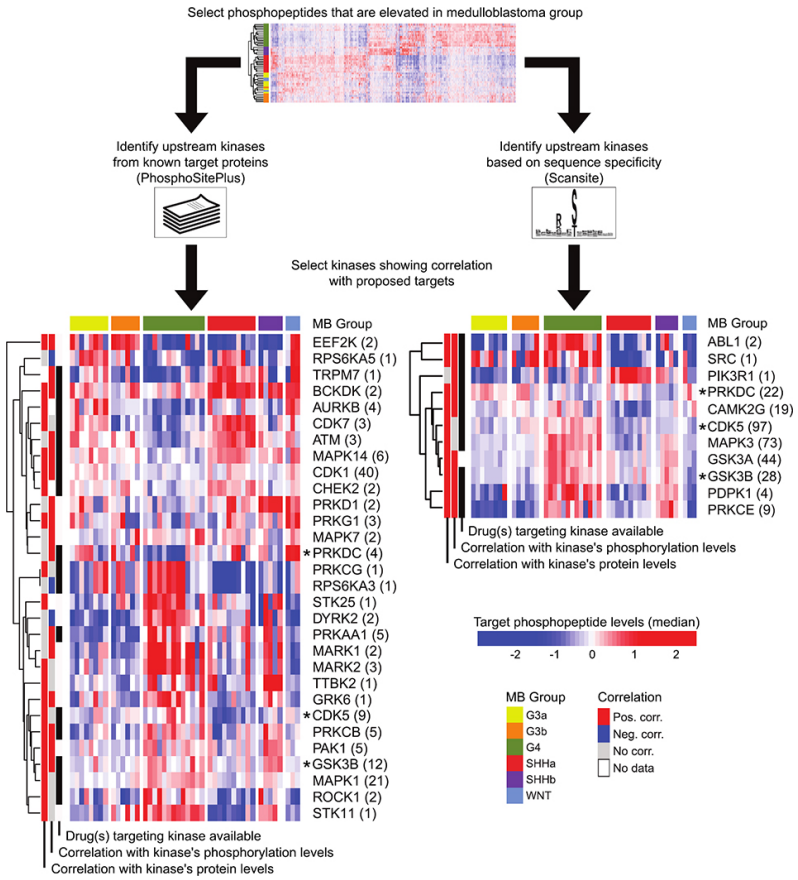


Figure 5: Medulloblastoma subtypes differ in kinase regulation and substrate levels.
Two different methods used to analyze the kinome are shown. On the left, upstream kinases are predicted from differential phosphopeptides between subgroups using PhosphoSitePlus database (Hornbeck et al., 2015). On the right, kinases are predicted from scoring differential phosphopeptides using sequence specificity motifs from Scansite (Obenauer et al., 2003). Heatmaps show the median levels of peptides matched to an upstream kinase, with the number of peptides matching each kinase shown in parentheses. Kinases found by both methods are annotated with an asterisk. Bars on the side of the heatmaps indicate whether target peptides correlate with protein or phosphorylation levels of upstream kinases; and if DrugBank (Law et al., 2014) lists any drugs targeting the kinases. See also STAR Methods and Table S5.

5.3.5 MYC-active medulloblastoma cell lines have phosphorylated MYC and PRKDC

To test the finding that pS62 MYC is enriched in Group 3 MYC-active medulloblastoma, we measured MYC expression by Western blot in medulloblastoma cell lines reported to be *MYC*-amplified (D425, D458 and D556) compared to two that are not (DAOY and D283)(Bigner *et al.*, 1990). *MYC*-amplified lines were highly enriched for MYC protein and pS62 MYC, while DAOY and D283 were not (Figure 6A-C). Consistent with the kinome analysis, pS2056 PRKDC correlates with pS62 MYC and pT58/pS62 MYC in these lines (Figures 6A-C). To explore possible functional consequences of this finding,

we first examined the localization of PRKDC and MYC. Immunofluorescence identified that pS62 MYC and pS2056 PRKDC co-localize in the nucleus (Figure 6D).
We next examined the role of PRKDC in *MYC*-amplified medulloblastoma cell lines. PRKDC inhibitors are radio-sensitizing agents, but have limited cytotoxic range (Ciszewski *et al.*, 2014; Sunada *et al.*, 2016). We find that that the PRKDC inhibitor NU7441 preferentially sensitizes MYC-active cell line D458 to radiation (Figures 6E-F). Irradiation reduced the IC_{50} of the *MYC*-amplified D458 from 28 μM to 2.7 μM (Figure 6G), but did not change the IC_{50} of NU7441 in DAOY. These data suggest that cell lines with pS62 MYC depend on PRKDC activity for survival in response to DNA damage, and that PRKDC inhibition may radio-sensitize MYC-active tumors.

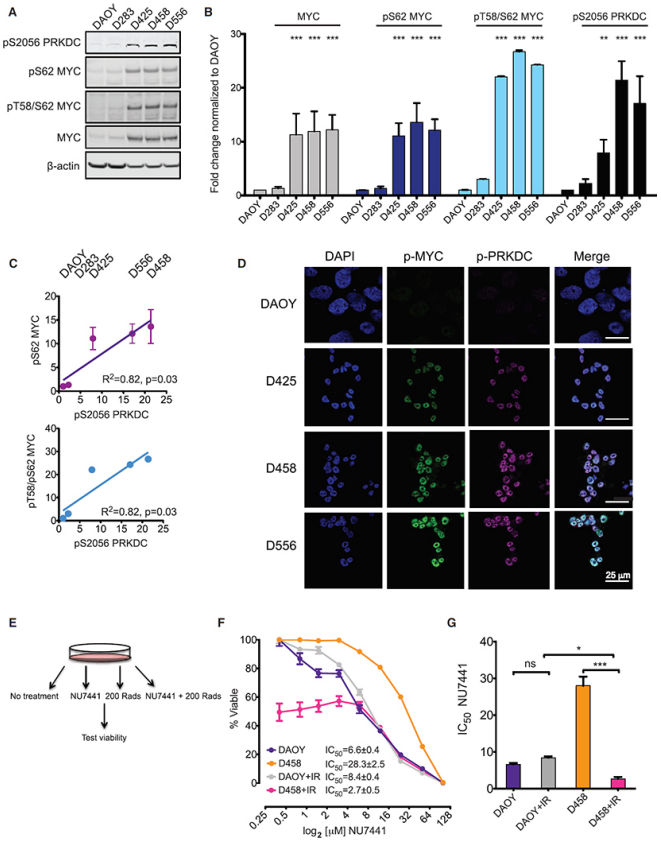


Figure 6: MYC status correlates with PRKDC phosphorylation, and predicts increased sensitivity to PRKDC inhibition with irradiation. (A) Representative Western blots of medulloblastoma cell lines performed in at least triplicate. Antibody against pT58/S62 MYC detects either site or both sites. (B) Quantification of Western blots represented as fold change compared to expression in DAOY cells. All proteins are normalized to β-actin. Significance was determined via one-way ANOVA with a Dunnett multiple comparison test to compare the normalized signal for each antibody across the five cells lines; **p < 0.001, ***p < 0.0001. Error bars represent mean normalized signal ± SEM. (C) Correlations between normalized means for specified antibodies determined by calculating a Pearson correlation coefficient. Error bars indicate ± SEM, and are depicted but are not visible for some data points because of scale. (D) Representative confocal images of indicated cell lines showing pS62 MYC and pS2056 PRKDC. (E) Experimental design of dose-response curve of PRKDC inhibitor NU7441 for 18 hours prior to irradiation. (F) Viability assay of medulloblastoma cell lines treated as indicated. Plotted is the mean of 6 biological replicates ± SEM for each dose; note, small error bars for D458 are depicted but obfuscated by trend lines. (G) Histogram of mean IC50 values ± SEM for DAOY vs D458 treated with NU7441 ± irradiation. *p < 0.01; ***p < 0.0001; ns = not significant; IR = irradiation.

5.3.6 **Integrative modeling**

To search for common patterns in the genomic, proteomic, and phospho-proteomic data, we adopted an integrated modeling approach (Figure 7). Omics Integrator (Tuncbag *et al.*, 2016) searches a vast network of physical interactions for sets of proteins and genes from disparate 'omic data that are likely to represent pathways altered in a disease process. Applied to signals that differ significantly between G3a and G3b, it identified coherent proteomic changes in proteins that physically associate with MYC, and up-regulation of known MYC transcriptional targets (Figure 7A). In addition, the networks highlighted ribosomal, mitochondrial and cell-cycle regulatory proteins. A similar analysis for SHH (Figure 7B) identified calcium, glutamate and Ras signaling pathways that were upregulated in SHHb. These networks integrated disparate data types including: mutated genes (*TP53*, *PIK3R1* and *NOTCH1*) and kinases with genomic/proteomic/PTM changes (CAMKs (CAMK2A/G, CAMK4), protein kinase A and C (PRKACB, PRKCA), ERK2 (MAPK1), ribosomal protein kinase 6 (RPSK6KA1/2), glycogen synthase kinase B (GSK3B), PI3Ks (PIK3R1, PIK3CA), DGKI, and CDK2). The networks also supported the role of several predicted kinases (PIK3R1, CAMK2G, GSK3B and MAPK1).

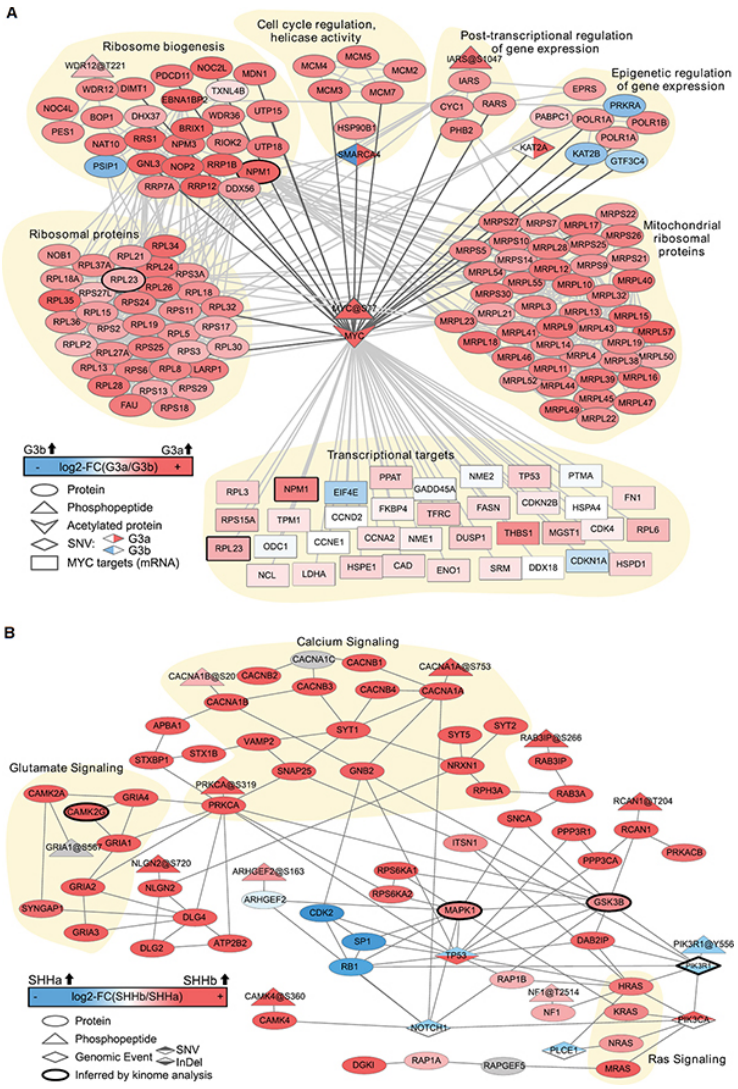


Figure 7: Network methods relying on known protein-protein interactions identify pathways relating to SHHb and G3a tumors. Omics Integrator output showing network views of proteins, posttranslational modifications, and genomic alterations associated with G3a (A) and SHHb (B). Node shapes indicate data type and colors indicate log2-based fold change between groups as described in the legends. Phosphopeptides are labeled with their phosphorylation sites (based on RefSeq) after the '@' symbol. Nodes associated with selected pathways are highlighted with yellow background. Grey nodes were added by Omics Integrator and have no associated proteomic data for our samples. (A) mRNA levels of transcriptional targets of MYC are shown at the bottom. Thick borders highlight proteins that are also shown as direct transcriptional targets of MYC. SNVs, indicated by diamonds, are color-coded to show which subtype the genomic alteration was seen in: G3a, red; G3b, blue. (B) Kinases that were also found by our independent Kinome Analysis (Figure 5) are highlighted with thick borders. The color of genomic lesions (diamonds) indicates the subtype in which they occur: SHHb, red; SHHa, blue. The location of color in a diamond indicates the type of genomic lesion: upper triangle, SNV; lower triangle, indel.

5.4 Discussion

Molecular subgroups identified using mRNA expression and DNA methylation array data have recently been accepted by the WHO as the international standard for medulloblastoma diagnosis (Louis, Ohgaki, *et al.*, 2016; Louis, Perry, *et al.*, 2016). Here we show that, at the highest level, proteomic and phospho-proteomic data sets revealed similar subgroup assignments to these consensus subgroups. This finding contrasts with global proteomic studies in breast and colon cancer where molecular subgroups were not as consistently durable across data types, and subgroup compositions are dependent on the data type used (Zhang *et al.*, 2014; Mertins *et al.*, 2016). The consistency of the medulloblastoma subgroups may in part be due to the distinct developmental states of the cell of origin of medulloblastoma subgroups, as well as the subgroup-specific driver genetic events (Gibson *et al.*, 2010; Northcott *et al.*, 2012).

Our data identify heterogeneous molecular mechanisms within the subgroups that are not evident in the transcriptome or genome. The two clusters of SHH samples revealed by our proteomic data (SHHa and SHHb) reflect a different distinction from the known age-based split in SHH samples (Kool *et al.*, 2014; Cavalli *et al.*, 2017; Schwalbe *et al.*, 2017). Adult patients in our cohort (ages 23-35) predominantly clustered with SHHa, but the infant and childhood patients were spread across both subsets. The different signals that emerged from DNA methylation and proteomics may reflect the differing sensitivity of these methods to specific biological processes. DNA methylation data likely reflect the developmental state of the cells of origin at the onset of oncogenesis (Horvath *et al.*, 2015; Lu *et al.*, 2016). By contrast, the proteomic data will be strongly affected by post-transcriptional changes including RNA stability, protein stability, translational regulation, and signaling pathways.

The proteomic data we gathered from SHH tumors have important clinical implications. SHHa samples had expression signatures and molecular alterations including *PTCH1* mutations that are consistent with activation of the canonical SHH pathway. SHHb tumors also had SHH pathway activating mutations, but also were enriched for pathways typically associated with Group 4, such as glutamate, calcium, and Ras signaling. For example, we saw very consistent increases in many of the proteins associated with the glutamatergic synapse (Figures 3C, 7B). While these pathways are known features of some gliomas and Group 4 medulloblastoma (Arcella *et al.*, 2005; Cavalli *et al.*, 2017), they are not typically associated with SHH medulloblastoma. SHHb-like patients may therefore also benefit from any future therapies developed for Group 4 tumors. Proteomic data identifies one potential candidate therapeutic target: CD47 protein is enriched in Group 4 and SHHb tumors. CD47 is a membrane protein involved in several processes, including vesicle-mediated transport, and is an anti-phagocytic cell surface ligand (Brown and Frazier, 2001; Jaiswal *et al.*, 2009). Currently, an anti-CD47 antibody is being tested for efficacy in Group 3 medulloblastoma (Gholamin *et al.*, 2017). While *CD47* RNA levels are not significantly different between the SHH groups ($p = 0.06$, t-test), the mean CD47 protein level for SHHb samples is much higher than the mean for SHHa and Group 3 samples. These data suggest that anti-CD47 therapies may be particularly effective on SHHb tumors. The discrepancy

between protein and mRNA levels in CD47 is not unique, as the correlation of global protein levels with mRNA expression was significantly lower in SHHb samples. It will be important for future studies to examine whether there exist fundamental differences in the efficiency of protein translation or in the stability of proteins in SHHb tumors. These observations highlight the importance of proteomic studies for discovery of biomarkers.

Group 3 medulloblastoma are clinically diverse and it has been clear that MYC activation has important therapeutic consequences. Previously, we demonstrated that transcriptional profiles can distinguish low- and high-risk Group 3 patients (Cho *et al.*, 2011; Tamayo *et al.*, 2011). More recent studies have proposed additional subtypes of Group 3. Cavalli *et al.* (2017) proposed three subtypes, and Northcott *et al.* (2017) identified eight subtypes of Group 3 and Group 4. However, genomic, epigenomic, and transcriptional data do not directly measure activation of MYC, which can occur through several mechanisms. In breast cancer, for example, more than 40% of tumors show increased levels of MYC protein, but the fraction with increased MYC mRNA ranges from 22-35% and only 13-22% have an amplification of the *MYC* locus (Chen and Olopade, 2008). In breast cancer, higher MYC protein levels without genomic amplification have been explained by elevated levels of pS62, a phosphosite associated with more stable and transcriptionally active MYC (Janghorban *et al.*, 2014). Using the direct measurement of MYC post-translational modifications, we were able to refine the transcriptional signature of MYC-activated Group 3 tumors. It is clear from the pathway-level analysis that G3a tumors have higher levels of MYC activity compared to those in the G3b cluster. Consistent with these data, we see pS62 on MYC in the proteomic data and by staining both of FFPE tumor sections and medulloblastoma cell lines; and we found that pS62 MYC was primarily expressed in the nucleus. More work will be needed to understand the functional consequences of the observed post-translational modifications.

Through systematic analysis of the phospho-proteomic data, we have identified several kinases that should be studied further to understand their therapeutic implications. PRKDC was predicted by our kinome analysis and its levels were elevated in many Group 3 and WNT samples. While there is no prior evidence for PRKDC's role in medulloblastoma, *in vitro* experiments with various tumor cell lines show that PRKDC promotes MYC stability (An *et al.*, 2008). PRKDC and MYC are known to function together in the presence of DNA damage (Cui *et al.*, 2015). Notably, we have shown that in the presence of high levels of endogenous MYC activity in medulloblastoma cell lines, PRKDC inhibition functions as a radiation sensitizer, but not in cells with low MYC levels. Indeed, our data suggest that the radio sensitizing effects of PRKDC inhibitors may be dependent on MYC status, and furthermore that PRKDC inhibition may serve as radiation sensitizer of MYC-active G3a medulloblastoma in clinical trials.

In conclusion, our results show that quantitative mass spectrometry-based proteomics reveals molecular mechanisms within medulloblastoma subgroups that are not evident through analysis of genome, epigenome, or transcriptome. Protein expression and post-translational modifications represent the functional state of the cancer cells, a reflection of the influence that somatic mutations and other genetic and epigenetic alterations

have to alter the cellular state during progression from normal to a cancerous state. Kinome analysis is a particularly sensitive method to identify specific kinases for therapeutic targeting. Finally, differential modification of proteins through post-translational modifications offers new biomarkers for specific medulloblastoma subtypes. Our integrative exploration of medulloblastoma furthermore provides the clinical and research communities with a wealth of data that may help advance strategies for patient selection and treatments for this devastating disease.
It has come to our attention that a parallel study by Ayrault and colleagues also identifies two subgroups of SHH medulloblastoma. Their pathway analysis also finds SHHa samples to be enriched for DNA replication processes, and SHHb to be enriched for neuronal and neurotransmitter genes.

5.5 Acknowledgements

Funding: U01-CA184898 (E.F., J.P.M., S.L.P.), U54-HD090255 (S.L.P.), R01-CA109467 (S.L.P., J.P.M.), R01-CA121941 (J.P.M.), R01-GM074024 (J.P.M.), U24-CA194107 (J.P.M.), U24-CA210004 (J.P.M.), U01-CA217885 (J.P.M., P.T.), R01-CA154480 (P.T.), U24-CA210986 (S.A.C., M.A.G.), U24-CA210979 (D.R.M.), T32-HL007901 (T.C.A.), 1U54HD090255 (Boston Children's Hospital IDDRC), R01-CA196228 (R.C.S.), R01-CA186241 (R.C.S.), U54-CA209988 (R.C.S.), Swedish Research Council Dnr 2014-323 (F.M), Swedish Society for Medical Research (F.M.), AACR NextGen Grant for Transformative Cancer Research (P.A.N.), American Lebanese Syrian Associated Charities (P.A.N.), St. Jude Children's Research Hospital (P.A.N.), and ICGC.
We thank Jessica Rusert, Robert Wechsler-Reya, Anthony Gitter, Pamela Milani, Miriam Adam, Lukas Chavez, and all members of the Lipton Lab for helpful discussions.

5.6 Author Contributions

Conceptualization, E.F., J.P.M., S.L.P.; Methodology, T.E., T.C.A., M.P.G., S.A.C.; Software, T.E., A.L., K.K., M.P.G., K.C.; Validation, T.C.A., T.E., M.P.G., J.O.L., E.L.M.; Formal Analysis, T.E., M.P.G., T.C.A., C.K.M., D.R., H.Z., J.J., D.R.M., J.S.; Investigation, F.M., E.L.M., L.C.T., J.P.F., T.C.A., C.J.D., E.L.M.; Resources, S.M.P., P.A.N., M.K., A.K.; Data Curation, T.C.A., F.M., K.K., T.E., H.Z.; Writing – Original Draft, T.C.A., T.E., M.P.G., E.F., J.P.M., S.L.P.; Writing – Review & Editing, T.C.A., M.P.G., T.E., E.F., J.P.M., S.L.P., S.M.P., S.A.C., F.M., M.A.G., K.K., P.A.N., M.K., C.J.D., R.C.S., J.O.L.; Visualization, T.E., M.P.G., C.K.M., F.M., K.K., T.C.A.; Supervision, E.F., S.L.P., J.P.M., P.T., S.M.P., S.A.C., M.N., D.R.M., P.M., M.A.G., K.C., R.C.S., J.O.L.; Project Administration, E.F., S.L.P., J.P.M., P.T., T.C.A.; Funding Acquisition, E.F., S.L.P., J.P.M., P.T.

5.7 Methods

5.7.1 **Patient samples**

Primary medulloblastoma patient samples, including FFPE slides, were obtained with informed consent as previously published in Northcott *et al.* (2017). All samples were

de-identified. Tumor samples of 50 mg were freeze-fractured using Covaris cryoPREP CP02 at setting “impact level 4”. The pulverized samples were aliquoted for the downstream methods.

5.7.2 **Cell Lines**

Medulloblastoma cell lines D425, D458 and D556 were a kind gift from Dr. Darell Bigner (Duke University). DAOY and D283 were obtained from American Tissue Culture Collection.

5.7.3 **Proteomic profiling**

The global proteome and phospho-proteome were processed according to adapted protocols from our previous studies (Mertins *et al.*, 2016; Huang *et al.*, 2017). In brief, cryo-pulverized tumor tissue from each patient was lysed at 4°C using 8M urea lysis buffer. Extracted proteins were reduced using dithiothreitol and alkylated with iodoacetamide before digestion using LysC for two hours followed with trypsin overnight. Both digestion steps were performed at a 1:50 enzyme:protein ratio. For relative quantification of the global proteome and phospho-proteome by liquid chromatography tandem mass spectrometry (LC-MS/MS), 400 µg per patient, as measured on protein level (BCA protein concentration determination kit; before digestion) was labeled with 10-plexing tandem mass tags (TMT-10; Thermo Scientific) following the manufactures instructions. All 45 patients were run in 5 total TMT-10 plexes, with each plex including 9 patient samples and an internal reference sample. Samples were assigned to plexes in a semi-randomized manner, such that consensus subgroups were split across TMT-10 plexes (assignments in Table S1). The internal reference sample was composed of equal amounts of peptide material from 40 of the 45 patients, representing all subgroups, and was included in each TMT10-plex to provide a common standard for precise relative quantitation. Isobarically-labeled peptides were combined and fractionated using high-pH reversed phase chromatography into 24 fractions. From each fraction, 5% of the material was evaluated for its proteomic content. The remaining 95% material was combined into 12 fractions which were each enriched for phosphopeptides using immobilized metal affinity chromatography (IMAC)(Mertins *et al.*, 2013). The flow-throughs after IMAC enrichment were collected and further concatenated into 4 fractions that were each enriched using antibodies for acetylated peptides (see Key Resource Table). In parallel to the global proteome, phospho-proteome, and acetylome, an additional 500ug of TMT-labeled peptides per patient were enriched for phosphotyrosine peptides using phospho-tyrosine antibodies (see Key Resource Table) and analyzed as a single fraction on the mass spectrometer. All proteomic based data were collected on a Lumos mass spectrometer (Thermo Fisher Scientific) and the resulting spectra were searched using Spectrum Mill (Agilent, version 12.212). All mass spectra contributing to this study can be downloaded in the original instrument vendor format from the MassIVE online repository.

5.7.4 **Sequencing and DNA Methylation Array Data Collection**

Whole genome sequencing and DNA methylation data sets reported here are previously published in Northcott *et al.* (2017). Consensus subgroup assignments (Figure 1) were provided by DKFZ, and methylation-based subgroups were assigned previously in Northcott *et al.* (2017).

5.7.5 **Western blots**

Cells were cultured as previously described (Weeraratne *et al.*, 2012). For Western blots, 1 million cells were plated in 10 cm dishes and harvested after 48 hours. Proteins were normalized using Pierce BCA Protein Assay Kit and 50 µg protein was loaded per well into NuPAGE Novex gels with Bolt running buffer (see Key Resource Table for specifics). Proteins were transferred using iBlot2 system. Blots were probed for primary antibodies and visualized using the Licor Odyssey system on the Odyssey CLx Infrared Imaging System according to manufacturer's directions. Antibodies dilutions for Western blots are listed on Key Resource Table. Antibody staining was quantified using Image Studio Lite for Western blots. Densitometry was performed by comparing raw densitometry for each antibody to actin on unmodified images.

5.7.6 **Immunofluorescence of cell lines**

Immunocytochemistry was performed as previously reported (Weeraratne *et al.*, 2012) with the following modifications: cells were plated at 30,000 cells/well in 500 µl of media on glass coverslips. Slides were imaged using Zeiss 710 Confocal Microscope in the IDDRC and analyzed using Fiji Image J. All antibodies were used at 1:50 concentration for immunofluorescence.

5.7.7 **Drug dose response assay**

Adherent DAOY cell lines were plated at 2,500 cells per well and suspended D458 were plated at 20,000 cells per well in 96-well plates in 75 µL of culture media. Each sample was plated in sextuplicate. NU7441 was added 24 hours after plating at 4x concentrations in 25 µL. CellTiter-Glo Luminescent Cell Viability Assay was used to measure viability 18 hours after drug addition according to manufacturer's directions in white opaque plates. Luminescence data was measured by EnSight Multimode Plate Reader using Kaleido 1.2 software with 0.1 second measurement time. For irradiation, the cells were exposed to 200 rads of gamma radiation and assayed for cell viability 5 hours later.

5.7.8 **Immunofluorescence staining of FFPE slides**

Antigen retrieval was achieved by pressure-cooking in citrate buffer pH 6 (Sigma) for 10 minutes. Antibodies used for staining are as follows (details in Key Resource Table): rabbit polyclonal c-Myc S62 specific phospho-antibody 1:25 (Zhang *et al.*, 2012) and rabbit polyclonal c-Myc pT58 antibody 1:50 (Applied Biological Material) incubated overnight at 4°C. Secondary antibody was Alexa Fluor 594 1:500 (Invitrogen) and DAPI at 1:5000 (Sigma) incubated for 1 hour at room temperature. ProLong Gold mounting

media (LifeTech) was used and allowed to cure for at least 24 hours. Images were taken with a Hamamatsu digital camera (Japan) mounted on a Leica fluorescence microscope (Wetzlar, Germany) at 40x. Representative images (Figure 4B) were acquired at 63x on a Zeiss LSM 880 laser-scanning confocal microscope (Germany).

5.7.9 **Processing of DNA Methylation Array Data**

We used the *minfi* R library (Aryee *et al.*, 2014) to process the IDAT files into quantile normalized beta values. The probes' beta values were then collapsed to gene symbols using the means of gene-associated probes.

5.7.10 **Processing of genomics data**

This section describes the (re)processing of genomics data (RNA-seq, WGS) for the medulloblastoma cohort that is the subject of this study. All genomics data were available prior to this study and published elsewhere (Northcott *et al.*, 2017). Details about sequencing protocols can be found in the corresponding publications. We decided to re-process all data using the latest best-practice pipelines developed at the Cancer Genome Analysis (CGA) group of the Broad Institute.

5.7.11 **Processing of RNA-seq data**

Bam files were unaligned and converted to FASTQ using Picard (http://broadinstitute.github.io/picard/). All further RNA-seq data processing described below was conducted in FireCloud, a cloud-based computing environment developed and maintained at the Broad Institute. Briefly, RNA-seq reads (50 bp) were aligned to GRCh37 (UCSC hg19) genome assembly using STAR aligner (Dobin *et al.*, 2013). For each sample we assessed QC metrics using RNA-seqC (DeLuca *et al.*, 2012). Transcript expression was quantified as Transcripts Per Kilobase Million (TPM) using RSEM (Li and Dewey, 2011).

5.7.12 **Processing of Affymetrix expression array data**

Expression array data was quantile normalized using the preprocessCore R library (function "normalize.quantiles").

5.7.13 **Processing of WGS data**

The processing of 44 WGS tumor-normal pairs described below was accomplished by chaining together modules implemented in GDAC Firehose (http://gdac.broadinstitute.org/). Somatic variant calling modules were based on the Genome Analysis Tool Kit v3 (GATK) following 'GATK Best Practices' workflows for variant discovery (DePristo *et al.*, 2011).
Paired-end (PE) sequencing reads (100 bp) in FASTQ format were aligned to GRCh37 (UCSC hg19) genome assembly using BWAMem. Base quality score recalibration (BQSR) to correct systematic errors made by the sequencer was done using GATK's 'BaseRecalibrator' and 'PrintReads' programs. Local realignment of insertions and deletions (indels) to correct alignment artifacts was performed with GATK's

'IndelRealigner' program. Cross-sample contamination was assessed by GATK 'ContEst' program.
Mutation calling. Somatic single nucleotide variants (SNVs) and indels were called using Mutec2. Resulting VCF files were annotated and converted to MAF format by Oncotator (Ramos *et al.*, 2015). Oxidative artifacts contributing to SNV calls were assessed by D-ToxoG. SNVs and indels were further filtered for commonly observed germline variants using a panel of normal (PoN)(Costello *et al.*, 2013).
Copy number calling. To derive somatic copy number variants, we used a GATK v4 workflow that comprised four steps: 1) Proportional coverage per read group was calculated by the 'CalculateTargetCoverage' program. 2) A panel of normals (PoN) was created from the normal samples which is meant to encapsulate sequencing noise and common germline variants. The program 'CombineReadCounts' combined proportional read counts from all normal samples into a single file. The PoN file stores information like the median coverage target and was generated using 'CreatePanelOfNormals' using Principle Component Analysis (PCA) to calculate systematic noise. 3) Normalization of tumor coverage by PoN target medians and PoN principle components using 'NormalizeSomaticReadCounts'. Resulting data were log2 transformed. 4) Segmentation of groups of contiguous targets with the same copy ratio using the 'PerformSegmentation' program.
To derive gene-centric copy number variant calls we used GISTIC2.0. Contiguous gene copy number ratios were corrected for diploidy; a value of zero indicates the presence of two gene copies.

5.7.14 **Normalization of proteomics data**

Relative expression data derived from proteomics profiling (proteome, phospho-proteome, acetylome) were separately normalized by sample using robust z-scores (z_r): $z_r = (x - M)/MAD$. In this expression, M is the median expression of the sample; and MAD is the median absolute deviation of the sample. To allow for a better comparison across data types, WNT samples were excluded from these analyses as they lacked all but the proteomic data, and there was insufficient high-quality tissue to perform the additional assays.

5.7.15 **Normalization of RNA-seq expression data**

TPM values were first converted to relative scale by median-row normalization. Resulting ratios were then transformed to robust z-scores as described above.

5.7.16 **Quantification of Immunofluorescence staining**

Mean immunofluorescence densities were generated using OpenLab 5.5 software (Improvision) by using the ROI tool to quantify 50 to 100 nuclei in 3 to 4 fields of view at 40x magnification. Scatter and bar plots were generated using GraphPad Prism. Error bars represent SEM and significance reported is from two-tailed unpaired t-tests.

5.7.17 **Quantification of Western Blots**

Raw densitometry for each antibody was normalized to actin using Image Studio Lite. All quantification was performed in at least triplicate on original unmodified blots. Normalized signals were compared between cell lines by arbitrarily normalizing samples against the DAOY line. One-way ANOVA with a Dunnett multiple comparison test was performed to compare the normalized signal for each antibody across the five cells lines using GraphPad Prism 5. Error bars represent mean normalized signal ± SEM.

5.7.18 **Dose response curve**

Luminescence for each well was measured and the mean signal for each condition ± SEM plotted. Because adherent (DAOY) and suspended (D458) cells were compared, we normalized mean luminescence to 100% for the non-irradiated treatment group for each cell line. A dose-response curve and the IC50 were determined by non-linear regression with GraphPad Prism.

5.7.19 **Survival Curves**

Group 3 samples were divided into G3a and G3b groups. G3a consists of the samples in our cohort assigned to G3a and the samples from the Cho *et al.* (2011) cohort predicted to be G3a. The same scheme applies to G3b. Kaplan-Meier curves were generated for each of these groups, for both overall survival (OS) and progression-free survival (PFS). All Kaplan-Meier curves and log rank test p values were generated with the lifelines Python package (Davidson-Pilon *et al.*, 2017). Samples whose death event were not observed within the 10 years following initial diagnosis were right-censored for Figure 4D, and the full outcome follow up Kaplan-Meier plots are available in Figure S5E.

5.7.20 **Group 3 Cohort Expansion**

To expand our PFS and OS analysis for G3a and G3b we used the c1 and c5 samples from the Cho *et al.* (2011) cohort which are most like Group 3. Since only array data were available for the Cho samples, we also used array data for our current Group 3 cohort to better normalize the two data sets.
Following the methods in Cho *et al.* (2011), we projected the expression data for both cohorts into “gene set space” using a single sample version of GSEA (ssGSEA), using the Hallmarks (H), Curated Gene Sets (C2), Motif Gene Sets (C3), and Oncogenic Gene Sets (C6) collections from the MSigDB. Using the G3a, G3b labels for the current cohort, we determined the 10 most differentially enriched gene sets for each subtype (Figure S5D, main text) according to the Information Coefficient (IC) as defined in Kim *et al.* (2016). These top sets were used as features to train a Bayesian cumulative log-odds predictor as previously described (Tamayo *et al.*, 2011). The predictor was applied to the projected Cho *et al.* (2011) c1 and c5 samples to assign G3a and G3b labels and those labels were used to construct a combined cohort PFS and OS analysis. We used the Cho *et al.* (2011) labels for the current cohort as given in Figure 2B of the main text to construct a combined c1/c5 labeled cohort for PFS and OS analysis. This analysis

did not include sample MB166 and MB278 from G3b which were labeled c4. The c1/c5 labeling of the current cohort used the classifier developed for Cho *et al.* (2011), with two minor modifications: (1) we restricted the analysis to genes contained in both data sets, and (2) the IPA_KCNIP2_DN gene set used as a feature for c5 in Cho *et al.* (2011) was replaced by BIOCARTA_PS1_PATHWAY.

5.7.21 **Glutamate Pathway modeling**

For Figure 3C, the gene sets were collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG). Protein levels were summarized using the median log2-normalized protein level for a gene for all samples in a group (SHHa or SHHb). When multiple genes are in a gene set, the average of those gene values is shown. We only included genes with data for at least 75% of samples from both groups. A summary of all the genes included in Figure 3C is included in Table S4.

5.7.22 **Consensus Clustering**

We used the ConsensusClusterPlus R library (Wilkerson and Hayes, 2010) to perform consensus clustering on our data sets. For each data set we varied k, the number of clusters, from 2 to 9 and ran the algorithm with 1,000 subsamples for all combinations of two clustering methods (hierarchical clustering and k-means) and three distance metrics (Euclidean, Spearman, Pearson). All consensus heatmaps are available in the Supplement (Figure S1). Consensus clustering of proteomic data indicated that this data type produced highly consistent clusters with k = 5. Runs with k < 5 showed a clear substructure within the clusters identified. Notably, for all clusterings with k < 5, the SHHb samples clustered with Group 4 samples rather than the SHHa group. Runs with k > 5 largely included very small clusters or samples that showed pronounced cluster promiscuity. We also calculated PAC (proportion of ambiguously clustered pairs) scores (Senbabaoglu, Michailidis and Li, 2014) for our consensus clustering runs (Figure S1) and found confirmation that, for proteomics data, k = 5 was the optimal number of clusters when trying to avoid clusters with three or fewer samples. Using the PAC scores we also determined that hierarchical clustering and Pearson-based distance was the best parameter setting for this *k*. With these settings, one sample, MB247, frequently shifted between G3a and G3b. We decided to assign it to G3a based on results from other unsupervised methods such as PCA.

5.7.23 **Dimensionality Reduction**

Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE) were employed to display our high-dimensional data sets in two dimensions. We used the R function “princomp” in the “stats” library for PCA, and the “Rtsne” package for tSNE. Running Rtsne, we used the function’s default parameters along with the maximum perplexity the program would allow for (perplexity = (number of samples – 1) / 3).

5.7.24 **Differential analysis**

We used two types of analyses to find differential features in our proteomic data sets. One strategy sought to extract features specific to a single disease group in comparison to the rest of the samples; the other approach selected features that differ between pairs of groups. We performed univariate t-tests (using “t.test” in R) to find features between each disease group and all the other samples. For pairwise comparisons across all groups, we performed ANOVA along with R’s “TukeyHSD” method to calculate p values. P values from both methods were corrected for multiple hypothesis testing using the Benjamini-Hochberg/FDR method implemented in R's “p.adjust” function. Due to concerns associated with the small number of samples (in particular for comparisons involving SHH samples), we aimed to focus the comparison on the most homogeneous groups possible. We therefore withheld sample MB136 from these statistical tests, as it clustered stably with SHHb in the proteomic data but showed similarities with SHHa in other data types.
We are aware that the small sizes of our cohort's disease groups present challenges in the selection of statistical tests and estimation of significance of results. We selected the methods mentioned above as they were able to identify differential features with reasonable confidence and allowed us to prioritize features to feed into downstream analysis methods (e.g. Omics Integrator). These downstream methods served as a second filter and none of our conclusions are directly based on the results of these tests. For statistics on individual univariate comparisons we used a Mann Whitney U test (these instances are indicated as such in the main text). Across the different testing methods, we considered FDR-adjusted p values of less than 0.05 as significant, but chose more stringent thresholds for certain analyses to limit the number of features.

5.7.25 **Global Correlation Analysis**

To correlate proteomic and RNAseq data, we considered only those data with no missing values and identified matching proteins and transcripts based on gene symbols. In doing so, splice isoforms were collapsed by their means. For each sample, we then calculated the Spearman correlation for the 8,674 proteins and their corresponding mRNAs.

5.7.26 **Functional Annotation of Data Sets**

To project data sets into the space of gene sets representing functional information, we performed hypergeometric overlap tests (background size context dependent) and used an in-house implementation of single sample Gene Set Enrichment Analysis (ssGSEA)(Subramanian *et al.*, 2005; Barbie *et al.*, 2009). For the latter, we used the following parameter settings: rank-based sample scoring (“sample.norm.type”: ranks), area under curve statistic to calculate scores (“statistic”: area.under.RES), default weight (0.75), 500 permutations, z-score pre-processing (“correl.type”: z.score), and normalized enrichment score output (“NES”).
We used the gene sets available in the Molecular Signature Database (MSigDB v6.0)(Liberzon, 2014; Liberzon *et al.*, 2015), primarily focusing on canonical pathways (C2CP), GO terms (C5), and Hallmark sets (H). P values from these methods were

corrected for multiple hypothesis correction using the Benjamini-Hochberg/FDR correction implemented in R's “p.adjust” function. FDR-adjusted p values less than 0.05 were considered significant.

5.7.27 **Integrative Network Analysis**

We used the latest version of Omics Integrator 2 (OI2 v2.24, https://github.com/fraenkel-lab/OmicsIntegrator2) to construct disease-focused, integrative networks using proteomic and genomic data (Tuncbag *et al.*, 2016). Omics Integrator begins by mapping a set of proteins of interest onto the nodes in a network of physical interactions ("interactome") among proteins. The nodes are assigned “prizes”, reflecting their importance (see below for details). The interactions were derived from public databases using the iRefIndex v14 collection of interactions (Razick, Magklaras and Donaldson, 2008). Each interaction is associated with a cost that is lowest for the most reliable interactions (calculated as 1 minus the edge score provided by iRefIndex). We also added previously published site-protein (cost: 0.25) and site-kinase (as published in PhosphoSitePlus, (Hornbeck *et al.*, 2015)) edges (cost: 0.4) to the interactome to allow the algorithm to find site-kinase interactions in the solution that are not included in the iRefIndex interactome. The algorithm seeks to identify subnetworks that contain many disease relevant nodes (based on prizes) while still avoiding using too many low confidence edges (based on costs). For more details about this method, see Tuncbag *et al.* (2016).

We tailored the selection and definition of prizes to the type of network we sought to detect. We created networks focusing on the differences between G3a and G3b (“G3 network”), as well as SHHa and SHHb (“SHH network”). Input nodes were selected based on differential protein or peptide levels for the relevant sample group comparison in each network. For the G3 network, proteomic features needed to pass the same FDR threshold (FDR < 0.05) in each data set. For the SHH network, proteins and phosphopeptides in pSTY needed to pass a more stringent threshold (FDR < 0.005) due to the large number of differential features, while the threshold for acetylated peptides and phospho-tyrosine peptides was set to 0.05. Prizes for nodes based on proteomic data sets were calculated as the absolute value of the log2-based fold changes (based on the sample groups’ means) from pairwise comparisons. Genomic events were each assigned a fixed prize at an arbitrary value of 2.5 and each CNA and mutation found in at least one sample in the relevant group was given a prize. This value was used as we wanted to lend strong weight to genomic alterations, but not have them outweigh strong proteomic signals. We therefore opted for a value between the median and the 3rd quartile in the proteomic prize distributions.

Data for protein and acetylation events were collapsed to gene symbols of the relevant proteins, keeping the highest prized features as inputs for OI2. Due to the large number of differential phosphosites, these sites were treated as individual nodes in the OI2 network formulation and the required edges were added to the interactome (site-protein edges, cost: 0.25), splitting doubly phosphorylated sites into individual sites.

We ran OI2 with the selected disease-associated nodes and their prizes (see Table S6) as inputs on all combinations of the following parameters: Gs = (50k, 100k, 500k, 1M, 2.5M, 5M, 10M), Bs = (0.5, 1, 10), and Ws = (1, 3, 6, 10). We evaluated these networks to find a parameter set (W = 1, B = 1, G = 2.5M) that produced a network that was not

dominated by hub-nodes, and had a reasonable balance of input nodes and those added by the algorithm for connectivity. Using this parameter set, we performed additional calculations to assess the robustness and specificity of nodes in the network. To determine the robustness of nodes, we added Gaussian noise (standard deviation: 0.05, mean: 0) to the edges in the interactome before each of 200 runs. We then calculated the robustness score for each node as the fraction of times a node appeared across the networks. To calculate specificity, we assigned our prizes to randomly selected, degree-matched nodes in the interactome before each of 200 runs. The specificity score was then calculated as the one minus the fraction of times a node appeared across these networks. (More details on and rationale around these scores can be found in (Tuncbag *et al.*, 2016). For downstream analysis, we removed nodes from the OI2 output that were neither specific nor robust (threshold: 0.75). To link the network clusters to functional annotations we used hypergeometric tests as described above.
We used Cytoscape v3.3 (Shannon *et al.*, 2003) for visualizations of our networks. Nodes received log2-fold change-based colors based on the source of their prizes, and shapes according to the data type they represent (protein, genomic lesion, phosphosite, acetylated protein – see figure legend for details). Nodes in the network that did not receive a prize but were still included in the network received the log2-FC based color of the corresponding proteomic value if data was available (or were colored grey otherwise). For the final display items, we rearranged nodes primarily based on pathway associations rather than the clustering of the network used for the initial analysis.

5.7.28 **Kinome Analysis**

This analysis sought to identify the kinases potentially responsible for phosphorylating phosphopeptides of interest. We first selected phosphopeptides that were either specific to one of the clusters found in proteomic data or significantly different between a pair of clusters (FDR < 0.01). We then used two different methods to identify upstream kinases:
(1) PhosphoSitePlus: Cell Signaling Technologies' PhosphoSitePlus database (Hornbeck *et al.*, 2015) is a curated resource of experimentally determined peptide targets of specific kinases (downloaded February 2017). We searched this database for matches to phosphopeptides identified in our analyses.
(2) Scansite: the Scansite platform (Obenauer, Cantley and Yaffe, 2003) contains sequence specificity motifs for kinases that were derived from oriented peptide library screens. For this analysis, we used 15 amino acid long peptide sequences (phosphosite ± 7 amino acids) surrounding our phosphorylation events as input into Scansite (Scansite 3 Web Service) and used its most stringent setting (high stringency) to get only the best motif matches. For results to be reported with this setting, a motif's score for a peptide sequence needs to be in the top percentile of scores from an empirical score distribution that is based on all potentially matching sites in the vertebrate proteome.
For each of these two methods, we then collapsed all peptides matched to a kinase (median across matching peptides) and calculated the Spearman correlation of this 'peptide-profile' to the kinase's protein and/or phosphopeptide-levels (if available). If the correlation with the kinase's protein levels or any of its phosphopeptides was higher

than 0.4 (or less than -0.4 for phosphosites) it was labeled as correlating (or anticorrelating).
We checked DrugBank.ca (Law *et al.*, 2014) for drugs that are known to target any of the kinases nominated by our analysis. Whenever this information was available, we also annotated whether any FDA approved products were listed for a drug. We provide a full table of all kinase-drug matches in the Supplement (Table S5).
To identify disease-related kinases that we could not associate with peptides (as in the analyses described above) due to a lack of supporting data (no kinase motifs or known sites), we also extracted significantly differential kinases from all our proteomic data sets (FDR < 0.005). These kinases are indicated in Table S2.
We also used an R implementation (Wagih *et al.*, 2016) of the motif-x algorithm (Chou and Schwartz, 2011) to discover motifs. We performed the analysis for 15 amino acid-long peptide sequence windows around serine (S), threonine (T), and tyrosine (Y) residues in our phospho-proteomics data sets. As foreground sets we used peptides up/down in only a single proteomic group of samples and ran the algorithm for all our groups at various significance thresholds with all S, T, Y-centered peptides in our data set as background. The most specific motifs found by this approach were “SP” and “TP”.

5.8 References

An, J. *et al.* (2008) “DNA-dependent protein kinase catalytic subunit modulates the stability of c-Myc oncoprotein,” *Mol Cancer*. 2008/04/23, 7, p. 32. doi: 10.1186/1476-4598-7-32.
Arcella, A. *et al.* (2005) “Pharmacological blockade of group II metabotropic glutamate receptors reduces the growth of glioma cells in vivo,” *Neuro Oncol*. 2005/08/02, 7(3), pp. 236–245. doi: 10.1215/S1152851704000961.
Archer, T. C., Jin, J. and Casey, E. S. (2011) “Interaction of Sox1, Sox2, Sox3 and Oct4 during primary neurogenesis,” *Dev Biol*. 2010/12/15, 350(2), pp. 429–440. doi: 10.1016/j.ydbio.2010.12.013.
Archer, T. C., Mahoney, E. L. and Pomeroy, S. L. (2017) “Medulloblastoma: Molecular Classification-Based Personal Therapeutics,” *Neurotherapeutics*. 2017/04/08, 14(2), pp. 265–273. doi: 10.1007/s13311-017-0526-y.
Arnold, H. K. *et al.* (2009) “The Axin1 scaffold protein promotes formation of a degradation complex for c-Myc,” *EMBO J*. 2009/01/10, 28(5), pp. 500–512. doi: 10.1038/emboj.2008.279.
Aryee, M. J. *et al.* (2014) “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays,” *Bioinformatics*. 2014/01/31, 30(10), pp. 1363–1369. doi: 10.1093/bioinformatics/btu049.
Barbie, D. A. *et al.* (2009) “Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1,” *Nature*. 2009/10/23, 462(7269), pp. 108–112. doi: 10.1038/nature08460.
Bigner, D. D. *et al.* (1990) “Amplification of the c-myc Gene in Human Medulloblastoma Cell Lines and Xenografts,” *Cancer Research*.
Brown, E. J. and Frazier, W. A. (2001) “Integrin-associated protein (CD47) and its ligands,” *Trends Cell Biol*. 2001/04/18, 11(3), pp. 130–135.
Cavalli, F. M. G. *et al.* (2017) “Intertumoral Heterogeneity within Medulloblastoma

Subgroups," *Cancer Cell*. 2017/06/14, 31(6), p. 737–754 e6. doi: 10.1016/j.ccell.2017.05.005.
Chen, Y. and Olopade, O. I. (2008) "MYC in breast tumor progression," *Expert Rev Anticancer Ther*. 2008/10/18, 8(10), pp. 1689–1698. doi: 10.1586/14737140.8.10.1689.
Cho, Y.-J. *et al.* (2011) "Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29(11), pp. 1424–30. doi: 10.1200/JCO.2010.28.5148.
Chou, M. F. and Schwartz, D. (2011) "Biological sequence motif discovery using motif-x," *Curr Protoc Bioinformatics*. 2011/09/09, Chapter 13, p. Unit 13 15-24. doi: 10.1002/0471250953.bi1315s35.
Cibulskis, K. *et al.* (2013) "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nat Biotechnol*. 2013/02/12, 31(3), pp. 213–219. doi: 10.1038/nbt.2514.
Ciszewski, W. M. *et al.* (2014) "DNA-PK inhibition by NU7441 sensitizes breast cancer cells to ionizing radiation and doxorubicin," *Breast Cancer Research and Treatment*. doi: 10.1007/s10549-013-2785-6.
Costello, M. *et al.* (2013) "Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation," *Nucleic Acids Res*. 2013/01/11, 41(6), p. e67. doi: 10.1093/nar/gks1443.
Cui, F. *et al.* (2015) "The involvement of c-Myc in the DNA double-strand break repair via regulating radiation-induced phosphorylation of ATM and DNA-PKcs activity," *Molecular and Cellular Biochemistry*. doi: 10.1007/s11010-015-2422-2.
Davidson-Pilon, C. *et al.* (2017) "Lifelines survival analysis in Python."
DeLuca, D. S. *et al.* (2012) "RNA-seqC: RNA-seq metrics for quality control and process optimization," *Bioinformatics*. 2012/04/28, 28(11), pp. 1530–1532. doi: 10.1093/bioinformatics/bts196.
DePristo, M. A. *et al.* (2011) "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nat Genet*. 2011/04/12, 43(5), pp. 491–498. doi: 10.1038/ng.806.
Diaz, R. J. *et al.* (2015) "Mechanism of action and therapeutic efficacy of Aurora kinase B inhibition in MYC overexpressing medulloblastoma.," *Oncotarget*. doi: 10.18632/oncotarget.3245.
Dobin, A. *et al.* (2013) "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*. 2012/10/30, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.
Edwards, N. J. *et al.* (2015) "The CPTAC Data Portal: A Resource for Cancer Proteomics Research," *J Proteome Res*. 2015/04/16, 14(6), pp. 2707–2713. doi: 10.1021/pr501254j.
Farrell, A. S. and Sears, R. C. (2014) "MYC degradation," *Cold Spring Harbor Perspectives in Medicine*. doi: 10.1101/cshperspect.a014365.
Gholamin, S. *et al.* (2017) "Disrupting the CD47-SIRPalpha anti-phagocytic axis by a humanized anti-CD47 antibody is an efficacious treatment for malignant pediatric brain tumors," *Sci Transl Med*. 2017/03/17, 9(381). doi: 10.1126/scitranslmed.aaf2968.
Gibson, P. *et al.* (2010) "Subtypes of medulloblastoma have distinct developmental origins," *Nature*. 2010/12/15, 468(7327), pp. 1095–1099. doi: 10.1038/nature09587.

Hanaford, A. R. *et al.* (2016) “DiSCoVERing innovative therapies for rare tumors: Combining genetically accurate disease models with in silico analysis to identify novel therapeutic targets,” *Clinical Cancer Research*, 22(15). doi: 10.1158/1078-0432.CCR-15-3011.
Hornbeck, P. V *et al.* (2015) “PhosphoSitePlus, 2014: mutations, PTMs and recalibrations,” *Nucleic Acids Res*. 2014/12/18, 43(Database issue), pp. D512-20. doi: 10.1093/nar/gku1267.
Horvath, S. *et al.* (2015) “The cerebellum ages slowly according to the epigenetic clock,” *Aging (Albany NY)*. 2015/05/23, 7(5), pp. 294–306.
Hovestadt, V. *et al.* (2014) “Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing,” *Nature*. 2014/05/23, 510(7506), pp. 537–541. doi: 10.1038/nature13268.
Huang, K. L. *et al.* (2017) “Proteogenomic integration reveals therapeutic targets in breast cancer xenografts,” *Nat Commun*. 2017/03/30, 8, p. 14864. doi: 10.1038/ncomms14864.
Jaiswal, S. *et al.* (2009) “CD47 is upregulated on circulating hematopoietic stem cells and leukemia cells to avoid phagocytosis,” *Cell*. 2009/07/28, 138(2), pp. 271–285. doi: 10.1016/j.cell.2009.05.046.
Janghorban, M. *et al.* (2014) “Targeting c-MYC by antagonizing PP2A inhibitors in breast cancer,” *Proc Natl Acad Sci U S A*. 2014/06/14, 111(25), pp. 9157–9162. doi: 10.1073/pnas.1317630111.
Kahn, M. (2014) “Can we safely target the WNT pathway?,” *Nat Rev Drug Discov*. 2014/07/02, 13(7), pp. 513–532. doi: 10.1038/nrd4233.
Kanehisa, M. *et al.* (2017) “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Res*. 2016/12/03, 45(D1), pp. D353–D361. doi: 10.1093/nar/gkw1092.
Karolchik, D. *et al.* (2004) “The UCSC Table Browser data retrieval tool,” *Nucleic Acids Res*. 2003/12/19, 32(Database issue), pp. D493-6. doi: 10.1093/nar/gkh103.
Kim, J. W. *et al.* (2016) “Characterizing genomic alterations in cancer by complementary functional associations,” *Nat Biotechnol*. 2016/04/19, 34(5), pp. 539–546. doi: 10.1038/nbt.3527.
Kool, M. *et al.* (2012) “Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas,” *Acta Neuropathol*. 2012/02/24, 123(4), pp. 473–484. doi: 10.1007/s00401-012-0958-8.
Kool, M. *et al.* (2014) “Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition,” *Cancer Cell*. 2014/03/22, 25(3), pp. 393–405. doi: 10.1016/j.ccr.2014.02.004.
Law, V. *et al.* (2014) “DrugBank 4.0: shedding new light on drug metabolism,” *Nucleic Acids Res*. 2013/11/10, 42(Database issue), pp. D1091-7. doi: 10.1093/nar/gkt1068.
Lawrence, R. T. *et al.* (2015) “The proteomic landscape of triple-negative breast cancer,” *Cell Rep*. 2015/04/22, 11(4), pp. 630–644. doi: 10.1016/j.celrep.2015.03.050.
Li, B. and Dewey, C. N. (2011) “RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome,” *BMC Bioinformatics*. 2011/08/06, 12, p. 323. doi: 10.1186/1471-2105-12-323.
Li, H. and Durbin, R. (2009) “Fast and accurate short read alignment with Burrows-

Wheeler transform," *Bioinformatics*. 2009/05/20, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.
Liberzon, A. (2014) "A description of the Molecular Signatures Database (MSigDB) Web site," *Methods Mol Biol*. 2014/04/20, 1150, pp. 153–160. doi: 10.1007/978-1-4939-0512-6_9.
Liberzon, A. *et al.* (2015) "The Molecular Signatures Database (MSigDB) hallmark gene set collection," *Cell Syst*. 2016/01/16, 1(6), pp. 417–425. doi: 10.1016/j.cels.2015.12.004.
Louis, D. N., Perry, A., *et al.* (2016) "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary," *Acta Neuropathol*. 2016/05/10, 131(6), pp. 803–820. doi: 10.1007/s00401-016-1545-1.
Louis, D. N., Ohgaki, H., *et al.* (2016) *WHO Classification of Tumours of the Central Nervous System*. Revised 4t. Geneva, Switzerland: WHO Press.
Lu, A. T. *et al.* (2016) "Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum," *Nat Commun*. 2016/02/03, 7, p. 10561. doi: 10.1038/ncomms10561.
Ma, Y. *et al.* (2004) "A biochemically defined system for mammalian nonhomologous DNA end joining," *Mol Cell*. 2004/12/03, 16(5), pp. 701–713. doi: 10.1016/j.molcel.2004.11.017.
McCubrey, J. A. *et al.* (2014) "GSK-3 as potential target for therapeutic intervention in cancer," *Oncotarget*. 2014/06/17, 5(10), pp. 2881–2911. doi: 10.18632/oncotarget.2037.
Mermel, C. H. *et al.* (2011) "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biol*. 2011/04/30, 12(4), p. R41. doi: 10.1186/gb-2011-12-4-r41.
Mertins, P. *et al.* (2013) "Integrated proteomic analysis of post-translational modifications by serial enrichment," *Nat Methods*. 2013/06/12, 10(7), pp. 634–637. doi: 10.1038/nmeth.2518.
Mertins, P. *et al.* (2016) "Proteogenomics connects somatic mutations to signalling in breast cancer," *Nature*. 2016/06/03, 534(7605), pp. 55–62. doi: 10.1038/nature18003.
Morrish, F. and Hockenbery, D. (2014) "MYC and mitochondrial biogenesis," *Cold Spring Harb Perspect Med*. 2014/05/03, 4(5). doi: 10.1101/cshperspect.a014225.
Northcott, P. A. *et al.* (2011) "Pediatric and adult sonic hedgehog medulloblastomas are clinically and molecularly distinct," *Acta Neuropathol*. 2011/06/18, 122(2), pp. 231–240. doi: 10.1007/s00401-011-0846-7.
Northcott, P. A. *et al.* (2012) "Medulloblastomics: the end of the beginning," *Nat Rev Cancer*. 2012/11/24, 12(12), pp. 818–834. doi: 10.1038/nrc3410.
Northcott, P. A. *et al.* (2017) "The whole-genome landscape of medulloblastoma subtypes," *Nature*, 547(7663). doi: 10.1038/nature22973.
Obenauer, J. C., Cantley, L. C. and Yaffe, M. B. (2003) "Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs," *Nucleic Acids Res*. 2003/06/26, 31(13), pp. 3635–3641.
Pei, Y. *et al.* (2016) "HDAC and PI3K Antagonists Cooperate to Inhibit Growth of MYC-Driven Medulloblastoma," *Cancer Cell*. 2016/03/16, 29(3), pp. 311–323. doi: 10.1016/j.ccell.2016.02.011.
Pfister, S. *et al.* (2009) "Outcome prediction in pediatric medulloblastoma based on DNA copy-number aberrations of chromosomes 6q and 17q and the MYC and MYCN loci," *J*

Clin Oncol. 2009/03/04, 27(10), pp. 1627–1636. doi: 10.1200/JCO.2008.17.9432.
Pozo, K. and Bibb, J. A. (2016) "The Emerging Role of Cdk5 in Cancer," *Trends Cancer*. 2016/12/06, 2(10), pp. 606–618. doi: 10.1016/j.trecan.2016.09.001.
Pugh, T. J. *et al.* (2012) "Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations," *Nature*. 2012/07/24, 488(7409), pp. 106–110. doi: 10.1038/nature11329.
Ramos, A. H. *et al.* (2015) "Oncotator: cancer variant annotation tool," *Hum Mutat*. 2015/02/24, 36(4), pp. E2423-9. doi: 10.1002/humu.22771.
Rauniyar, N. and Yates 3rd, J. R. (2014) "Isobaric labeling-based relative quantification in shotgun proteomics," *J Proteome Res*. 2014/10/23, 13(12), pp. 5293–5309. doi: 10.1021/pr500880b.
Razick, S., Magklaras, G. and Donaldson, I. M. (2008) "iRefIndex: a consolidated protein interaction database with provenance," *BMC Bioinformatics*. 2008/10/01, 9, p. 405. doi: 10.1186/1471-2105-9-405.
Robinson, G. *et al.* (2012) "Novel mutations target distinct subgroups of medulloblastoma," *Nature*. 2012/06/23, 488(7409), pp. 43–48. doi: 10.1038/nature11213.
Robinson, G. W. *et al.* (2015) "Vismodegib Exerts Targeted Efficacy Against Recurrent Sonic Hedgehog-Subgroup Medulloblastoma: Results From Phase II Pediatric Brain Tumor Consortium Studies PBTC-025B and PBTC-032," *J Clin Oncol*. 2015/07/15, 33(24), pp. 2646–2654. doi: 10.1200/JCO.2014.60.1591.
Schwalbe, E. C. *et al.* (2017) "Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study," *Lancet Oncol*. 2017/05/27, 18(7), pp. 958–971. doi: 10.1016/S1470-2045(17)30243-7.
Senbabaoglu, Y., Michailidis, G. and Li, J. Z. (2014) "Critical limitations of consensus clustering in class discovery," *Sci Rep*. 2014/08/28, 4, p. 6207. doi: 10.1038/srep06207.
Shannon, P. *et al.* (2003) "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res*. 2003/11/05, 13(11), pp. 2498–2504. doi: 10.1101/gr.1239303.
Staal, J. A. *et al.* (2015) "Proteomic profiling of high risk medulloblastoma reveals functional biology," *Oncotarget*. 2015/05/15, 6(16), pp. 14584–14595. doi: 10.18632/oncotarget.3927.
Subramanian, A. *et al.* (2005) "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A*. 2005/10/04, 102(43), pp. 15545–15550. doi: 10.1073/pnas.0506580102.
Sunada, S. *et al.* (2016) "Nontoxic concentration of DNA-PK inhibitor NU7441 radio-sensitizes lung tumor cells with little effect on double strand break repair," *Cancer Science*. doi: 10.1111/cas.12998.
Tamayo, P. *et al.* (2011) "Predicting relapse in patients with medulloblastoma by integrating evidence from clinical and genomic features.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29(11), pp. 1415–23. doi: 10.1200/JCO.2010.28.1675.
Taylor, M. D. *et al.* (2012) "Molecular subgroups of medulloblastoma: the current consensus," *Acta Neuropathol*. 2011/12/03, 123(4), pp. 465–472. doi: 10.1007/s00401-011-0922-z.
Tuncbag, N. *et al.* (2016) "Network-Based Interpretation of Diverse High-Throughput

Datasets through the Omics Integrator Software Package," *PLoS Comput Biol*. 2016/04/21, 12(4), p. e1004879. doi: 10.1371/journal.pcbi.1004879.
Wagih, O. *et al.* (2016) "Uncovering Phosphorylation-Based Specificities through Functional Interaction Networks," *Mol Cell Proteomics*. 2015/11/18, 15(1), pp. 236–245. doi: 10.1074/mcp.M115.052357.
Wang, X. *et al.* (2011) "Phosphorylation regulates c-Myc's oncogenic activity in the mammary gland," *Cancer Res*. 2011/01/27, 71(3), pp. 925–936. doi: 10.1158/0008-5472.CAN-10-1032.
Weeraratne, S. D. *et al.* (2012) "Pleiotropic effects of miR-183~96~182 converge to regulate cell survival, proliferation and migration in medulloblastoma ," *Acta Neuropathol.* . doi: 10.1007/s00401-012-0969-5.
Wilkerson, M. D. and Hayes, D. N. (2010) "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking," *Bioinformatics*. 2010/04/30, 26(12), pp. 1572–1573. doi: 10.1093/bioinformatics/btq170.
Zhang, B. *et al.* (2014) "Proteogenomic characterization of human colon and rectal cancer," *Nature*. 2014/07/22, 513(7518), pp. 382–387. doi: 10.1038/nature13438.
Zhang, H. *et al.* (2016) "Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer," *Cell*. 2016/07/04, 166(3), pp. 755–765. doi: 10.1016/j.cell.2016.05.069.
Zhang, H., Meltzer, P. and Davis, S. (2013) "RCircos: an R package for Circos 2D track plots," *BMC Bioinformatics*. 2013/08/14, 14, p. 244. doi: 10.1186/1471-2105-14-244.
Zhang, X. *et al.* (2012) "Mechanistic insight into Myc stabilization in breast cancer involving aberrant Axin1 expression," *Proc Natl Acad Sci U S A*. 2011/08/03, 109(8), pp. 2790–2795. doi: 10.1073/pnas.1100764108.

Chapter 6 Developmental Basis of SHH Medulloblastoma Heterogeneity

Contributing Authors (alphabetical order): Joseph Benetatos, Jennifer A. Cotter, Shawn M. Davidson, Laura Donovan, Ernest Fraenkel, Julie Galindo, Livia Garzia, **Maxwell P. Gold**, Andrey Korshunov, Andrew M. Masteller, Jill P. Mesirov, Winnie Ong, Noel R. Park, Scott L. Pomeroy, Raul A. Saurez, Michael D. Taylor, Maria C. Vladoiu, Adam D. Walker, Robert Wechsler-Reya,

This section is the latest draft of the manuscript describing the developmental basis of SHH medulloblastoma. My primary contributions to this work were leading the data analysis and hypothesis generation for the project. I relied significantly on many talented associates and collaborators for wet lab experiments and data interpretation.

6.1 Abstract

Medulloblastoma (MB) is one of the most common malignant pediatric brain tumors. The sonic hedgehog (SHH) subtype accounts for 30% of MB cases and likely arises from mutations in granule cell precursors (GCPs), neuronal progenitors of the cerebellar cortex that differentiate into granule neurons. SHH MB is extremely heterogeneous, but it is unknown whether this heterogeneity relates to the tumors' developmental origins. To investigate this question, we performed single-nucleus RNA-Sequencing on seven highly differentiated SHH MB with extensively nodular histology and observed malignant cells resembling each stage of granule neuron development. Using novel computational approaches, we connected these results to published datasets and found that established molecular subtypes of SHH MB are enriched for specific developmental cell types. Additionally, some genomic copy number variations are associated with specific developmental stages, and we observed distinct metabolic and histological profiles for tumors containing cells resembling late-stage granule neurons. This work details computational and experimental approaches that can be repurposed for analysis of tumor cell differentiation in other cancers.

6.2 Introduction

Medulloblastoma (MB) is one of the most common malignant pediatric brain tumors. The standard treatment regimen of surgical resection, radiation, and chemotherapy has led to favorable short-term outcomes in aggregate (Gajjar et al., 2021; Kool et al., 2012), but unfortunately these therapies can cause neurological side effects and increased risks of secondary cancers (Olivier et al., 2019; Salloum et al., 2019). Thus, there is an urgent need for more targeted, less toxic therapies, which requires a better understanding of the heterogeneity within and between MB tumors (Cavalli et al., 2017; Hovestadt et al., 2019; Riemondy et al., 2022; Vladoiu et al., 2019).

The World Health Organization recognizes both histological and molecular heterogeneity in MB (Louis et al., 2021; Orr, 2020). The four primary histological categories are classic, large cell anaplastic (LCA), desmoplastic/nodular (DNMB), and medulloblastomas with extensive nodularity (MBEN). DNMB histology is characterized by tightly packed cells interrupted by nodules filled with a lower density of differentiated neuron-like cells. Tumors with widespread nodularity are designated as MBENs.

In addition to histological heterogeneity, there are four consensus molecular subgroups recognized by the MB research community (Taylor et al., 2012): WNT, SHH, Group 3, and Group 4. SHH MBs represent 30% of cases and have an overactive sonic hedgehog pathway caused by germline or acquired mutations (Northcott et al., 2012). Granule cell precursors (GCPs) of the cerebellum are the proposed cell of origin for these tumors (Kim et al., 2003; Schüller et al., 2008; Yang et al., 2008). During normal development, the GCPs proliferate in response to SHH in the external granule layer (EGL) (Wallace, 1999; Wechsler-Reya and Scott, 1999) before differentiating into granule neurons (GN) (Komuro et al., 2001), which then migrate to the internal granule layer (IGL) (Komuro and Rakic, 1995; Komuro and Yacubova, 2003; Rakic, 1971).

Many groups have characterized the molecular heterogeneity of SHH MBs. Analysis of methylation and transcriptional data revealed four consensus subtypes: SHH-1 (β), SHH-2 (ɣ), SHH-3 (α), and SHH-4 (δ) (Cavalli et al., 2017; Garcia-Lopez et al., 2021). Additionally, Archer *et al.* found proteomic clusters of SHH MB tumors (Archer et al., 2018) and Korshunov *et al.* identified a transcriptional subtype of SHH MBEN tumors with exceptionally good outcomes (Korshunov et al., 2020). Multiple single-cell RNA sequencing (scRNA-seq) studies have characterized the cell types in SHH MBs (Hovestadt et al., 2019; Riemondy et al., 2022; Vladoiu et al., 2019). They observed undifferentiated progenitors resembling cerebellar GCPs and differentiated NeuN+ cells.

Since SHH MBs are proposed to originate from GN progenitors, we hypothesized that the inter- and intra-tumoral heterogeneity in these samples is related to their developmental origins. Single-cell clusters and molecular subtypes have been described as “differentiated,” but not associated with specific stages of GN development. Precise annotations would allow for more biologically informed discussions about the clinical and therapeutic relevance of specific cell types. For

example, there is great interest in using differentiation therapy to treat SHH MB by inducing cycling progenitors to differentiate into postmitotic neurons (Breitman et al., 1980; Cheng et al., 2020; Cicconi and Lo-Coco, 2016), and more knowledge about the drivers of differentiation in these tumors would help inform target identification.

We reasoned that the relationship between development and intra-tumoral heterogeneity would be particularly pronounced in tumors with MBEN histology because they contain widespread differentiated nodules. Therefore, we performed single-nucleus RNA sequencing (snRNA-seq) on seven SHH MBs with the MBEN histology. We identified cells mimicking every stage of cerebellar GN development and then used computational techniques to relate these MBEN cell types to previously described examples of SHH MB heterogeneity. Specifically, we detail novel insights about tumor subtypes, copy number variations, metabolism, and histology. Overall, this work highlights computational and experimental approaches that can be used to investigate connections between tumor heterogeneity and known developmental trajectories.

6.3 Results

6.3.1 Tumor Cells in Medulloblastomas with Extensive Nodularity (MBEN) Recapitulate Granule Neuron Development

Prior scRNA-seq studies of SHH MB have not included any tumors with MBEN histology (Hovestadt et al., 2019; Riemondy et al., 2022; Vladoiu et al., 2019). While rare, such tumors are of particular interest due to their occurrence in very young patients and their unusual morphology, characterized by large regions of differentiated cells. We reasoned that a deeper analysis of MBEN tumors might provide insight into the relationship between normal GN development and tumor differentiation, so we performed snRNA-seq on seven MBEN tumors (Supplementary Table 1).

First, we clustered the data and annotated the corresponding cell types (Figure 1A). Tumor cells represent 92% of the high-quality nuclei, while the remaining nuclei are from cell types that commonly infiltrate SHH MBs, such as macrophages and microglia (Figure 1A, Supplementary Figure 1A). The malignant nuclei are heterogeneous and Louvain clustering revealed eight groups of cells (Supplementary Figure 2A). Our data confirm previous reports of tumor cells resembling non-cycling GCPs (GLI2+/TOP2A-), cycling GCPs (GLI2+/TOP2A+), and differentiated cells (NeuN+) (Figure 1B). Additionally, we performed pseudotime analysis on the malignant nuclei and identified a clear trajectory from cycling GCPs to differentiated cells (Figure 1B).

Despite these patterns, pseudotime and UMAP plots are not necessarily reflective of genuine biological differentiation. Fortunately, canonical GN differentiation is well studied in both humans and rodents, and there are known marker genes for each developmental stage (Figure 1C). GCPs express the SHH pathway marker GLI2 and proliferate in the EGL. In the normal developing cerebellum, the actively cycling GCPs

(TOP2A+) are located in the outer EGL, while the non-cycling GCPs reside in the inner EGL (Legué et al., 2015; Wallace, 1999; Wechsler-Reya and Scott, 1999). The GCPs can differentiate into GN, where they are SEMA6A+ for the short time they are migrating tangentially within the EGL (Kerjan et al., 2005). The GN then express the GRIN2B glutamate receptor as they migrate radially across the molecular layer (ML) (Akazawa et al., 1994; Tárnok et al., 2008; Watanabe et al., 1994). Once they reach the IGL and start to mature, GRIN2B is replaced by the GRIN2C receptor (Cathala et al., 2000; Cull-Candy et al., 1998; Losi et al., 2002; Takahashi et al., 1996).

To assess the relationship between our tumor cells and normal GN differentiation, we re-analyzed the nuclei along the potential GCP-to-GN trajectory. Remarkably, we observed tumor nuclei expressing key markers from every stage of GN differentiation and maturation (Figure 1D, Supplementary Figure 1D) (Abe et al., 2011; Miyazaki et al., 2003; Sato et al., 2005; Suzuki et al., 2005). These transcriptomic patterns suggest that some MB tumor cells retain the capacity to recapitulate canonical GN development.

Since there are known biases in single-cell and single-nucleus sequencing (Ding et al., 2020), we also performed scRNA-seq on fresh cells from one MBEN tumor (Supplementary Figure 1E). The scRNA-Seq cells do not express the same exact markers as the snRNA-Seq nuclei, but we still observe clusters of malignant cells with markers for each stage of GN development (Supplementary Figure 1F).

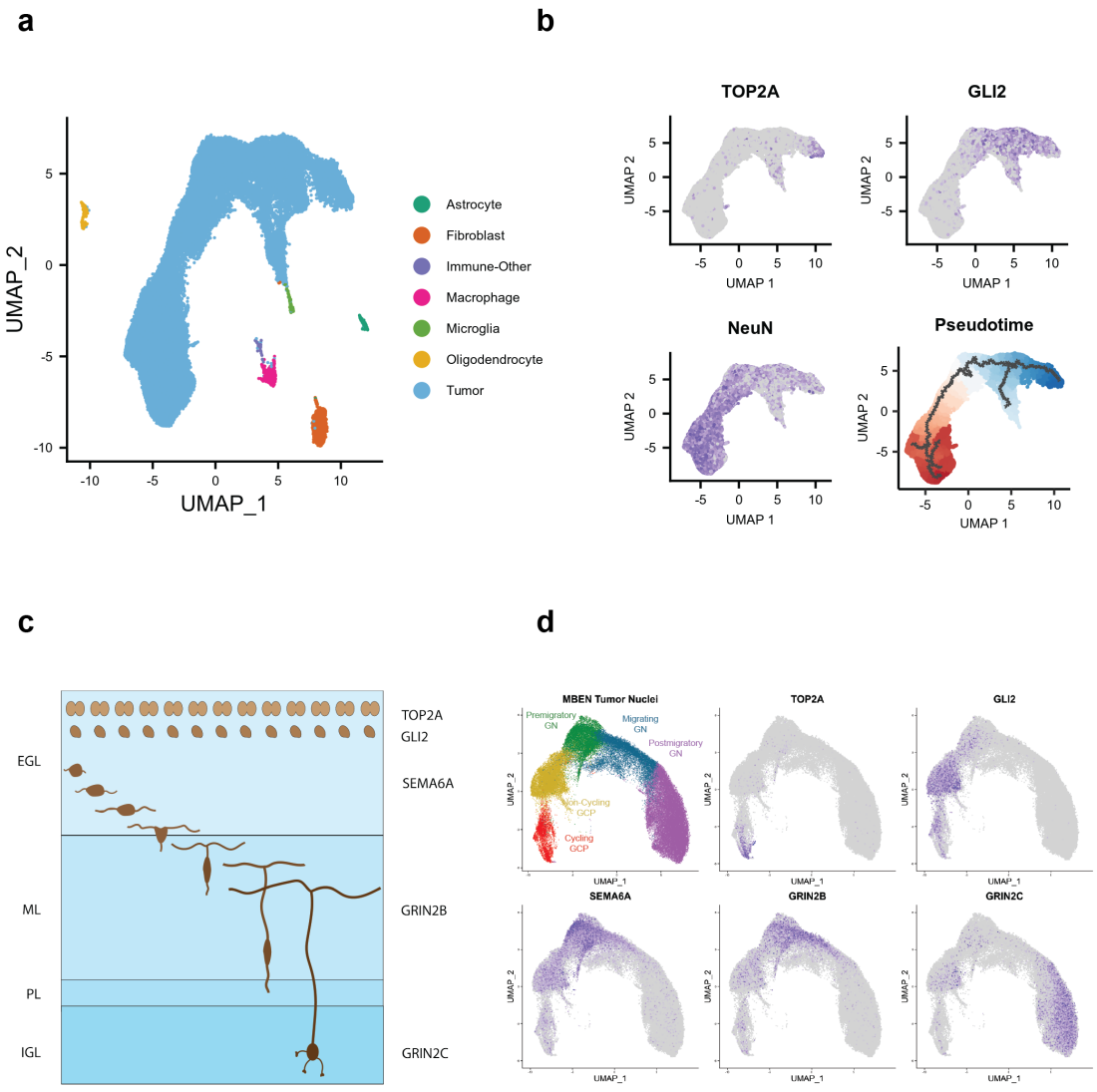


Figure 1) MBEN Tumor Cells Mimic Granule Neuron Development

A) Summary UMAP of Seven tumors with MBEN histology. Malignant and non-malignant cell types are labeled in the legend.

B) Marker Genes and Pseudotime for Malignant Cells. Tumor cells express markers for cycling GCPs (TOP2A+/GLI2+), non-cycling progenitors (TOP2A-/GLI2+), and differentiated neurons (NeuN+). Bottom right image shows pseudotime analysis rooted at cycling cells. Pseudotime increases from dark blue to white and then dark red. The black lines represent trajectories identified by *monocle3*.

C) Summary of Granule Neuron Development. Granule cell precursors (GCPs) proliferate in the outer portion of the external granule layer (EGL), while non-cycling progenitors lie in the middle portion of the EGL. As the granule neurons (GN) differentiate, they express SEMA6A as they migrate tangentially across the inner portion of the EGL. The GN then turn and migrate radially across the molecular layer (ML), during which they express the glutamate receptor GRIN2B. Once the GN reach their final location in the internal granule layer (IGL), GRIN2C replaces GRIN2B.

D) MBEN Tumor Cells Resemble Stages of Granule Neuron Development. UMAP plot for malignant tumor cells along potential GCP to GN trajectory. There are tumor cells that express markers for each stage of GN development: cycling GCPs (TOP2A+), non-cycling GCPs (TOP2A-/GLI2+), premigratory GN (SEMA6A+), migrating GN (GRIN2B+), and postmigratory GN (GRIN2C+). Feature plots use a minimum cutoff at the 80th percentile for each marker to highlight the cells with the highest expression.

6.3.2 **Clustering of Gene Set Signatures Reveals Connections Between Granule Neuron Development and SHH MB Heterogeneity**

Since we observed tumor cells expressing markers from each stage of GN development, we sought to understand how these MBEN cell types relate to previously published examples of SHH MB heterogeneity. To accomplish this, we developed the computational approach outlined in Figure 2A. This method is based on generating gene signatures from relevant developmental and tumor datasets and then using a large compendium of expression data from SHH MBs to identify which signatures are activated in the same patients.

We defined signatures from six SHH MB studies (Supplementary Figure 2A) by identifying the top 100 marker genes from each cell type or molecular subtype (Supplementary Table 2). Additionally, we generated signatures for each stage of canonical GN development using published data from human brain tissue (Okonechnikov et al., 2021) and P14 mouse cerebella (Vladoiu et al., 2019) (Supplementary Figure 2B). We then used Gene Set Variation Analysis (GSVA) (Hänzelmann et al., 2013) to calculate an activation score for each signature in all 223 SHH MBs from the MAGIC cohort (Cavalli et al., 2017) (Supplementary Table 3).

To identify relationships among the signatures, we performed consensus clustering on the signature scores. Figure 2B summarizes the co-clustering pattern for 1000 trials. We observe a clear correspondence between the MBEN cell types and the known stages of GN development and this pattern holds true across many clustering parameters and gene signature sizes (see Methods, Supplementary figure 2D).

Additionally, we analyzed how the signatures of GN development relate to cell types identified from SHH MBs with non-MBEN histology. Specifically, we investigated the SHH-A, SHH-B, and SHH-C clusters from Riemondy *et al.* (Riemondy et al., 2022). We confirmed their findings that the SHH-A and SHH-B signatures correspond to cycling and non-cycling GCPs respectively. We also identified a novel association between the SHH-C2 signature and premigratory GN, which is supported by previous studies showing that the SHH-C2 cells populate the differentiated nodules in SHH MBs (Riemondy et al., 2022).

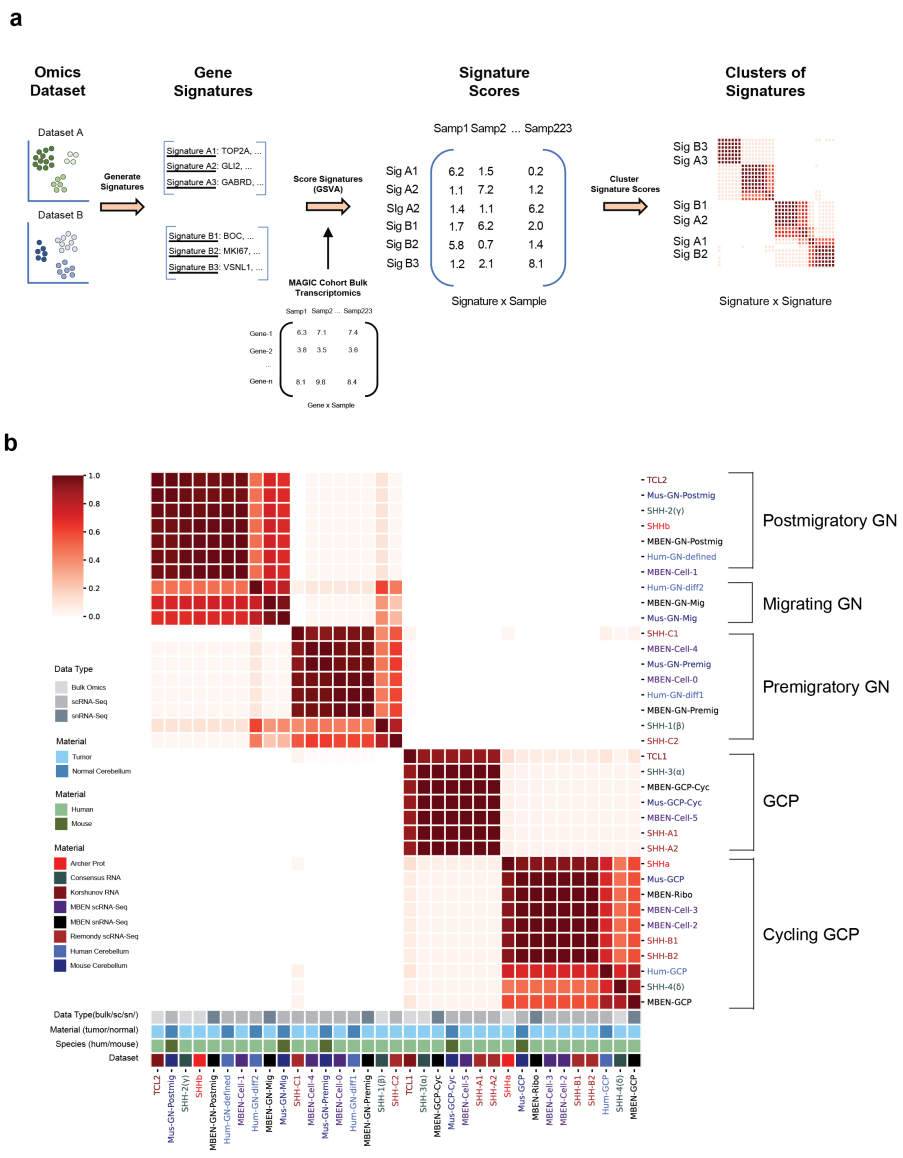


Figure 2) Clustering of Gene Signature Activation Scores

A) Clustering of Gene Signature Scores Approach. The first step in this method is collecting the bulk or single-cell datasets of interest (Supplementary Figure 2A). For each dataset, clustering was used to identify relevant cell types or molecular subtypes. Then the top 100 marker genes were identified for each cluster and considered the gene signature. Those marker genes were used as gene sets for Gene Set Variation Analysis (GSVA), which was run on bulk transcriptomics data from 223 SHH tumors in the MAGIC cohort. This method converts the “genes by samples” matrix into a “signatures by samples” matrix by scoring each gene set signature for each of the 223 tumors. Consensus clustering was then run on those signature scores to identify which signatures are activated in the same SHH tumors.

B) Signature Score Consensus Clustering Summary. Co-clustering results of 1000 samples of the “signatures by samples” output dataset. In each case, signatures were clustered using data from 70% of SHH MBs. Beige signifies signatures that never cluster together and stronger red coloring indicates signatures that cluster together more often. The legend at the bottom indicates the dataset of origin, the species of the cells (human or mouse), the data type (scRNA-Seq, snRNA-Seq, or bulk), and material type (SHH MB tumor or healthy cerebellum). The groups of signatures on the right were manually annotated using the known cerebellar cell types from human and P14 mouse samples.

6.3.3 **Consensus Subtypes of SHH MB are Associated with Specific Developmental Stages**

The signatures we identified from normal GN development and MBEN tumors provide an opportunity to connect these cell types to other examples of SHH MB heterogeneity. Clustering of DNA methylation and transcriptomics data revealed four consensus subtypes of SHH MB with distinct clinical and molecular features: SHH-1 (β), SHH-2 (ɣ), SHH-3 (α), and SHH-4 (δ) (Cavalli et al., 2017). The 223 SHH tumors from the MAGIC cohort are annotated with a SHH subtype, so we re-analyzed the GSVA activation scores for our MBEN cell types to investigate potential associations (Figure 3A). The SHH-3 samples have the highest average cycling GCP signature scores (0.34 in SHH-3 vs. -0.11 in others), while the SHH-4 tumors are associated with the GCP signature (0.26 vs. -0.10). Both infant subtypes (SHH-1and SHH-2) have high scores for differentiated cells but are enriched for specific developmental stages; SHH-1 samples have the highest premigratory GN scores (0.43 vs. -0.08), while SHH-2 tumors have significant upregulation of the postmigratory GN signature (0.50 vs. -0.21). It is noteworthy that the subtype with the least favorable outcomes (SHH-3) also has the highest cycling GCP scores (Figure 3A).

We then analyzed the correlations among the MBEN cell type signatures and found that the postmigratory GN signature has a strong negative correlation ($r = -0.57$) with the progenitor score (i.e. sum of Cycling GCP score and GCP score) (Figure 3B). This suggests that tumors with more postmigratory GNs may have fewer progenitor cells. However, this pattern does not exist for the premigratory GN signature, which has no significant relationship with the progenitor score GN ($r = -0.02$) (Figure 3B).

6.3.4 **Genomic Associations with Specific Developmental Stages**

SHH MB tumors frequently contain copy number variations (CNVs), where large parts of chromosomal arms are lost or gained (Kool et al., 2014). Since these chromosomal alterations are quite common and affect many genes at once, we hypothesized some CNVs could affect tumor cell differentiation. We investigated potential associations between large chromosomal CNVs and the MBEN cell types using data from the MAGIC cohort. We considered individual cell type signatures and aggregate features that combine related developmental stages (see Methods). The strongest association is between a loss of chromosome 10q and the proliferation score (cycling GCP score minus the GCP score) (t-test $p < 0.01$) (Figure 3C). This pattern of 10q loss leading to higher cycling GCP scores is consistent for every subtype, suggesting that loss of chromosome 10q, which contains the *SUFU* and *PTEN* genes, may drive progenitor cells to remain proliferative.

Moreover, there is a strong negative relationship between a loss of chromosome 9q, which contains the *PTCH1* and *ELP1* genes, and the signature for late-stage GNs (i.e. sum of migrating and postmigratory GN scores) (t-test $p < 0.01$) (Figure 3D). This

pattern is most prominent in the SHH-2 patients, where the samples with extremely high activation scores for late-stage GNs rarely have loss of chromosome 9q. This trend is not observed for other common CNVs or premigratory GNs (Supplementary Figure 3A-C), suggesting that loss of chromosome 9q may inhibit tumor cells from progressing to the later stages of GN differentiation.

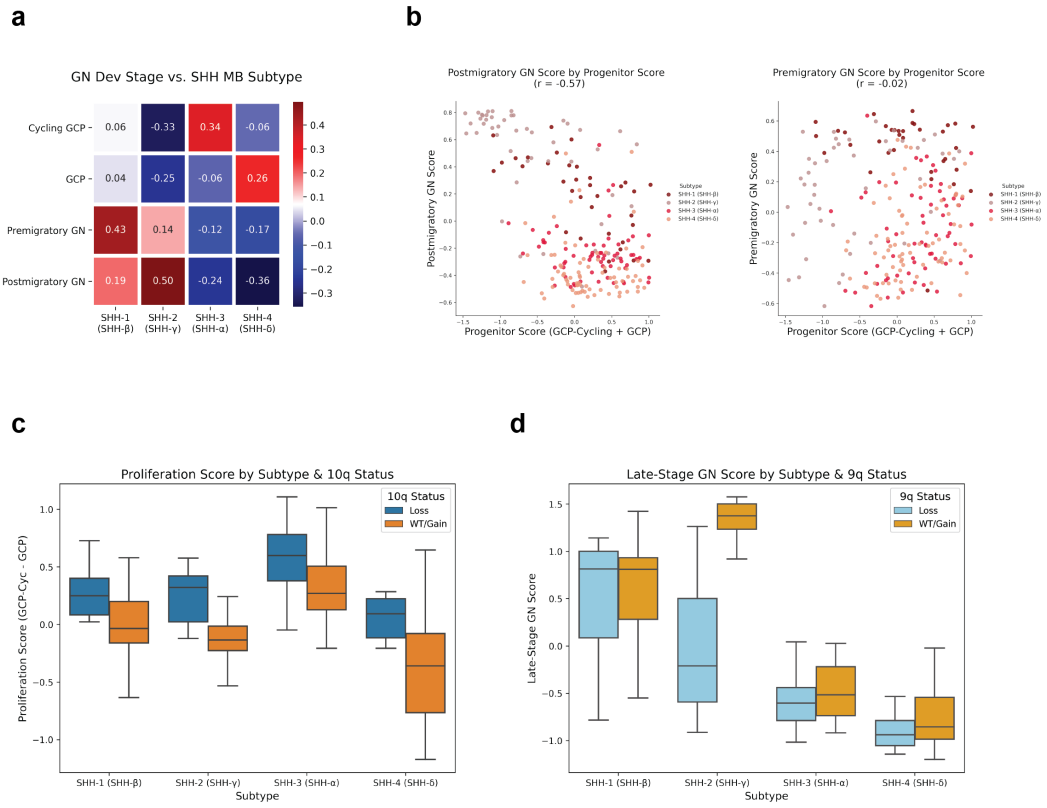


Figure 3: Genomic and Transcriptomic Associations Between SHH MB and GN Development
A) Mean Cell Type Activation Per Cell Type. Each box represents the mean GSVA activation for a given consensus subtype and an MBEN cell type resembling a specific GN development stage. Red indicates higher activation, while blue signals lower GSVA scores.
B) Associations Between Progenitor Score and Differentiated Cell Types. For both plots, each dot indicates a single sample from the MAGIC cohort and is colored by consensus subtype (SHH-1, SHH-2, SHH-3, or SHH-4). The left figure shows the postmigratory GN score on the y axis and the progenitor score (cycling GCP + non-cycling GCP) on the x axis. These two features have a significant negative correlation. The right plot uses the same x-axis, but the y-axis indicates the premigratory GN score. The progenitor score is not significantly correlated with premigratory GN.
C) Boxplot of Association Between Chromosome 10q and Cycling GCPs. Boxplots reflect GSVA scores for SHH MB tumors from MAGIC cohort. Y axis is the Proliferation score (cycling GCP score minus the GCP score). The X axis is separated by consensus subtype and further divided by 10q status (loss in blue, wt or gain in orange). For all subtypes, samples with 10q loss show higher differences between cycling and non-cycling GCP scores.
D) Boxplot of Association Between Chromosome 9q and Late-Stage GNs. Y axis is the late-stage GN score (migrating GN + postmigratory GN). The x-axis is separated by consensus subtype and further divided by 9q status (loss in light blue, wt or gain in light orange). SHH-2 samples show a substantial difference in postmigratory GN signature based on 9q status.

6.3.5 **The SHHb Proteomic Subtype is Associated with Tumor Cells Mimicking Late-Stage Granule Neurons**

Archer *et al.* identified proteomic subtypes of SHH MB that are not observed when clustering RNA or methylation data (Archer et al., 2018). The SHHb subtype is characterized by proteins related to specific neuronal functions like glutamatergic synapses and axon guidance. We re-analyzed the Archer proteomics data and found that even though synaptic proteins, like PCLO and DLG4, are upregulated in SHHb tumors, many markers of neuronal differentiation, such as NEUROD1 (Miyata et al., 1999) and SEMA6A (Kerjan et al., 2005), show similar expression across the groups (Figure 4A). This suggests that the SHHb proteomic subtype is not simply a proxy for tumors with differentiated cells.

To better understand the relationship between the snRNA-seq results and the SHHb phenotype, we used the SHHb proteomic markers (Supplementary Table 2) to calculate an activation score for each MBEN nucleus (see Methods). The highest activation occurs in tumor cells mimicking the migrating and postmigratory stages of granule neuron development (Figure 4B). We then analyzed the cell type composition of each tumor individually and observed a striking pattern: most MBEN tumors contain all GN cell types, but two samples only contain nuclei resembling GCPs and premigratory GN (Figure 4C). These results suggest that the SHHb subtype is driven by the presence of cells resembling the latest stages of GN differentiation. This would explain why early differentiation markers have similar expression between SHHa and SHHb tumors, while proteins related to late-stage developmental processes (e.g. synaptogenesis) are significantly enriched in SHHb samples.

We then sought to investigate how common these late-stage GN cells are in non-MBEN tumors. We re-analyzed published scRNA-Seq data from 14 SHH MBs and in no sample did we observe a cluster with high expression of the postmigratory GN markers GABRD and VSNL1 (Supplemental Figure 4B). This lack of expression is unlikely to be due to the differences between scRNA-Seq and snRNA-Seq because the one MBEN tumor in our cohort with scRNA-Seq data clearly shows a group of GABRD+/VSNL1+ cells (Supplemental Figure 4C). It is noteworthy that even though the published non-MBEN tumors do not contain cells mimicking late-stage GNs, every sample has a cluster of SFRP1+ cells resembling undifferentiated progenitors. Additionally, many tumors have cells expressing the premigratory GN markers STMN2 and SEMA6A. These findings suggest that non-MBEN SHH MBs still contain cells resembling the earliest stages of GN development, but MBEN tumors are more likely to have cells mimicking late-stage GNs that are associated with the SHHb subtype.

6.3.6 **FMRP-Induced Post-Transcriptional Regulation Helps Explain SHHb Proteomic Phenotype**

The association between the proteomic SHHb subtype and late-stage GNs raises another question: why does the presence of late-stage GNs result in clear clustering in the proteomic data, while having much less of an impact on RNA or methylation data? We hypothesized that this could be due to post-transcriptional regulation occurring in the SHHb-specific cell types. To test this theory, we re-analyzed 8674 genes from the Archer cohort by rank-normalizing the protein and RNA data for each sample and then calculating a rank difference (protein rank – RNA rank) for each gene to get a rough proxy for post-transcriptional regulation (see Methods).

We explored many gene sets and found that synaptic genes have especially high rank differences in the SHHb tumors, but not the SHHa tumors (Supplemental Figure 4A). We then specifically analyzed targets of FMRP, a protein encoded by the *FMR1* gene that regulates translation of RNAs related to synaptic plasticity and axon guidance (Antar et al., 2006; Darnell et al., 2011; Napoli et al., 2008). In SHHb tumors, synaptic genes targeted by FMRP have significantly higher rank differences than all other genes and show significantly higher rank differences than the other synaptic genes (Figure 4D). No such trend occurs in SHHa tumors where all four gene sets show similar rank differences around 0. Thus, positive rank differences for both synaptic genes and FMRP targets suggest that the SHHb phenotype may be especially strong in proteomic data due to post-transcriptional regulation specific to functions of late-stage GNs.

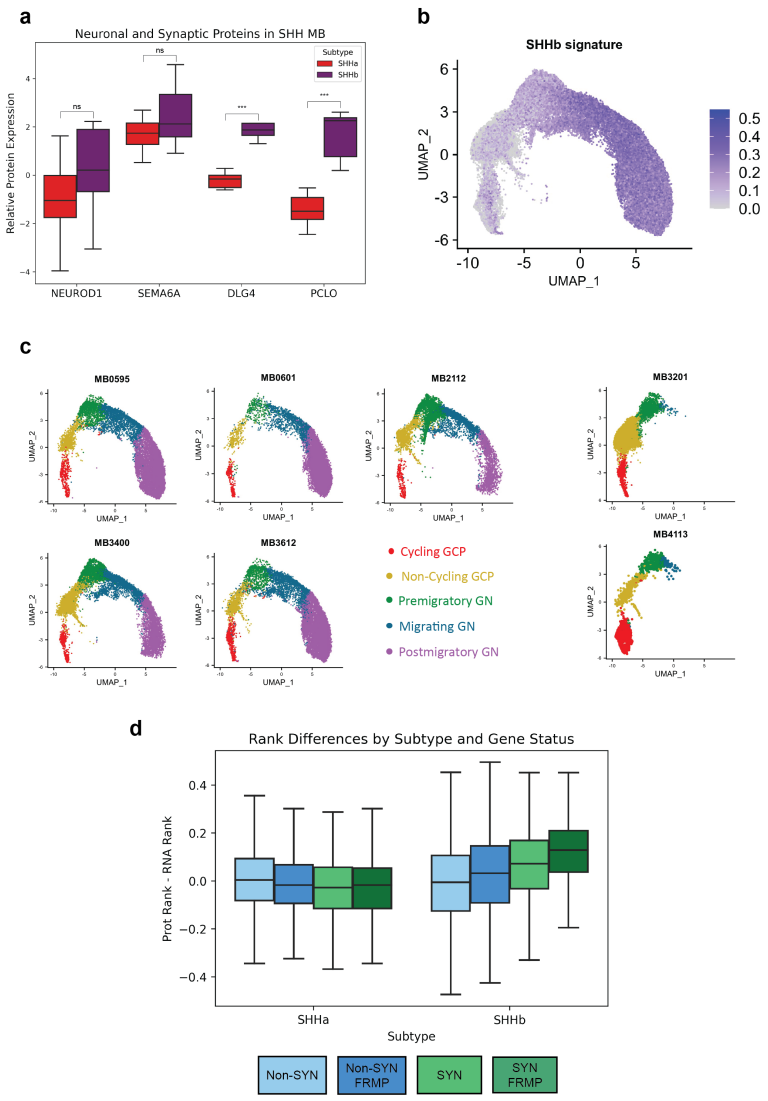


Figure 4: SHHb Proteomic Subtype Associated with Late-stage GNs.
A) Neuronal Protein Expression in SHH MB. Relative protein expression from Archer *et al.* data for markers of early differentiation (NEUROD1 and SEMA6A) and synapses (DLG4 and PCLO). Synaptic proteins are significantly upregulated in SHHb tumors compared to SHHa (t-test adj-p < 0.01), while NEUROD1 and SEMA6A are not.
B) SHHb Signature Scores for MBEN Nuclei. Activation scores for each MBEN nucleus using the top 100 SHHb marker proteins (see Methods). Scores were filtered using a minimum cutoff of zero. SHHb activation is highest in tumor cells mimicking migrating and postmigratory GN.
C) Tumor Cell Types per Sample. Five MBEN samples contain all types of cells. Two MBEN tumors do not contain late-stage GNs (migrating and postmigratory GN).
D) Rank Differences by Synapse and FMRP status. For each SHH tumor from Archer et al (2018), the RNA and Protein data were rank normalized and a rank difference (protein rank – RNA rank) was calculated for each gene. For every gene in every SHHa tumor, the mean rank difference was calculated across samples and then the same procedure was applied to SHHb tumors. The genes were then divided into four categories: Non-synaptic genes not targeted by FMRP (Non-SYN), non-synaptic genes targeted by FMRP (Non-SYN FMRP), synaptic genes not targeted by FMRP (SYN) and synaptic genes targeted by FMRP (SYN FMRP). For SHHb tumors, “SYN” and “SYN FMRP” genes have high rank differences. For SHHa tumors, all categories show rank differences around zero.

6.3.7 **Desmoplastic/Nodular (DNMB) Histology in SHH MB Reflects Granule Neuron Development**

We sought to understand how the neuronal MBEN cells from our snRNA-Seq data relate to the differentiated nodular regions observed in some SHH MBs. Eberhart *et al.* compared DNMB histology to the layers of the developing cerebellum (Figure 1C) (Eberhart et al., 2001), hypothesizing that the cycling internodular regions represent the progenitor cells of the EGL and that the nodules themselves represent the differentiated GN of the IGL. This model has been a useful framework for understanding DNMB histology, but our snRNA-seq data suggests it can be improved since our cohort includes two MB tumors with extensive nodularity that contain no cells resembling the postmigratory GN that populate the IGL (Figure 3C).

Thus, we propose an alternative model where most SHH MB tumors with DNMB histology are composed of nodules containing NeuN+ cells mimicking premigratory GN; these regions most closely correspond to the most internal part of the EGL, rather than the IGL (Figure 1C). Additionally, we posit that a subset of SHH MBs contain nodules that recapitulate the later migrating and postmigratory stages of GN development.

To test this hypothesis, we performed cyclic immunofluorescence (CyCIF) (Lin et al., 2016, 2018) on SHH MBs to detect relative protein levels for four markers related to GN development: Ki67 (cycling cells), MAP2 (all GNs), CNTN1 (late-stage GN axons), and VSNL1 (late-stage GN axons and dendrites). These proteins allow us to distinguish premigratory GNs (MAP2+/VSNL1-) from late-stage GNs (MAP2+/VSNL1+) (Supplementary Figure 5A). We ran this experiment on sections from eight tumors with known SHHa/SHHb proteomic subtype calls (Archer et al., 2018). All nodular tumors contain cells resembling premigratory GN (MAP2+/VSNL1-), while only the SHHb samples have areas resembling late-stage GNs (MAP2+/VSNL1+) (Figure 5A). Three of the four SHHb tumors have large VSNL1+ regions, while MB206 only contains a small area of VSNL1+ cells. These findings support our proposed model and highlight that MB nodules vary in their developmental stage.

6.3.8 **Significant Variability in VSNL1 Staining Between and Within Tumors**

Korshunov *et al.* found that the TCL2 transcriptional subtype of MBEN tumors contain diffuse VSNL1 staining and have exceptional outcomes (Korshunov et al., 2020). Based on our snRNA-seq data, VSNL1 is only expressed in tumor cells mimicking migrating and postmigratory GNs (Supplementary Figure 5A). This suggests that the TCL2 subtype may be identifying samples that are primarily composed of late-stage GNs. Given the potential clinical relevance of VSNL1 staining, we wanted to better understand the variability of VSNL1+ cells in SHH MBs. We performed immunohistochemistry targeting VSNL1 on FFPE slides from an additional seven MB with DNMB or MBEN histology and observed significant heterogeneity between and

within tumors (Supplementary Figure 5B). Samples like CHLA-3 contain no VSNL1, while others like CHLA-10 are almost entirely composed of VSNL1+ regions.

We further investigated this heterogeneity by running our CyCIF panel on CHLA-5, which showed significant regional variability in VSNL1 staining between sections (Figure 5B). One region is almost entirely composed of MAP2+/VSNL1- nodules, while another area contains mostly MAP2+/VSNL1+ cells. This suggests that local microenvironment may affect the differentiation stage in distinct regions of the same tumor.

6.3.9 **Tumor Cell Spatial Organization Can Recapitulate the Developing Cerebellum**

Since we observed tumor cells mimicking the expression patterns of differentiating GNs, we also wanted to investigate whether these cells spatially organize like the developing cerebellum. First, we confirmed the established phenomenon that Ki67+ cells are primarily located in internodular areas (Figure 5B). In multiple samples, we noticed that this pattern is more extreme for VSNL1+ nodules, which rarely contain Ki67+ cells, compared with MAP2+/VSNL1- nodules which will sometimes be infiltrated by cycling cells (Supplementary Figures 5C, 5D).

Additionally, in CHLA-10, tumor cells show a region that closely corresponds to the appearance of the developing cerebellum (Figure 5C, 5D). There is an outer layer resembling the EGL, which contains Ki67+ cells at the outer edge and MAP2+/VSNL1- cells resembling premigratory GNs at the inner edge. Within that area is a region analogous to the molecular layer, filled with CNTN1+/VSNL1+ axonal processes and very few cells. Lastly, the central region is composed of VSNL1+ cells mimicking the postmigratory GNs of the IGL. Of note, this pseudo-cerebellar structure lacks a Purkinje cell layer as GCP stem cells are not capable of differentiating into Purkinje neurons.

We also analyzed the staining patterns for many distinctive cellular patterns observed in MBEN tumors, such as parallel rows of nuclei surrounded by neuropil. These structures also roughly mimic the developing cerebellum by having central VSNL1+ cells surrounded by their VSNL1+/CNTN1+ axons (Supplementary Figure 5E). In total, these imaging results suggest that both simple and complex nodular structures routinely observed in SHH MB can be explained by malignant cells recapitulating specific stages GN development.

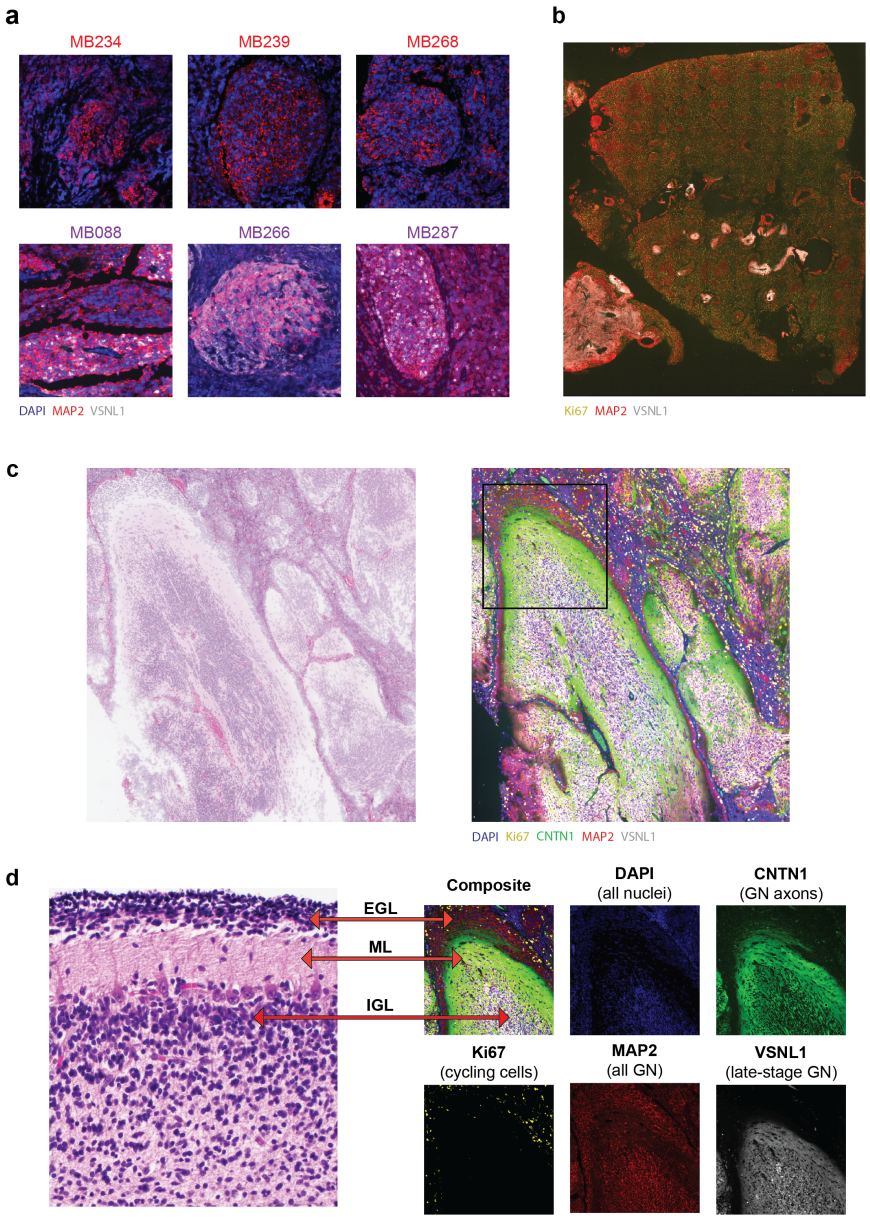


Figure 5: SHH MB Nodules Recapitulate Granule Neuron Development
A) Examples of VSNL1- and VSNL1+ nodules. Staining for DAPI (blue), MAP2 (red), and VSNL1 (white). The top row shows nodules with MAP2+/VSNL1- cells resembling premigratory GN in tumors with SHHa proteomic subtype. The three tumors with the SHHb proteomic subtype are on the bottom row and contain differentiated regions that are MAP2+/VSNL1+ mimicking the later stages of GN development.
B) Ki67 and VSNL1 anticorrelate in CHLA-5: Ki67 (yellow), MAP2 (red) and VSNL1 (white). Tissue section on bottom left is primarily composed of MAP2+/VSNL1+ cells and has very few cycling cells. The larger tissue on the right contains many Ki67+ cells and MAP2+/VSNL1- nodules.
C) Tumor cells Mimic Cerebellar Structure in CHLA-10: Left image contains H&E stain from pseudo-cerebellar region in CHLA-5. Right image shows CyCIF staining same region.
D) Zoomed in region highlights cerebellar layers: Left image is H&E stain from developing cerebellum. The right images contain the boxed section from Figure 5C. Outer layer resembles the EGL and contains Ki67+ cycling cells and MAP2+/VSNL1- cells mimicking premigratory GN. The next layer is like the molecular layer (ML), which has few nuclei and CNTN1+/VSNL1+ axons. The white interior region represents the internal granule layer (IGL) and is filled with VSNL1+ cells mimicking postmigratory GN.

6.3.10 **Tumors with Late-Stage Granule Neurons have Distinct Metabolic Profiles**

We are particularly interested in SHH MB metabolism because many metabolites, like glutamate, are drivers of GN migration and differentiation (Consalez et al., 2021; Komuro et al., 2013). Genomic and transcriptomic heterogeneity of SHH MBs has been well studied, but there are very few published papers analyzing SHH MB metabolism and these projects typically focus on tumor type or subtype classification using bulk metabolomics(Bennett et al., 2018; Wilson et al., 2009).

We used MALDI Imaging Mass Spectrometry (MALDI-IMS)(Stoeckli et al., 2001) to collect spatial metabolomics data on ten tumor sections collected from nine individual SHH MB patients (Supplementary Figure 6A). Seven of the nine patients were included in the snRNA-seq cohort (Figure 1), while the other two samples are from tumors with DNMB histology. From the snRNA-seq data, we know that five samples contain tumor cells resembling late-stage GNs and we sought to understand the metabolic differences between these tumors and the others. First, we calculated the mean metabolite levels for each section and performed differential analysis. The most upregulated metabolite in tumors with late-stage GNs is N-Acetylaspartic acid (NAA), which is synthesized by neurons and has extremely high concentrations in the brain (Tallan, 1957; Tallan et al., 1956). The most downregulated metabolites are related to nucleic acids, like UDP and adenine.

We then implemented joint graphical lasso (Danaher et al., 2014) (see Methods) to generate networks of conditionally dependent metabolites in each tumor to better understand how metabolite co-expression patterns change across samples. Some core metabolite relationships, like glutamate-glutamine, were observed in the networks from every sample. We used betweenness centrality, a metric of a node's influence within a network, to identify metabolites that are especially important for each tumor. When comparing the networks from samples with late-stage GNs to the other networks, taurine has the largest mean centrality difference. Figure 6A shows select edges and highlights metabolic relationships with taurine that are overrepresented in the tumors with late-stage GNs. The prominence of taurine in these tumor networks is noteworthy because of prior literature linking taurine to GCP differentiation and GN migration (Maar et al., 1995; Sturman et al., 1985) . One study specifically shows that depleting dietary taurine in mother cats impedes GN development in newborn kittens, a phenotype that can be abrogated by directly feeding taurine to the kittens (Sturman et al., 1985).

Given the significance of taurine to the late-stage GN network and GN development, we further analyzed the spatial distribution of taurine. We used bivariate Moran's I statistic to assess what metabolites correlate or anticorrelate with taurine expression in a given cellular neighborhood (see Methods). We found that there is a strong negative association between taurine and guanine in tumors with late-stage GNs (Figure 6B). This anticorrelation is visually striking in the tumors with late-stage GNs, but is not observed in other samples (Figure 6C).

These results suggest that taurine may be playing a specific role in tumors with late-stage GNs. We assessed this hypothesis expanding our CyCIF panel to include an

antibody that detects taurine and using this new panel on three MBEN tumors. We found that taurine staining is more prominent in VSNL1+ regions compared to MAP2+/VSNL1- regions. This trend is strongest in CHLA-5 where there is high taurine abundance within the VSNL1+ nodules, but not within the MAP2+/VSNL1- nodules (Figures 6D). Additionally, we re-stained the pseudo-cerebellar structure from CHLA-10 with this new panel and observed high taurine staining in the central region mimicking the IGL and around the edges of the structure (Supplementary Figure 6C). There appears to be a region with strong taurine staining between the pseudo-ML and the pseudo-EGL (Figure 6E). These imaging experiments further support that taurine is associated with tumor cells mimicking late-stage GNs.

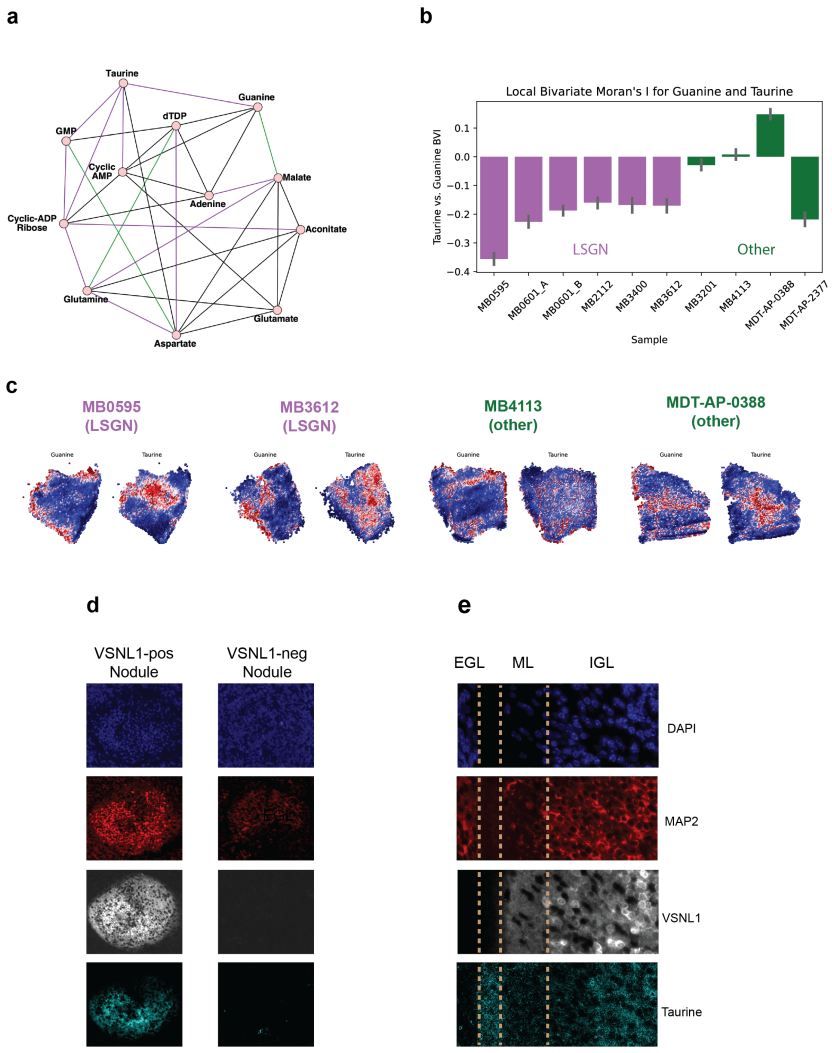


Figure 6: Metabolic Features of Differentiation in SHH MBs
A) Select Edges from Joint Graphical Lasso Analysis: Purple edges appear in at least 50% of the networks for tumors with late-stage GNs, but not for the other samples. Green edges only appear in the consensus network for the other samples. Black edges appear in consensus networks from both groups.
B) Barplot of Local Bivariate Moran's I between Taurine and Guanine: The local bivariate moran's I was calculated between the guanine and the spatial lag of taurine for each spot in each section (see Methods). Barplot indicates the mean statistic and error bars were generated through bootstrapping. Tumors with late-stage GNs (purple) have a strong negative relationship not consistently observed in the other samples (green).
C) Taurine and Guanine Anti-Correlate in Tumors with late-stage GNs: For four sections, the relative expression values are plotted for guanine and taurine. Each plot shows the metabolite values, clipped at the 3rd and 97th percentiles. The tumors with late-stage GNs show clear spatial anti-correlation between guanine and taurine. This pattern is not consistently found in the other tumors.
D) Taurine Stains in VSNL1+ Nodules in CHLA-5: DAPI (blue), MAP2 (red), VSNL1 (white) and taurine (cyan) stains. Left panel shows VSNL1+ nodule where internal region is clearly positive for taurine. Right panel shows MAP2+/VSNL1- nodule that does not show significant taurine staining.
E) Vertical layers of pseudo-cerebellar structure from CHLA-10: Zoomed in region from orange box in Supplementary Figure 6C. Furthest left region resembles EGL with MAP2+/VSNL1- cells. On the far right is a region similar the IGL that contains MAP2, VSNL1, and Taurine. In between is a region divided in two, where the right section contains VSNL1+ axons and the left one stains for taurine. There appears to be a layer with high taurine levels between the pseudo-ML and pseudo-EGL.

6.4 Discussion

We performed snRNA-seq on seven MBEN tumors and found malignant cell types mimicking every stage of cerebellar granule neuron development. By re-analyzing published data using this developmental perspective, we were able to elucidate the causes of previously unexplained molecular and histological phenomena. Specifically, we found that each consensus subtype of SHH MB is enriched for a specific developmental stage and that the proteomic SHHb subtype is likely caused by the presence of tumor cells resembling late-stage GNs. Additionally, a spectrum of recognized histological patterns, such as layered nodules and linear arrays of tumor nuclei in MBEN tumors, can now be understood as tumor cells mimicking the structure of the IGL of the developing cerebellum.

This work is the first to describe tumor cells mimicking late-stage GNs and presents significant progress in our understanding of the biological causes of tumor differentiation in SHH MBs. While preparing our manuscript, we became aware of a parallel study carried out by Ghasemi *et al.* (personal communication). Strikingly, they identify the same cell types that we do and observe similar spatial patterns. Additionally, Ghasemi *et al.* collected snRNA-Seq data from three tumors from the Archer *et al.* proteomics cohort (Archer et al., 2018) and consistent with our predictions, these SHHb tumors (MB088, MB266, and MB287) all contain cells resembling late-stage GNs.

It is still not well understood why some SHH MBs are primarily composed of differentiated cells and other tumors have none. Based on this study, we believe that tumor microenvironment and genomics are important factors. Tumors with extensive nodularity rarely contain large CNVs and frequently occur in infants, whose brains are actively developing. We hypothesize the tumor microenvironment in these young patients may contain pro-development factors that can induce some malignant cells to escape the progenitor state and follow GN differentiation. If these tumor cells truly recapitulate GN development, they could promote the differentiation of nearby malignant cells through the release or expression of factors like glutamate and CNTN1 (Tárnok et al., 2008; Xenaki et al., 2011). This could induce a feed-forward loop whereby more maturation leads to more cells producing pro-differentiation molecules.

We also observed that CNVs are associated with distinct developmental stages. We highlighted alterations to chromosome 9q and 10q, which are significant because they contain the key SHH pathway genes *PTCH1* and *SUFU*. Loss of 9q or 10q can activate the SHH pathway, but these CNVs may also promote tumorigenesis by inhibiting differentiation. *PTEN* and *NEURL1* are both located on chromosome 10q and negatively regulate GCP cycling (Adachi et al., 2021; Teider et al., 2010; Zhu et al., 2017). Additionally, chromosome 9q contains *NTRK2*, which plays a vital role in granule cell migration (Borghesani et al., 2002; Zhou et al., 2007) and maturation (Suzuki et al., 2005). There are many SNPs and CNVs with unknown effects on SHH MBs and it is worth further exploration to determine the potential developmental impact of those mutations.

Our work reinforces the importance of understanding differentiation state for prognosis. Korshunov *et al.* showed that MBENs with diffuse VSNL1 staining have excellent outcomes, while the other MBEN tumors have similar survival rates to samples with DNMB histology (Korshunov et al., 2020). Our snRNA-Seq data shows that VSNL1 is exclusively expressed in cells resembling late-stage GNs, suggesting these cell types may have greater clinical relevance than other differentiated cells. This hypothesis is supported by our GSVA analysis, which shows that the late-stage GN score negatively correlates with the progenitor score; by contrast, the score for premigratory GNs, an earlier stage of differentiation, has no relationship with the progenitor score (Figure 3B). Together, these results suggest that not all SHH MB differentiation has the same prognostic relevance, and that differentiation stage could potentially be used to stratify individual patient risk more accurately.

These molecular and prognostic insights are important for the development of therapeutics. The similarities we observe between tumor cells and GN development suggest that the current understanding of human cerebellar development can be leveraged for target identification for differentiation therapies. Extracellular factors like CNTN1, glutamate, and taurine promote differentiation and migration during canonical GN development. Unfortunately, there are no established models of MBEN tumors for us to use in this study, but future experiments can test whether these molecules can also induce differentiation in SHH MBs.

In summary, this work characterizes the differentiated cells in SHH MBs, establishes the key transcriptomic and metabolomic patterns of those cells, and uses these findings to help explain the biological basis of observed molecular subtypes and histological patterns. It is unlikely that this study includes every malignant cell type related to SHH MB biology, but we do show that most tumor cells in pediatric SHH MBs can be associated with some stage of GN development. We hope that these findings promote further research into connections between SHH MB tumorigenesis and cerebellar GN development and that a deeper understanding of this relationship will ultimately enable novel therapeutic approaches.

6.5 Methods

6.5.1 **Preparation of single-cell suspensions**

Fresh patient tumors were collected at the time of surgical resection. Tumor tissue was mechanically dissociated followed by collagenase-based enzymatic dissociation as previously described (Vladoiu et al., 2019).

6.5.2 **Preparation of single-nuclei suspensions**

Nuclei were isolated from fresh, snap frozen tumor tissues as previously described (Nagy et al., 2020). Frozen tissues were dounced in 1 ml of chilled lysis buffer (lysis buffer; 10mM Tris-HCl (pH 7.4), 10mM NaCl, 3 mM $MgCl_2$, 0.05% NP-40 detergent) 5 times with a loose pestle, 10 times with a tight pestle and lysed for 10 minutes on ice. 5 ml of chilled wash buffer (wash buffer; 5% BSA, 0.04U/µL RNase inhibitor, 0.25% glycerol) was added to the sample, passed through a 40 µm cell strainer and centrifuged at 500 x g for 5 minutes at 4ºC. After pelleting, the nuclei were resuspended in 5-10 ml of wash buffer. After two washes, single-nuclei suspensions were passed through a 20 µm cell strainer, pelleted, and resuspended in PBS with 0.05% BSA.

6.5.3 **Single-cell and single-cell RNA library preparation and sequencing**

Single cell and single nuclei suspensions were assessed with a trypan blue count. For each sample, 10,000-15,000 cells or nuclei were loaded using the Chromium Controller in combination with the Chromium Single Cell 3' V3 and V3.1 Gel Bead and Chip kits (10X Genomics). Individual cells or nuclei were partitioned into gel beads-in-emulsion (GEMS), followed by reverse transcription of barcoded RNA and cDNA amplification. Individual single-cell libraries with indices and Illumina P5/P7 adapters were generated with the Chromium Single Cell 3' Library kit and Chromium Multiplex kit. The libraries were sequenced on an Illumina Novaseq6000 sequencer.

6.5.4 **Human Medulloblastoma Tissue Collection (CHLA)**

6.5.5 **VSNL1 immunohistochemistry**

VSNL1 staining was performed on a Ventana BenchMark Ultra (Ventana Medical Systems, Tuscon, AZ) on 4-micron sections of paraffin-embedded medulloblastoma tissue. Briefly, slides were deparaffinized and underwent antigen retrieval protocol with cell conditioning 1 (CC1), followed by application of primary antibody (VSNL1, mouse monoclonal, OTI4A6, #MA5-26516, Thermo Fisher Scientific Inc.), at a dilution of 1:1600. 3,3’-diaminobenzidine (DAB) chromogen (ultraView Universal DAB detection kit,Ventana Medical Systems, Tucson, AZ) was used for visualization and counterstaining with hematoxylin was performed.

6.5.6 **Cyclic Immunofluorescence**

FFPE slides of medulloblastoma samples were processed before antibody staining by performing baking, deparaffinization, antigen retrieval, tissue permeabilization, autofluorescence photobleaching, and background imaging with DAPI staining. Slides were baked at 60 °C for one hour. Deparaffinization was completed with the following washes, each at three minutes: xylene (2x), 1:1 ratio of xylene to ethanol, ethanol (2x), 95% ethanol, 70% ethanol, 50% ethanol, RODI water (2x). Slides were submerged in 1x Tris-EDTA and heated in boiling water under pressure of a pressure cooker, then washed in 1x PBS for 10 minutes. The tissue was permeabilized with 1x PBS + 0.1% TritonX-100 for 10 minutes. Slides were washed in three changes of 1x PBS + 0.1% Tween20 for 10 minutes/wash.

Autofluorescence photobleaching was carried out by submerging slides in a solution of 4.5% hydrogen peroxide, 25 mM sodium hydroxide, and 1x PBS, while heating the samples and solution to 37 °C and exposing the tissue to direct full visible spectrum LED light. Slides were then washed three times in 1x PBS + 0.1% Tween20 for 10 minutes/wash.

Endogenous non-target proteins were blocked with a one-hour wash of Odyssey Blocking Buffer (PBS). Slides were then washed (1x PBS + 0.1% Tween20, 10 min), incubated in DAPI for 10 minutes, and then washed again (1x PBS + 0.1% Tween20, 10 min).

Glycerol (10% in 1x PBS) was used to mount coverslips onto the slides for imaging. Background imaging was captured with the same excitation and emission settings as was later used for fluorescent antibody imaging: (ex,em); D360/40x, ET460/50M; HQ480/40X, 535/50M; 560/40X, D630/60M; 628/40X, 692/40M.

Following background autofluorescence imaging with DAPI staining for registration, samples were incubated in the unconjugated primary antibodies AB1, AB2, AB3 overnight at 4 °C. Samples were then washed three times in 1x PBS + 0.1% Tween20 for 10 minutes/wash. Secondary antibodies with their respective fluorophores were added for one hour at room temperature in a dark humidity chamber.

Fluorescently-labeled tissue slides were imaged on a TE2000 inverted microscope using the excitation emission filters described above with 10x magnification and 0.30 NA lens with a resolution of 1.546 pixels/μm.

After the final round of IHC imaging, slides were photobleached and then stained with hematoxylin and eosin. The slides stained with H&E were imaged on an Aperio AT2 slide scanner at 40x magnification.

6.5.7 **Antibody Validation and Panels**

Antibodies were validated using reference tissues (Supplementary Figure 6C). Two CyCIF panels were used for this study and are described below. Panel 1 includes

taurine and was run on FFPE tissue from CHLA-5, CHLA-10, and BCH-96. The other CyCIF images were generated using Panel 2. MAP2, VSNL1, and Taurine were detected using secondary antibodies. Ki67 was directly conjugated to AF647 before purchasing, while CNTN1 was directly conjugated after purchase using the AF555 kit from Abcam (ab269820).

These antibodies allow for detecting stages of GN development. Ki67 is marker of the cycling progenitor cells (Gerdes et al., 1983, 1984) and MAP2 is expressed by postmitotic granule neurons (Przyborski and Cambray-Deakin, 1995). CNTN1 is a cell surface marker localized the GN dendrites and axons during development, but the dendritic expression is lost as the GNs mature (Virgintino et al., 1999). VSNL1 is a calcium-sensor expressed in GNs (Braunewell and Klein-Szanto, 2009) and our snRNA-Seq data indicates it expressed exclusively in late-stage GNs.

Panel 1:
Round 1: DAPI, MAP2 (488), VSNL1 (555), Taurine (647)
Round 2: DAPI, CNTN1 (555), Ki67 (647)

Panel 2:
Round 1: DAPI, MAP2 (488), VSNL1 (647)
Round 2: DAPI, CNTN1 (555), Ki67 (647)

6.5.8 **MALDI Slide preparation and Matrix Coating**

Fresh tissues were harvested and flash-frozen on dry ice, then stored at 80°C. 10µm-thin tissue sections were cut on a cryostat (Leica CM3050S, Wetzlar, Germany) in serial sections for MALDI and IF. Tissue sections were thaw-mounted on indium tin oxide (ITO)-coated glass slides (Bruker Daltonics, Bremen, Germany) and desiccated under vacuum for 10 mins before matrix coating.

Slides were subsequently sprayed with negative ionization matrix N-(1-Naphthyl) ethylenediamine dihydrochloride (NEDC, Sigma #222488) using an HTX TM-Sprayer (HTX Technologies, LLC). Concentration of 10 mg/ml was used in 70% MeOH. The sprayer parameter used was 80°C temperature, 0.1 ml/min flow rate, 1000 mm/min velocity, 2 mm track spacing, 10 psi pressure, and 3 liters/min gas flow rate. Slides were coated on the same day as MALDI imaging.

6.5.9 **MALDI imaging**

For MALDI-FTICR scanning, the matrix-coated slides were immediately loaded into a slide adapter (Bruker Daltonics, Bremen, Germany) and then into a solariX XR FTICR mass spectrometer with a 9.4T magnet (Bruker Daltonics, Bremen, Germany) with

resolving power of 120,000 at *m/z* 500. The laser focus was set to 'small,' and the *x-y* raster stepsize of 50µm was used using Smartbeam-II laser optics. A spectrum was accumulated from 200 laser shots at 1000 Hz. The ions were accumulated using the "cumulative accumulation of selected ions mode" (CASI) within an *m/z* range of 70-300 Daltons before being transferred to the ICR cell for a single scan.

6.5.10 **Hematoxylin and Eosin (H&E) staining for MALDI slides**

Histological staining was performed on the same slides after MALDI using Meyer and Briggs' Hematoxylin (Sigma #MHS32) and Eosin (Sigma #HT110332). Matrix was washed off and slide was fixed with cold MeOH for 5 minutes, washed with PBS 3 times and water, then submerged in hematoxylin for 15 minutes. Slides were transferred into warm water for 15 minutes, then dehydrated in 95% EtOH for 30 seconds, incubated with eosin for 1 minute, then dehydrated again stepwise with 95% for 1 minute and 100% EtOH for 1 minute, and cleared using Xylene for 2 minutes. Slides were mounted using Cytoseal 60 mounting media (Thermo Scientific #8310-4) and imaged and visualized using a Hamamatsu Nanozoomer with NDP.view2 software (Hamamatsu).

6.5.11 **snRNA-Seq and scRNA-Seq data processing**

Quality control was performed for each tumor individually. The filtered counts matrix was used to create a Seurat object (version 4.1.0) (Hao et al., 2021), removing any genes present in less than five nuclei and any nuclei that contain less than 300 features. Nuclei were removed from the dataset if they met any of the following criteria: below 5th percentile of UMI or features detected, above 95th percentile of UMI or features detected, mitochondrial genes represent more than 5% of the counts, or in the top 10% of DoubletScore determined by Scrublet (Wolock et al., 2019). These strong filters left 71,008 high-quality nuclei for analysis.

For the seven snRNA-seq samples, the filtered objects were merged and log-normalized using a scale factor of 10000. The top 2500 variable genes were identified and the number of UMI counts were regressed out during scaling. Integration was performed using *Harmony* (Korsunsky et al., 2019) and the top 50 dimensions were used for UMAP plotting and Louvain clustering (resolution = 0.25). Non-malignant cells were annotated using marker genes (Supplementary Figure 1A). Tumor cells were identified by their clustering pattern and expression of known SHH MB genes (Riemondy et al., 2022). Additionally, Gashemi *et al*. observed similar cell types in their cohort and detected copy number variations in these cells providing further evidence of their malignant status (personal communication). Clusters 6 and 12 were not included in re-clustering of malignant cells because each cluster was primarily associated with a single sample (MB4113 for 6 and MB2112 for 12) (Supplementary Figure 1B).

Supplementary Figure 1C details the re-clustering of the snRNA-seq MBEN samples. To generate this plot, nuclei in clusters 0,1,2,3,4 or 5 were selected and the first 50 harmony dimensions were used to calculate nearest neighbors and a UMAP plot. Louvain clustering was then performed on these cells with a resolution of 0.25. Almost every new cluster could be associated with a developmental stage using known marker genes (Figure 1D). Cluster 6 was merged with cluster 0 to represent the postmigratory GN. For cluster 4, the top marker genes are related to ribosomes. It is unknown whether this cluster corresponds to stage of GN development, so it was excluded from development-related figures. The high ribosomal content indicates these cells could be a sequencing artifact, but similar ribosomal cells are found in our scRNA-seq MBEN sample, the SHH-B2 cells from the Riemondy cohort (Riemondy et al., 2022) , and the P14 mouse cerebella (Vladoiu et al., 2019) (Supplementary Figure 2C) suggesting they may be biologically relevant.

The scRNA-seq datasets were processed using similar filters and parameters. The same quality control metrics were used, except the mitochondrial percentage filter was set to 25%. The two scRNA-seq datasets (one MBEN tumor and the P14 mouse dataset from GSE118068), were each analyzed by themselves, so no integration methods were used. Instead, these sample were log-normalized using a factor 10,000 and scaled, and then dimensionality reduction was performed using Principal Component Analysis on the 2500 most variable genes. The top 40 PCs were used for clustering and UMAP plots. For the SHH MBEN tumor, malignant cells were identified by their clustering pattern and expression of SHH MB markers, and these cells were re-clustered using a resolution of 0.2 (Supplementary Figure 1E). For the P14 mouse, the granule neuron lineage was identified using marker genes and these cells were re-clustered with a resolution of 0.3 and annotated using known markers of GN development (Supplementary Figure 2B).

6.5.12 **Pseudotime Analysis**

Pseudotime analysis was performed using *monocle3* (Trapnell et al., 2014) with a minimal branch length of 15 cells. The root node was set by selecting the nuclei from the cycling GCP cluster (cluster 5) with the highest value for UMAP component 2.

6.5.13 **Plotting with Seurat**

Plots from *Seurat* were made using the DimPlot and FeaturePlot functions, adding on specific parameters from *ggplot2* (Ginestet, 2011) that are described in the code. For all feature plots, the baseline plot was re-made with a custom function that orders the cells by their expression of the relevant gene to help ensure that cells expressing a given

marker are visible in the plot. Some feature plots use a minimum cutoff to highlight cells with high expression and these instances are described in the figure legends.

6.5.14 **Clustering of Gene Set Signature Scores**

Gene set signatures were created for 8 datasets (Supplementary Figure 2A): Archer proteomic subtypes (Archer et al., 2018), Consensus SHH MB subtypes (Cavalli et al., 2017), Korshunov MBEN transcriptional subtypes (Korshunov et al., 2020), human GN development (Okonechnikov et al., 2021), MBEN snRNA-seq, MBEN scRNA-seq, Riemondy SHH MB scRNA-seq (Riemondy et al., 2022), and P14 Mouse scRNA-seq (Vladoiu et al., 2019). For the snRNA-seq and scRNA-seq MBEN datasets, marker genes were identified using the *Seurat* FindMarkers function with default parameters. For these cases, we performed down-sampling before differential analysis so that each cluster was represented by an equal number of cells. The same procedure was used for the P14 mouse to identify the top differential genes for each stage of granule neuron development. Mouse genes were mapped to human orthologs using the biomartR package (Drost and Paszkowski, 2017). In a small number of instances, genes were not in the BioMart database. In those cases, if the capitalized version of the mouse gene was present in the MAGIC cohort transcriptomic data, that gene was also included. Otherwise, the gene was not included and the next highest-ranking gene took its place in the signature.

For the consensus subtypes, the transcriptomic microarray data from Cavalli *et al.* was rank normalized and marker genes were identified by performing t-tests between the subtype of interest and the other SHH tumors (Cavalli et al., 2017). For the Archer *et al.* subtypes, t-tests was used to identify the differential proteins between SHHa and SHHb (Archer et al., 2018). The markers for the Riemondy single-cell clusters and Korshunov *et al.* MBEN subtypes were taken from their respective supplemental materials (Supplementary Table 3 for Riemondy *et al.* and Supplementary Tables 1 & 2 for Korshunov *et al.*).

From these marker gene lists, a gene signature was created by taking the top n genes (50,100, 150, & 200). In some cases, n was larger than the number of marker genes for published datasets. When this occurred, all marker genes were included for the given signature. These signatures are summarized in Supplementary Table 2.

All genes sets of size n were used together as inputs for Gene Set Variation Analysis (GSVA), which takes in gene-level data and calculates enrichment scores for each gene set (Hänzelmann et al., 2013; Sonja et al., 2014). GSVA was run on transcriptomic data from the 223 SHH tumors of the MAGIC cohort (Cavalli et al., 2017). *ConsensusClusterPlus* was used to run consensus clustering on the activation scores to better understand what signatures are activated in the same samples (Wilkerson and Hayes, 2010). This method repeatedly subsamples items (i.e. gene signatures) and features (i.e. SHH tumors) and clusters the data to determine which signatures cluster

together. 1000 sub-samplings were run using the k-means algorithm and Euclidean distance metric. For each run, all of the gene signatures were used, but the SHH samples (i.e. features) were subsampled for the following percentages: 0.3, 0.5, 0.7 and 0.9. For each parameter set, the method was run from k=2 to k=10 and the optimal k was found to be 5 clusters for all parameter sets using an elbow plot.

The consensus clustering plots were made using the clustermap function in *seaborn* (Waskom, 2021) using "correlation" as the distance metric. All gene set sizes and subsampling percentages revealed similar high-level trend (Supplementary Figure 2D) with four primary groups representing developmental stages: cycling progenitors, non-cycling progenitors, premigratory GN, and migrating/postmigratory GN (late-stage GNs).

6.5.15 **Copy number variation (CNV) analysis**

Copy number variation data for chromosomal arms was included in the MAGIC cohort data from Cavalli *et al.* (Cavalli et al., 2017). T-tests were performed to determine associations between CNVs and GSVA gene set signatures from MBEN snRNA-seq cohort. The following gene sets were considered: Cycling GCP, GCP, Ribosomal, Premigratory GN, Migrating GN, Postmigratory GN, Proliferation (Cycling GCP score – GCP score), Progenitor (Cycling GCP score + GCP score), and Late-Stage GN (Migrating GN score + Postmigratory GN score). For each CNV/signature pair two t-tests were performed. First, all samples with a loss of the chromosome arm were compared with tumors having a WT or Gain, and then a second t-test was performed to compare the samples with a gain to the other tumors. If no tumors had a loss or gain, that test was not performed, leaving 648 comparisons. Multiple associations were significant using the Bonferroni-corrected alpha of 1.54 E -5. These analyses were performed using gene set signature scores based on 100 genes, but the relationships shown in Figure 3C are also significant for signatures of size 50, 100, and 200.

6.5.16 **Post-transcriptional regulation analysis**

Proteomic and transcriptomic data from Archer *et al.* (Archer et al., 2018) was re-analyzed, considering the 8674 genes identified in all tumors for both assays. For each sample, the protein and RNA values were rank-normalized from 0 to 1 and a rank difference (protein rank – RNA rank) was calculated for each gene as a proxy for post-transcriptional regulation. Thus, a gene with a high protein rank in a given sample and a low RNA rank would have a strong positive rank difference, suggesting possible post-transcriptional upregulation. Genes were categorized as synaptic if they appear in the GO_SYNAPSE gene set from MSIGDB (Subramanian et al., 2005) and as FMRP targets if they appear in the stringent list in Supplementary Figure 2A from Darnell *et al.* (Darnell et al., 2011).

For SHHb analysis, the mean protein rank, RNA rank, and rank difference were calculated for every gene by taking the average value for the five SHHb tumors. The same procedure was applied to the nine SHHa tumors for SHHa analysis. Synaptic and non-synaptic gene ranks were compared using a two sample t-test in *scipy* (Oliphant, 2007). To determine potential post-transcriptional regulation, rank differences were compared to mean of 0 using a one-sample t-test.

6.5.17 **Image registration and processing**

Individual images were taken with 10% horizontal and vertical overlap and then stitched together with the Microscopy Image Stitching Tool (MIST) (Chalfoun et al., 2017) through *FIJI* (Schindelin et al., 2012). Image registration between CyCIF rounds was performed using the *MultiStackReg* plugin in *FIJI* (Schindelin et al., 2012) using the DAPI channels and the rigid body transformation. When the two images were not identical sizes, the two images were cropped slightly around the edges to enforce identical sizing. In a small number of select images, there were clear visual artifacts with extremely high fluorescence. In such cases, a mask was created and subtracted from the original image.

6.5.18 **MALDI Data Processing**

MALDI intensity data was analyzed for 4677 consensus m/z peaks identified by the *Isoscope* package (Wang et al., 2022). For each m/z value in each MALDI spot, the intensity was assigned to the maximum intensity of peaks within 2 PPM of the m/z value. These data were then loaded into a *scanpy* object (Wolf et al., 2018), and each spot was normalized using the total ion count (TIC) method, whereby every m/z intensity is divided by the sum of the intensities for that spot. The data was then scaled, and dimensionality reduction was performed using PCA. The spots were clustering using leiden clustering with a resolution of 0.2 and considering the 10 nearest neighbors and first 10 PCs. Some clusters (2,4,9,10,11,12,13,&14) were primarily around the edges and thus removed as they were likely artifacts. Additionally, there were clusters of spots in each sample that were visible artifacts compared to H&E staining and these clusters were removed as well. This resulted in 52,393 high-quality spots that strongly resemble the tissue structure from H&E stains (Supplementary Figure 6A).

Our analysis was focused on high-quality annotated metabolites. First, m/z peaks were associated with known metabolites using the database from Supplementary Table 4 and a window of 2 PPM. These features were then further filtered to only include metabolites that have an intensity greater than 100,000 in more than 50% of the MALDI spots. These cutoffs produced 56 high-quality metabolites for network and correlation analysis.

6.5.19 **Joint graphical lasso**

Joint graphical lasso (Danaher et al., 2014) analysis was performed using the *gglasso* python package (version 0.1.9). Graphical lasso (Friedman et al., 2008) uses observational data to learn a sparse approximation of the precision matrix, where each 0 represents two metabolites whose expression levels are independent of each other when considering the intensities of the other metabolites. This leaves of matrix of non-zero values, which can be represented as a network with edges between conditionally dependent metabolites. Joint graphical lasso allows for approximating multiple precision matrices at one time, while sharing information between the related datasets.

The group graphical lasso algorithm was used for the TIC-normalized MALDI for the 56 annotated metabolites. A parameter sweep was performed for lambda1, which controls sparsity, and lambda2, which promotes similarity across networks, using the following parameters: lambda1 (0.05, 0.1, 0.15, 0.2, 0.25) and lambda2 (0.01, 0,02, 0,03, 0,04 and 0.05). The optimal parameters (lambda1 = 0.1, and lambda2 = 0.01) were the minimal empirical bayes score, which was calculated using the ebic function from the *gglasso* package with a gamma parameter of 20 to promote sparser networks.

A consensus network was created for the sections with late-stage GNs by considering any edge present in at least 50% of those sections and the same procedure was applied to generate a network for the other four sections. Betweenness centrality was calculated for these consensus networks using the *networkx* python package (version 2.6.3). The union of the two consensus networks was output to a gml file and uploaded into Cytoscape (version 3.7.2) (Shannon et al., 2003), which was used to create Figure 6A.

6.5.20 **Bivariate Moran's I analysis**

Bivariate Moran's I analysis was performed using the *pysal* (version 2.4.0) package to analyze what metabolites correlate with the spatial lag of taurine. Weights were determined using the 24 nearest spots (i.e. two levels out on the MALDI spot grid) and then row normalized. The Local Moran's Bivariate I statistic was calculated between guanine and taurine for each spot using the Moran_Local_BV function. The resulting local bivariate Moran's I values were plotted with *seaborn* with error bars determined through the default bootstrapping approach implemented in the package.

6.6 References

Abe, H., Okazawa, M., and Nakanishi, S. (2011). The Etv1/Er81 transcription factor orchestrates activity-dependent gene regulation in the terminal maturation program of

cerebellar granule cells. Proc. Natl. Acad. Sci. U. S. A.
Adachi, T., Miyashita, S., Yamashita, M., Shimoda, M., Okonechnikov, K., Chavez, L., Kool, M., Pfister, S.M., Inoue, T., Kawauchi, D., et al. (2021). Notch signaling between cerebellar granule cell progenitors. ENeuro.
Akazawa, C., Shigemoto, R., Bessho, Y., Nakanishi, S., and Mizuno, N. (1994). Differential expression of five N-methyl-D-aspartate receptor subunit mRNAs in the cerebellum of developing and adult rats. J. Comp. Neurol.
Antar, L.N., Li, C., Zhang, H., Carroll, R.C., and Bassell, G.J. (2006). Local functions for FMRP in axon growth cone motility and activity-dependent regulation of filopodia and spine synapses. Mol. Cell. Neurosci.
Archer, T.C., Ehrenberger, T., Mundt, F., Gold, M.P., Krug, K., Mah, C.K., Mahoney, E.L., Daniel, C.J., LeNail, A., Ramamoorthy, D., et al. (2018). Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups. Cancer Cell.
Bennett, C.D., Kohe, S.E., Gill, S.K., Davies, N.P., Wilson, M., Storer, L.C.D., Ritzmann, T., Paine, S.M.L., Scott, I.S., Nicklaus-Wollenteit, I., et al. (2018). Tissue metabolite profiles for the characterisation of paediatric cerebellar tumours. Sci. Rep.
Borghesani, P.R., Peyrin, J.M., Klein, R., Rubin, J., Carter, A.R., Schwartz, P.M., Luster, A., Corfas, G., and Sergal, R.A. (2002). BDNF stimulates migration of cerebellar granule cells. Development.
Braunewell, K.-H., and Klein-Szanto, A.J. (2009). Visinin-like proteins (VSNLs): interaction partners and emerging functions in signal transduction of a subfamily of neuronal Ca2+-sensor proteins. Cell Tissue Res.
Breitman, T.R., Selonick, S.E., and Collins, S.J. (1980). Induction of differentiation of the human promyelocytic leukemia cell line (HL-60) by retinoic acid. Proc. Natl. Acad. Sci. U. S. A.
Cathala, L., Misra, C., and Cull-Candy, S. (2000). Developmental profile of the changing properties of NMDA receptors at cerebellar mossy fiber-granule cell synapses. J. Neurosci.
Cavalli, F.M.G., Remke, M., Rampasek, L., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A.S., et al. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell.
Chalfoun, J., Majurski, M., Blattner, T., Bhadriraju, K., Keyrouz, W., Bajcsy, P., and Brady, M. (2017). MIST: Accurate and Scalable Microscopy Image Stitching Tool with Stage Modeling and Error Minimization. Sci. Rep.
Cheng, Y., Liao, S., Xu, G., Hu, J., Guo, D., Du, F., Contreras, A., Cai, K.Q., Peri, S., Wang, Y., et al. (2020). NeuroD1 Dictates Tumor Cell Differentiation in Medulloblastoma. Cell Rep.
Cicconi, L., and Lo-Coco, F. (2016). Current management of newly diagnosed acute promyelocytic leukemia. Ann. Oncol.
Consalez, G.G., Goldowitz, D., Casoni, F., and Hawkes, R. (2021). Origins, Development, and Compartmentation of the Granule Cells of the Cerebellum. Front. Neural Circuits.
Cull-Candy, S.G., Brickley, S.G., Misra, C., Feldmeyer, D., Momiyama, A., and Farrant, M. (1998). NMDA receptor diversity in the cerebellum: Identification of subunits contributing to functional receptors. In Neuropharmacology, p.

Danaher, P., Wang, P., and Witten, D.M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. J. R. Stat. Soc. Ser. B Stat. Methodol.
Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y.S., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W., et al. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell.
Ding, J., Adiconis, X., Simmons, S.K., Kowalczyk, M.S., Hession, C.C., Marjanovic, N.D., Hughes, T.K., Wadsworth, M.H., Burks, T., Nguyen, L.T., et al. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat. Biotechnol.
Drost, H.G., and Paszkowski, J. (2017). Biomartr: Genomic data retrieval with R. Bioinformatics.
Eberhart, C.G., Kaufman, W.E., Tihan, T., and Burger, P.C. (2001). Apoptosis, neuronal maturation, and neurotrophin expression within medulloblastoma nodules. J. Neuropathol. Exp. Neurol.
Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics.
Gajjar, A., Robinson, G.W., Smith, K.S., Lin, T., Merchant, T.E., Chintagumpala, M., Mahajan, A., Su, J., Bouffet, E., Bartels, U., et al. (2021). Outcomes by Clinical and Molecular Features in Children With Medulloblastoma Treated With Risk-Adapted Therapy: Results of an International Phase III Trial (SJMB03). J. Clin. Oncol.
Garcia-Lopez, J., Kumar, R., Smith, K.S., and Northcott, P.A. (2021). Deconstructing Sonic Hedgehog Medulloblastoma: Molecular Subtypes, Drivers, and Beyond. Trends Genet.
Gerdes, J., Schwab, U., Lemke, H., and Stein, H. (1983). Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. Int. J. Cancer.
Gerdes, J., Lemke, H., Baisch, H., Wacker, H.H., Schwab, U., and Stein, H. (1984). Cell cycle analysis of a cell proliferation-associated human nuclear antigen defined by the monoclonal antibody Ki-67. J. Immunol.
Ginestet, C. (2011). ggplot2: Elegant Graphics for Data Analysis. J. R. Stat. Soc. Ser. A (Statistics Soc.
Hänzelmann, S., Castelo, R., Guinney, J., Kim, S.C., Seo, Y.J., Chung, W., Eum, H.H., Nam, D.-H., Kim, J., Joo, K.M., et al. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics *14*, 7.
Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell.
Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.K.L., Baumgartner, A., DeWitt, J.C., Groves, A., Mayr, L., Weisman, H.R., et al. (2019). Resolving medulloblastoma cellular architecture by single-cell genomics. Nature.
Kerjan, G., Dolan, J., Haumaitre, C., Schneider-Maunoury, S., Fujisawa, H., Mitchell, K.J., and Chédotal, A. (2005). The transmembrane semaphorin Sema6A controls cerebellar granule cell migration. Nat. Neurosci.
Kim, J.Y.H., Nelson, A.L., Algon, S.A., Graves, O., Sturla, L.M., Goumnerova, L.C., Rowitch, D.H., Segal, R.A., and Pomeroy, S.L. (2003). Medulloblastoma tumorigenesis diverges from cerebellar granule cell differentiation in patched heterozygous mice. Dev.

Biol.
Komuro, H., and Rakic, P. (1995). Dynamics of granule cell migration: A confocal microscopic study in acute cerebellar slice preparations. J. Neurosci.
Komuro, H., and Yacubova, E. (2003). Recent advances in cerebellar granule cell migration. Cell. Mol. Life Sci.
Komuro, H., Yacubova, E., Yacubova, E., and Rakic, P. (2001). Mode and tempo of tangential cell migration in the cerebellar external granular layer. J. Neurosci.
Komuro, Y., Fahrion, J.K., Foote, K.D., Fenner, K.B., Kumada, T., Ohno, N., and Komuro, H. (2013). Granule cell migration and differentiation. In Handbook of the Cerebellum and Cerebellar Disorders, p.
Kool, M., Korshunov, A., Remke, M., Jones, D.T.W., Schlanstein, M., Northcott, P.A., Cho, Y.J., Koster, J., Schouten-Van Meeteren, A., Van Vuurden, D., et al. (2012). Molecular subgroups of medulloblastoma: An international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. Acta Neuropathol.
Kool, M., Jones, D.T.W., Jäger, N., Northcott, P.A., Pugh, T.J., Hovestadt, V., Piro, R.M., Esparza, L.A., Markant, S.L., Remke, M., et al. (2014). Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer Cell.
Korshunov, A., Okonechnikov, K., Sahm, F., Ryzhova, M., Stichel, D., Schrimpf, D., Ghasemi, D.R., Pajtler, K.W., Antonelli, M., Donofrio, V., et al. (2020). Transcriptional profiling of medulloblastoma with extensive nodularity (MBEN) reveals two clinically relevant tumor subsets with VSNL1 as potent prognostic marker. Acta Neuropathol.
Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. ru, and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods.
Legué, E., Riedel, E., and Joyner, A.L. (2015). Clonal analysis reveals granule cell behaviors and compartmentalization that determine the folded morphology of the cerebellum. Dev.
Lin, J.R., Fallahi-Sichani, M., Chen, J.Y., and Sorger, P.K. (2016). Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. Curr. Protoc. Chem. Biol.
Lin, J.R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P.M., Santagata, S., and Sorger, P.K. (2018). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. Elife.
Losi, G., Prybylowski, K., Fu, Z., Luo, J.H., and Vicini, S. (2002). Silent synapses in developing cerebellar granule neurons. J. Neurophysiol.
Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., et al. (2021). The 2021 WHO classification of tumors of the central nervous system: A summary. Neuro. Oncol.
Maar, T., Morán, J., Schousboe, A., and Pasantes-Morales, H. (1995). Taurine deficiency in dissociated mouse cerebellar cultures affects neuronal migration. Int. J. Dev. Neurosci.
Miyata, T., Maeda, T., and Lee, J.E. (1999). NeuroD is required for differentiation of the granule cells in the cerebellum and hippocampus. Genes Dev.
Miyazaki, T., Fukaya, M., Shimizu, H., and Watanabe, M. (2003). Subtype switching of

vesicular glutamate transporters at parallel fibre-Purkinje cell synapses in developing mouse cerebellum. Eur. J. Neurosci.
Nagy, C., Maitra, M., Tanti, A., Suderman, M., Théroux, J.F., Davoli, M.A., Perlman, K., Yerko, V., Wang, Y.C., Tripathy, S.J., et al. (2020). Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Nat. Neurosci.
Napoli, I., Mercaldo, V., Boyl, P.P., Eleuteri, B., Zalfa, F., De Rubeis, S., Di Marino, D., Mohr, E., Massimi, M., Falconi, M., et al. (2008). The Fragile X Syndrome Protein Represses Activity-Dependent Translation through CYFIP1, a New 4E-BP. Cell.
Northcott, P.A., Jones, D.T.W., Kool, M., Robinson, G.W., Gilbertson, R.J., Cho, Y.J., Pomeroy, S.L., Korshunov, A., Lichter, P., Taylor, M.D., et al. (2012). Medulloblastomics: The end of the beginning. Nat. Rev. Cancer.
Okonechnikov, K., Joshi, P., Sepp, M., Leiss, K., Sarropoulos, I., Murat, F., Sill, M., Beck, P., Chan, K.C.-H., Korshunov, A., et al. (2021). Mapping pediatric brain tumors to their origins in the developing cerebellum. BioRxiv 2021.12.19.473154.
Oliphant, T.E. (2007). SciPy: Open source scientific tools for Python. Comput. Sci. Eng.
Olivier, T.W., Bass, J.K., Ashford, J.M., Beaulieu, R., Scott, S.M., Schreiber, J.E., Palmer, S., Mabbott, D.J., Swain, M.A., Bonner, M., et al. (2019). Cognitive implications of ototoxicity in pediatric patients with embryonal brain tumors. J. Clin. Oncol.
Orr, B.A. (2020). Pathology, diagnostics, and classification of medulloblastoma. Brain Pathol.
Przyborski, S.A., and Cambray-Deakin, M.A. (1995). Developmental regulation of MAP2 variants during neuronal differentiation in vitro. Dev. Brain Res.
Rakic, P. (1971). Neuron-glia relationship during granule cell migration in developing cerebellar cortex. A Golgi and electonmicroscopic study in Macacus rhesus. J. Comp. Neurol.
Riemondy, K.A., Venkataraman, S., Willard, N., Nellan, A., Sanford, B., Griesinger, A.M., Amani, V., Mitra, S., Hankinson, T.C., Handler, M.H., et al. (2022). Neoplastic and immune single-cell transcriptomics define subgroup-specific intra-tumoral heterogeneity of childhood medulloblastoma. Neuro. Oncol.
Salloum, R., Chen, Y., Yasui, Y., Packer, R., Leisenring, W., Wells, E., King, A., Howell, R., Gibson, T.M., Krull, K.R., et al. (2019). Late morbidity and mortality among medulloblastoma survivors diagnosed across three decades: A report from the Childhood Cancer Survivor Study. J. Clin. Oncol.
Sato, M., Suzuki, K., Yamazaki, H., and Nakanishi, S. (2005). A pivotal role of calcineurin signaling in development and maturation of postnatal cerebellar granule cells. Proc. Natl. Acad. Sci. U. S. A.
Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: An open-source platform for biological-image analysis. Nat. Methods.
Schüller, U., Heine, V.M., Mao, J., Kho, A.T., Dillon, A.K., Han, Y.G., Huillard, E., Sun, T., Ligon, A.H., Qian, Y., et al. (2008). Acquisition of Granule Neuron Precursor Identity Is a Critical Determinant of Progenitor Cell Competence to Form Shh-Induced Medulloblastoma. Cancer Cell.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for

integrated models of biomolecular interaction networks. Genome Res.
Sonja, H., Castelo, R., and Guinney, J. (2014). GSVA : The Gene Set Variation Analysis package for microarray and RNA-seq data. Bioconductor.Org 1–20.
Stoeckli, M., Chaurand, P., Hallahan, D.E., and Caprioli, R.M. (2001). Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. Nat. Med.
Sturman, J.A., Moretz, R.C., French, J.H., and Wisniewski, H.M. (1985). Taurine deficiency in the developing cat: Persistence of the cerebellar external granule cell layer. J. Neurosci. Res.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M. a, Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.
Suzuki, K., Sato, M., Morishima, Y., and Nakanishi, S. (2005). Neuronal depolarization controls brain-derived neurotrophic factor-induced upregulation of NR2C NMDA receptor via calcineurin signaling. J. Neurosci.
Takahashi, T., Feldmeyer, D., Suzuki, N., Onodera, K., Cull-Candy, S.G., Sakimura, K., and Mishina, M. (1996). Functional correlation of NMDA receptor ε subunits expression with the properties of single-channel and synaptic currents in the developing cerebellum. J. Neurosci.
Tallan, H.H. (1957). Studies on the distribution of N-acetyl-L-aspartic acid in brain. J. Biol. Chem.
Tallan, H.H., Moor, S., and Stein, W.H. (1956). N-Acetyl-L-aspartic acid in brain. J. Biol. Chem.
Tárnok, K., Czöndör, K., Jelitai, M., Czirók, A., and Schlett, K. (2008). NMDA receptor NR2B subunit over-expression increases cerebellar granule cell migratory activity. J. Neurochem.
Taylor, M.D., Northcott, P.A., Korshunov, A., Remke, M., Cho, Y.J., Clifford, S.C., Eberhart, C.G., Parsons, D.W., Rutkowski, S., Gajjar, A., et al. (2012). Molecular subgroups of medulloblastoma: The current consensus. Acta Neuropathol.
Teider, N., Scott, D.K., Neiss, A., Weeraratne, S.D., Amani, V.M., Wang, Y., Marquez, V.E., Cho, Y.J., and Pomeroy, S.L. (2010). Neuralized1 causes apoptosis and downregulates Notch target genes in medulloblastoma. Neuro. Oncol.
Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol.
Virgintino, D., Ambrosini, M., D'Errico, P., Bertossi, M., Papadaki, C., Karagogeos, D., and Gennarini, G. (1999). Regional distribution and cell type-specific expression of the mouse F3 axonal glycoprotein: A developmental study. J. Comp. Neurol.
Vladoiu, M.C., El-Hamamy, I., Donovan, L.K., Farooq, H., Holgado, B.L., Sundaravadanam, Y., Ramaswamy, V., Hendrikse, L.D., Kumar, S., Mack, S.C., et al. (2019). Childhood cerebellar tumours mirror conserved fetal transcriptional programs. Nature.
Wallace, V.A. (1999). Purkinje-cell-derived Sonic hedgehog regulates granule neuron precursor cell proliferation in the developing mouse cerebellum. Curr. Biol.
Wang, L., Xing, X., Zeng, X., Jackson, S.R.E., TeSlaa, T., Al-Dalahmah, O., Samarah,

L.Z., Goodwin, K., Yang, L., McReynolds, M.R., et al. (2022). Spatially resolved isotope tracing reveals tissue metabolic activity. Nat. Methods.
Waskom, M. (2021). seaborn: statistical data visualization. J. Open Source Softw.
Watanabe, M., Mishina, M., and Inoue, Y. (1994). Distinct spatiotemporal expressions of five NMDA receptor channel subunit mRNAs in the cerebellum. J. Comp. Neurol.
Wechsler-Reya, R.J., and Scott, M.P. (1999). Control of neuronal precursor proliferation in the cerebellum by sonic hedgehog. Neuron.
Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. Bioinformatics.
Wilson, M., Davies, N.P., Brundler, M.A., McConville, C., Grundy, R.G., and Peet, A.C. (2009). High resolution magic angle spinning 1H NMR of childhood brain and nervous system tumours. Mol. Cancer.
Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. Genome Biol.
Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst.
Xenaki, D., Martin, I.B., Yoshida, L., Ohyama, K., Gennarini, G., Grumet, M., Sakurai, T., and Furley, A.J.W. (2011). F3/contactin and TAG1 play antagonistic roles in the regulation of sonic hedgehog-induced cerebellar granule neuron progenitor proliferation. Development.
Yang, Z.J., Ellis, T., Markant, S.L., Read, T.A., Kessler, J.D., Bourboulas, M., Schüller, U., Machold, R., Fishell, G., Rowitch, D.H., et al. (2008). Medulloblastoma Can Be Initiated by Deletion of Patched in Lineage-Restricted Progenitors or Stem Cells. Cancer Cell.
Zhou, P., Porcionatto, M., Pilapil, M., Chen, Y., Choi, Y., Tolias, K.F., Bikoff, J.B., Hong, E.J., Greenberg, M.E., and Segal, R.A. (2007). Polarized Signaling Endosomes Coordinate BDNF-Induced Chemotaxis of Cerebellar Precursors. Neuron.
Zhu, G., Rankin, S.L., Larson, J.D., Zhu, X., Chow, L.M.L., Qu, C., Zhang, J., Ellison, D.W., and Baker, S.J. (2017). PTEN signaling in the postnatal perivascular progenitor niche drives medulloblastoma formation. Cancer Res.

Chapter 7 - Conclusions

This thesis highlights many applications for machine learning methods in neurological disease research. In this chapter, I detail the main conclusions, limitations, and potential future directions for each project.

7.1 Shallow Sparsely-Connected Autoencoders

Chapter 2 discusses SSCA, on a novel auto-encoder based method for gene set scoring. This method converts gene level expression data into scores for predefined gene sets, while simultaneously creating a biologically informed latent space that can be utilized for clustering or visualization. The gene set scores produced by SSCA lead to better supervised and unsupervised classification of blood cell types compared to existing methods. Additionally, the version of SSCA that uses a variational autoencoder (SSCVA) consistently outperforms the method using a standard autoencoder.

This work represents an exciting new algorithm for gene set scoring, but more work is necessary to characterize the advantages and limitations of these autoencoder-based approaches. First, we need to investigate the best way to use these autoencoders to get final gene set scores. In this thesis, I simply performed a parameter sweep over autoencoder parameters and used the encoder function from the optimal network to get gene set scores. Follow-up work has shown that the scores can be affected by the number of gene sets and their individual sizes, so better scoring metrics should account for these features. One potential solution would be through permutation testing, where the optimal network is retrained many times using shuffled edges and a gene set score is based on its activation in the real gene set node compared to the scores from the null distribution of randomly created gene sets.

The other primary limitation of this method is run time. Even though we have shown evidence that SSCVA can produce more meaningful latent spaces for scRNA-seq data than ssGSEA or GSVA, it still takes significantly longer to run than these rank-based methods. The biggest hurdle is the time it takes for hyperparameter optimization and network training. The run time of SSCVA limits its utility and engineering work to speed up identification of optimal parameters would greatly improve this method.

When this method was developed in 2018, there were not many published scRNA-seq studies from neurological conditions. In the last few years, however, many groups have published scRNA-seq datasets from the diseases studied in this thesis (Al-Dalahmah et al., 2020; Neftel et al., 2019; Vladoiu et al., 2019) and from other neurological conditions, such as Alzheimer's Disease (Mathys et al., 2019). Future work can apply the SSCVA method to study heterogeneity in these neurological conditions.

7.2 Cycling Cells in iPSC-derived neurons from HD

Chapter 3 focuses on multi-omic analysis of iPSC-derived neurons from HD and control samples. The most significant finding in this work is the identification of a neural stem cell like population of cycling cells that are observed in the HD models but not the control models. This finding is supported by epigenomic analysis, where ChIP-Seq for H3K4me3 shows that sites of active transcription are enriched for TFs related to the cell cycle. Additionally, validation experiments show that WNT signaling is driving the cell cycling in HD models.

These findings raise important questions about the developmental effects of the *HTT* mutation, but the clinical implications of these results are limited given the model system being used. It is not known whether this population of cycling cells exists in human HD or even mouse HD models, as it may simply be a product of the differentiation protocols being used. It will be crucial to follow up on these results by analyzing single-cell transcriptomics data from HD mouse models and human postmortem brains to determine if an analogous cell population is present. Additionally, single-cell ATAC-Seq (scATAC-Seq) could be used to better understand this cell population. One possible experiment would be to compare the cycling cells from HD iPSC-derived neurons to similar cells that exist from an earlier stage of differentiation in the control model. This would first require a longitudinal experiment, staining for markers of the cycling NSCs, to identify when these cells appear. Once this is determined, we could perform scATAC-Seq on the cells from this stage in control models and on the HD iPSC-derived neurons. Comparing the cycling cells in the HD model with the healthy NSCs, could reveal differentially accessible genomic regions and help us better understand the developmental implications of the *HTT* mutation.

7.3 SamNet 2.0 For Drug Response Analysis in Glioblastoma

Chapter 4 details subtype-specific responses to dasatinib in glioblastoma cell lines. Proteomic and phosphoproteomic data was collected from treated and untreated cell lines and the results highlight features of cell cycle inhibition in the sensitive mesenchymal cell lines, but not the resistant proneural cell lines. Additionally, an shRNA viability screen was performed in treated and untreated cells to identify targets for improving dasatinib efficacy in proneural samples. There were many significant screen hits, so we used a substantially improved version of the SamNet multicommodity flow algorithm to help prioritize targets for validation. This new SamNet 2.0 method discourages flow through hub nodes and incorporates baseline expression data to produce more realistic final networks. The resulting proneural network was significantly enriched for cell cycle genes, which suggested a focus on the shRNA hit WEE1. Validation experiments found that WEE1 inhibition significantly improved dasatinib efficacy in proneural models, but not in mesenchymal cell lines.

This work has biological and computational implications. For glioblastoma, this research presents an exciting new direction for potentially treating proneural glioblastomas. Still, it is not known how these results will translate to actual patients since the experiments

were performed in cell line models. The next step in this project would be to reproduce the validation experiments using mouse models of proneural glioblastoma.

The biggest computational contribution from this work was the newly developed SamNet 2.0 algorithm. The most significant change is the penalty term that discourages flow through hub nodes. Without this penalty, 33% of network flow goes through high-degree hub nodes, significantly biasing the final solutions toward highly-connected proteins that may not have any biological relevance. Additionally, SamNet 2.0 includes a term that allows for the incorporation of baseline RNA expression data in the algorithm. This biases the final network solutions away from genes with low expression, leading to output networks that favor genes that are expressed in a given subtype.

In addition to these changes, future work can make even more improvements to the SamNet framework. One project that was considered, but ultimately not completed, was to allow for the incorporation of phosphoproteomic data into the method. This would work by having each phosphosite assigned to a node in the directed network. These phosphosite nodes would have an incoming edge from any kinase that affected it and an outgoing edge to the associated protein (e.g. PKA -> FYN@S21 -> FYN). To make this work with the multicommodity flow framework, the incoming edges for phosphosites would have constraints, only allowing for flow if the phosphosite was a hit for the given subtype. This addition would allow for an even more comprehensive understanding of the subtype-specific effects of dasatinib and could serve as a framework for other phosphoproteomic studies.

7.4 Medulloblastoma Heterogeneity

Chapters 5 and 6 present significant advancements in our understanding of medulloblastoma tumor heterogeneity. Chapter 5 details proteomic and phosphoproteomic analysis of MB tumors. There are clear subdivisions in the SHH and G3 subtypes that are not observed when clustering RNA or methylation data. The Prize-Collecting Steiner Forest algorithm was helpful in integrating multiple omics data types to better understand the drivers of these new proteomic clusters. The SHH tumors are divided into SHHa, which resembles standard SHH-activated MB, and SHHb, which are enriched for genes related to neuronal pathways, like glutamatergic synapses. The Group 3 samples split into G3a and G3b, where the G3a samples has higher levels of phosphorylated MYC and worse outcomes than the G3b tumors.

Chapter 6 details a snRNA-seq study of medulloblastoma tumors with MBEN histology. These tumors contain cells that recapitulate cerebellar granule neuron development and I used computational approaches to connect these findings to existing single-cell studies and molecular subtypes of MB. The consensus subtypes of SHH MB are each associated with a specific developmental state and the SHHb proteomic subtype discussed in Chapter 5 is likely caused by the presence of tumor cells mimicking postmigratory granule neurons. We also showed this developmental perspective helps explain genomic, histologic, and metabolic heterogeneity in SHH MB.

These medulloblastoma projects are the most significant contributions from this thesis. The proteomic subtypes presented in Chapter 5 detail novel groups of MB tumors. There are clear implications for the Group 3 split as G3a samples have significantly worse outcomes, but it is unknown whether the SHHa/SHHb split is clinically relevant. The primary limitation of this work is the small sample size (45 tumors total), so a larger proteomics cohort is necessary to test the robustness of these subtypes and to better understand any potential clinical relevance of the SHH tumors.

The finding from Chapter 6 that SHH-activated MB tumor cells retain their capacity to follow granule neuron differentiation has led to meaningful insights about SHH MB tumor heterogeneity. Like the proteomics study, these findings are also based on a small sample size. This is understandable since the work focuses on one rare histological subtype of a rare brain tumor, but future validation experiments on additional tumors would be valuable. The most important implication of this work is the possibility of using these developmental findings for designing therapeutics. There is great interest in treating SHH MB by inducing tumor cells to differentiate into non-malignant neurons and this thesis suggests that we can use principles from canonical human granule neuron development for target identification. We present evidence that factors like CNTN1 and taurine, which drive granule neuron differentiation in the cerebellum, are also highly expressed in these tumors. Still, validation studies are required to show whether these molecules (or similar ones) could be used to induce tumor cell differentiation. It is particularly challenging to perform these validation experiments as there are not great cell line models for SHH MB and it is not known if the mouse models recapitulate this developmental phenotype. The next step in this project would be to analyze scRNA-seq data from MB mouse models to better understand the developmental potential of these cells. If these experiments show promising results, validation experiments can be performed on MB mice to test molecules for their capacity to induce tumor cell differentiation.

7.5 Overall Conclusions

For this thesis, I want to emphasize two primary conclusions. The first is that it is incredibly important to consider multiple biological data types when studying complex diseases. I show many examples of how the collection and integration of multiple biological data types yielded conclusions not possible from a single modality. The second conclusion is the importance of machine learning methods for neurological disease research. It can be extremely challenging to generate hypotheses from such small, noisy datasets and these methods are a valuable tool for gaining biological insights about these complicated conditions.