The Dissertation Committee for Amelia Weber Hall Certifies that this is the approved version of the following dissertation:

From genome to genotype: regulation of the genome in glioblastoma multiforme and atrial fibrillation

Committee:

Vishwanath R. Iyer, Supervisor

Haley O. Tucker

Jon Huibregtse

Lauren I. R. Ehrlich

Claus O. Wilke

From genome to genotype: regulation of the genome in glioblastoma multiforme and atrial fibrillation

by

Amelia Weber Hall, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2017

Dedication

Dedicated *in memoriam* to my mother,

Margaret Skinner Weber.

I’ll always be your little scientist.

Acknowledgements

None of this work would be possible without the support of my advisor, Dr. Vishy Iyer. He has gone above and beyond in his patience and support of me as a scientist and as a person. Many current and former members of the Iyer lab helped me find my footing when I joined, and facilitated my learning the plethora of wetlab and drylab techniques required to be an effective researcher in genomics. I am grateful for Dr. Bum-kyu Lee, Dr. Yunyun Ni, Dr. Damon Polioudakis, Dr. Daechan Park, Dr. Yaelim Lee, Dr. Dia Bagchi, Nathan Abell, Haridha Shivram and Anna Battenhouse for all of their help. Anna reintroduced me to the world of Unix and programming, and very patiently taught me how to be a competent analyst of bioinformatics data.

My committee members Dr. Haley Tucker, Dr. Jon Huibregtse, Dr. Lauren Ehrlich, and Dr. Claus Wilke have been advising me for several years, and I thank them profusely for their knowledge and kindness. Lauren and Haley in particular have always given me a lot of encouragement on my research, even when I felt hopeless or the work felt intractable, and that means a lot.

Over the years, I have engaged in a number of fruitful collaborations, both at UT, and at St David's Medical Center. The efforts from these collaborations to generate and analyze data have expanded my worldview, and exposed me to clinically oriented contexts in genomics. At St David's, I am grateful to Max Shpak, Matt Cowperthwaite and the staff at the NeuroTexas Institute, as well as Mitra Mohanty and the staff at Texas Cardiac Arrhythmia Institute. At UT Austin, I am grateful to the Aldrich lab, especially

Dr. Rick Aldrich, Dr. Tom Middendorf, Dr. D. Brent Halling, and former member Dr. Jenni Greeson-Bernier. You believed in me, even when I didn't know how to believe in myself. I am grateful to the Vokes lab, for years of helping me understand the proper context of developmental biology, and the Marcotte lab for being good friends and pushing me to consider my data through a more systems-oriented approach.

My family has always been very supportive of my scientific leaning, particularly my Mom and Dad, who always encouraged me to keep challenging myself and never shy away of learning things that were "too difficult." My sister Jane is always willing to listen to me ramble on about genomics and cancer, although her first love is botany. My close friends Christy, Emily, and Justin keep me optimistic and make sure I laugh regularly. Finally, my partner Dr. Brian McCann has been with me through everything: I would not be here today without his love and support.

From genome to genotype: regulation of the genome in glioblastoma multiforme and atrial fibrillation

Amelia Weber Hall, Ph.D.

The University of Texas at Austin, 2017

Supervisor: Vishwanath R. Iyer

The modern era of genomics has made sequencing a genome nearly routine. Genomics has amassed huge amounts of somatic and disease mutation data, as a result, the character sequence of the human genome has been extensively studied. This information is having an impact on the standard of care in the clinical sphere, with an increasing number of patients and clinicians turning to sequencing data as a determinant of treatment regimen. Knowledge of human protein coding genes and gene expression patterns is extensive, though not absolute. Venturing outside the relatively well-defined protein-coding regions of the genome, much is undetermined. Genome wide association studies (GWAS) have identified many genetic polymorphisms in non-coding regions on the genome that contribute to disease risk. Understanding the mechanisms by which a non-coding polymorphism can cause a phenotype demands an understanding of the physical organization and structure of chromatin in the eukaryotic nucleus.

Gene expression data from primary glioblastoma multiforme tumors (GBM) has uncovered the existence of four molecular subtypes, which affects prognosis and response to treatment. With the goal of gaining an understanding of transcriptional regulation in brain cancer, we profiled post-translational modifications of histone H3 in primary GBM tumors using ChIP-seq, and profiled gene expression in these tumors as well. We used a hidden Markov Model approach to abstract common co-occurrences of histone modifications into chromatin states. We were able to identify signatures consistent with known chromatin regulatory motifs, such as enhancers, and a bivalent state, marked by an active and repressive histone modification. These states regulated expression in a subtype-specific manner, with the proneural subtype showing a protective signature, and the mesenchymal and classical subtypes presenting a signature of invasive cellular migration and angiogenesis. The bivalent and enhancer states controlled a gene expression signature strongly suggestive of glioma stem cells (GSCs), the cells thought to be self-renewing in GBM.

As part of profiling gene expression in primary GBMs, we performed RNA-sequencing in primary normal human astrocytes and six GBM-derived commercially available cell lines. We identified widespread differences in expression between tumors and cell lines, as well as a gene interaction network that is common to tumors and cell lines, dominated by chromatin remodelers and Rho guanine exchange factors.

Finally, in a pilot study of 400 patients with atrial fibrillation (AF), we identified several SNPs associated with probability of success of cardiac ablation, a surgical therapy for AF. We propose that examining the local topology between a SNP of interest and any long-range contacts will help identify regulatory regions that allow a non-coding SNP to have an effect on gene expression, and thus phenotype.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Since the first complete sequence of the human genome in 2003, sequence variation (and its relationship to phenotype) has been the focus of many genotyping and whole genome sequencing studies [1]. Indeed, sequence variation controls a large proportion of genetic events that can result in phenotypic changes. While the relationship between coding regions and phenotype is generally well understood, the mechanism of how genetic variation in noncoding regions of the genome effects changes in phenotype remains obtuse [2]. Given that the adult human body contains over 200 distinct cell types, all derived from a single genome, at some level the mechanisms controlling cell fate and differentiation must be epigenetic in nature [3]. The effects of genetic polymorphisms in noncoding regions require genomic context to elucidate the mechanisms at play. This same genomic context can support efforts to identify the epigenetic mechanisms controlling cell fate and differentiation.

Projects such as ENCODE have amassed a large amount of data defining genome-wide profiles of DNA binding proteins, such as transcription factors, chromatin modifiers, and histone post-translational modifications. The initial data was derived from immortalized cell lines, and this data allowed for identification of patterns in DNA-binding proteins across multiple cell types [4]. Cell lines, both immortalized and primary, allow for identification of genetic pathways and epigenetic processes governing development and cell differentiation, including cancer development. However, cell lines are imperfect models for primary tissue [5, 6]. Thus there is a necessity to understand the epigenetic contribution to transcriptional regulation in uncultured tissue. What remains to be elucidated is an intimate understanding of how epigenetic signals, such as methylation or histone modifications, control cell fate in the adult as well as the developing embryo.

THE ORGANIZATION OF CHROMATIN WITHIN THE CELL NUCLEUS

Chromosomal domains; active and repressed compartments

In any given cell type, which genes are silent or expressed is determined in part by structure in the nuclear compartment. Chromosomal painting has revealed that the DNA from individual chromosomes tends to cluster together; each chromosome can also be divided into active and inactive sections [7]. Inactive regions of the genome were first identified by the presence of dark regions near the inner nuclear membrane. These highly compacted regions are termed heterochromatin, and they are transcriptionally silent [8]. Undifferentiated cells, such as embryonic stem cells, contain minimal inactive regions of this type, with the majority of their chromatin being "open" and accessible to the transcriptional machinery [9].

During the process of cellular differentiation, genes are successively silenced and heterochromatin appears at the nuclear periphery; this is also true for induced pluripotency [10]. These processes result in two compartments in differentiated cells: an "A" compartment, which contains active and open chromatin, and a "B" compartment, which contains highly compacted, silent chromatin, and localizes near the nuclear periphery (**Figure 1.1**). The B compartment is specifically marked by methylation of the K9 residue on histone H3 [11], or by methylation of the K27 residue on histone H3 [12]. These repressive marks promote compaction and specifically restrict access of the transcriptional machinery to the chromatin. The A compartment, in contrast, is specifically marked by trimethylation of the K4 residue of histone H3, as well as several other "activating" histone modifications, discussed in detail below.

Looping and topologically-associated domains (TADs)

Beneath the macro scale localization of chromosomes and compartments in the nucleus, there are topologically associated domains, or TADs. These range from 100 kilobases (kb) to several megabases (Mb) in size, and represent regions of chromatin that tend to interact with one another. TADs are important in timing the replication of large genomes [13], and in determining the functionality of enhancers, regulatory regions that drive the expression of distal genes. Potential mechanisms of action for noncoding somatic polymorphisms can be reduced in scope by considering what genes, SNPs, and regulatory regions are present on the same TAD. Regions that are linearly far away may be topologically close if they're located within the same TAD [14]. TADs can localize to the A or B compartments discussed above, and dimers of the multifunctional insulator binding protein *CTCF* define their boundaries [15]. In this context, *CTCF* defines large loops of active or repressed chromatin, with the elements contained on a given loop being more likely to interact, and be in the same compartment.

The 10nm fiber: "beads on a string"

When nuclear cell extracts are treated with a high salt solution and examined under high magnification, there is an appearance of circular beads on a string of DNA. The beads are nucleosomes – a nucleosome is a complex of eight histone proteins, and

147bp of DNA is wound around each complex, or octamer [16]. Since DNA is net negatively charged, it binds easily to the nucleosome octamers, which are net positively charged. Histone proteins have long terminal tails, which protrude from the nucleosome core and can be chemically modified at certain residues [17]. These chemical modifications are key to understanding the compartmentalization and function of chromatin in detail [18].

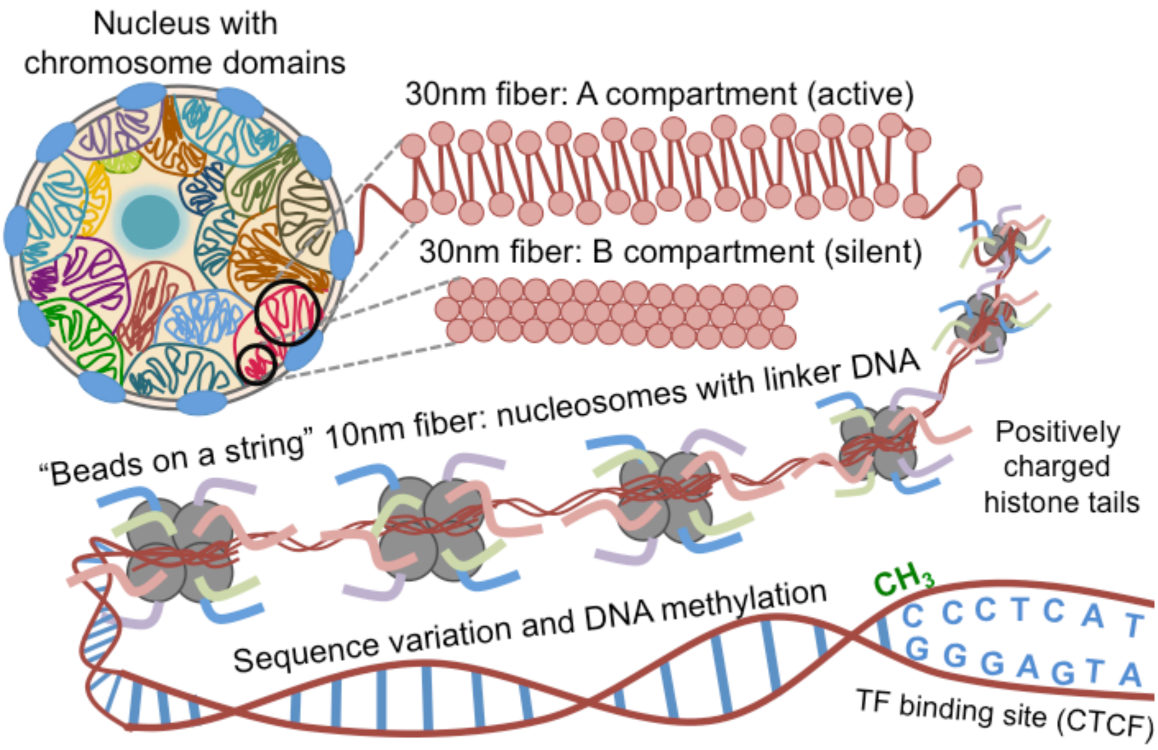

Nucleus with chromosome domains
30nm fiber: A compartment (active)
30nm fiber: B compartment (silent)
"Beads on a string" 10nm fiber: nucleosomes with linker DNA
Positively charged histone tails
Sequence variation and DNA methylation
CH3
C C C T C A T
G G G A G T A
TF binding site (CTCF)

Figure 1.1: Hierarchical regulation of the eukaryotic genome

Cartoon model of hierarchical levels of chromatin regulation. In the eukaryotic cell nucleus, chromosomes are organized into domains. Within these domains, there are "A" and "B" compartments. The lower inset into the nucleus indicates the tightly compacted organization of nucleosomes in the "B" compartments. The upper inset into the nucleus illustrates the organization of nucleosomes in the "A" compartment. The A compartment is then unwound into the 10nm fiber of nucleosomes connected by linker DNA, and finally into naked DNA, with one methylated cytosine indicated in green, as well as a TF binding site (for the multifunctional regulatory protein *CTCF*).

The nucleosome as a functional unit of transcriptional regulation

The histone code; histone tails are enriched for PTMs

Each nucleosome octamer contains four histone peptides (H3, H2A, H2B and H4) present in two copies each in a nucleosome octamers. Each peptide can be extensively chemically modified (**Figure 1.2**). As in other proteins, lysines, arginines, serines and threonines are commonly post-translationally modified. Because they are the most numerous, and also the most studied, the effect of histone H3 modifications on transcription and regulation of the genome is best understood at present [19]. These modifications can generally be grouped into activating and repressive modifications, which are specifically read and written onto the histone tails by a specific broad class of enzymes termed "chromatin remodelers." It should be noted that there are non-canonical histone variants, which can switch into nucleosomes in particular cellular contexts, such as *H2A.X* in DNA double strand break repair [20].

Active chromatin, histone modifications and chromatin modifiers

There are several modifications to histone H3, which result in a generally active state where the DNA underlying the chromatin can be transcribed. Tri-methylation of the K4 residue is mediated by the Trithorax complex, which was first discovered in *Drosophila*, and acts in opposition to the Polycomb repressor complex, discussed in the next section. In humans, there are six SET-domain methytransferases that can establish H3K4 trimethylation [21, 22]. The K4 residue can also be mono or di-methylated – monomethylation is specific for enhancer regions when combined with K27 acetylation [23]. Dimethylation specifically marks nucleosomes proximal to transcription factor binding loci [24], and tends to mark cell-type specific genes [25]. Mono- and di-methylation is catalyzed by the methyltransferase *KMT2D* that appears to be specific in this function [26].

Acetylation of histone 3 lysine residues 9 and 27 produces open chromatin. Acetylation of the positively charged lysine residue reduces some of the affinity of the local DNA for based on charge and produces a more open and accessible conformation of the chromatin [27]. K27 acetylation marks promoters and active enhancers [28], and is also a predictor of developmental state [29], while K9 acetylation broadly reflects promoter regions [30], and selectively marks regulatory elements such as active enhancers [31].

Acetylation is mediated by histone acetylases (HATs), which are frequently members of large multiprotein complexes that mediate effects on transcription, such as the SAGA complex [32]. Histone deacetylases, or HDACs, reverse lysine acetylation and are generally associated with repressive activity [33]. HATs and HDACs can have different catalytic specificity depending on their co-interacting proteins, so their specificity *in vitro* may not match their specificity *in vivo*.

Repressed chromatin, histone modifications and lysine demethylases

Silenced chromatin tends to accumulate near the nuclear periphery and is more compact than active chromatin. Trimethylation of the K9 or K27 residues of histone H3 is indicative of silenced chromatin, and this silencing is governed by two distinct mechanisms: the Polycomb repressor complex 2 (PRC2) mediates K27 trimethylation,

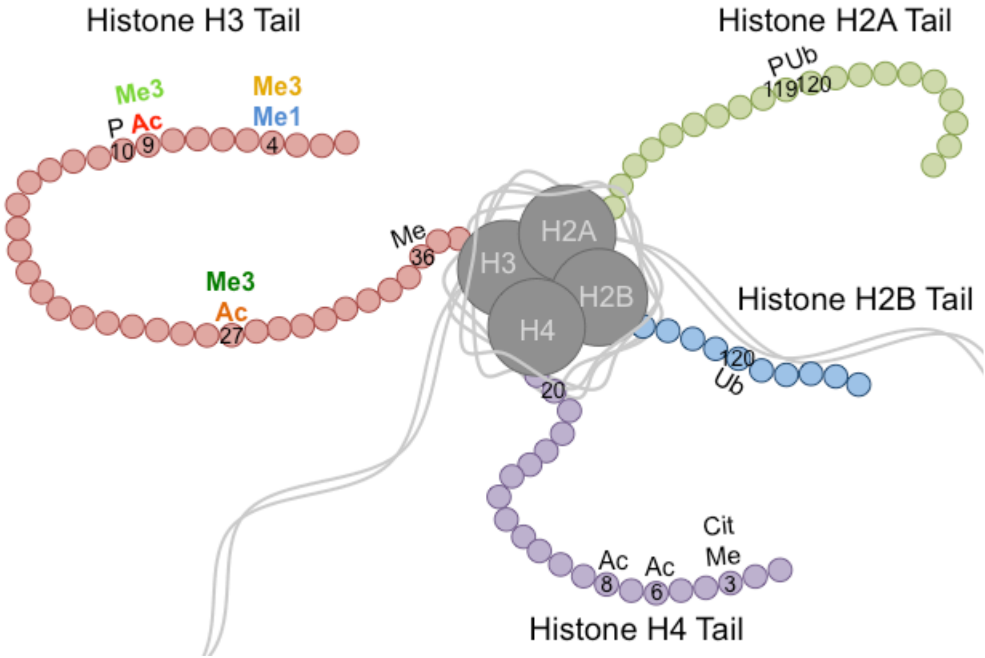


Figure 1.2: The modified nucleosome: common histone PTMs

Each of the eight histone peptides in a nucleosome octamer can be chemically postranslationally modified (PTM) at multiple positions. The six histone modifications illustrated in Chapter 2 are shown on the histone H3 tail in colored text. Common PTMs of other histone tails are displayed in black. P; phosphorylation, Ub; ubiquitylation, Cit; citrullination, Ac; acetylation, Me; methylation, Me1; mono-methylation, Me3; tri-methylation.

and *SUV39H1* catalyzes K9 trimethylation [34]. The PRC2 complex selectively silences genes temporally during development [12], and interacts with Polycomb repressor complex 1 to prevent DNA methylation of cytosines at these sites. As DNA methylation is generally silencing in nature [35], this is an indicator that polycomb silenced promoters are not permanently silent [36, 37]. Heterochromatin is also established during development, and is more strongly associated with the nuclear periphery than polycomb silencing [8]. The H3K9 methyltransferase *KMT1C* works in concert with a H3K4 demethylase (*KDM5A*) to maintain gene repression, so silencing of genes is not a passive process [38]. When polycomb or heterochromatin silencing is aberrantly removed from adult cells, or from tissue specific stem cells, cancer can be the result [39, 40].

GENOMIC APPROACHES TO STUDYING CHROMATIN STRUCTURE

Chromatin-immunoprecipitation followed by sequencing (ChIP-seq)

Given the massive number of proteins that bind to and modify DNA in some fashion, understanding their genome-wide binding patterns is important to understanding resulting gene regulatory effects. ChIP-seq allows for identification of the binding profiles of chemically modified histones, transcription factors, and components of the transcriptional machinery, such as RNA polymerase II (**Figure 1.3**). Coupled with RNA sequencing, ChIP-seq is a powerful technique for studying how DNA binding proteins regulate transcription of the genome [41].

However, it is a complex technique highly dependent on antibody specificity, and specific post-processing quality checking to determine where true binding signal exists [42]. When experiments are performed in primary tissues, chromatin degradation can be a serious issue that results in a low signal to noise ratio. As such, data derived from ChIP-seq experiments should be quality checked according to the parameters in the Methods section of Chapter 2, and any novel antibodies used for this technique should be validated using a successive Western blot and IP-Western to ensure the immunoprecipitation process is enriching for the protein or proteins of interest.

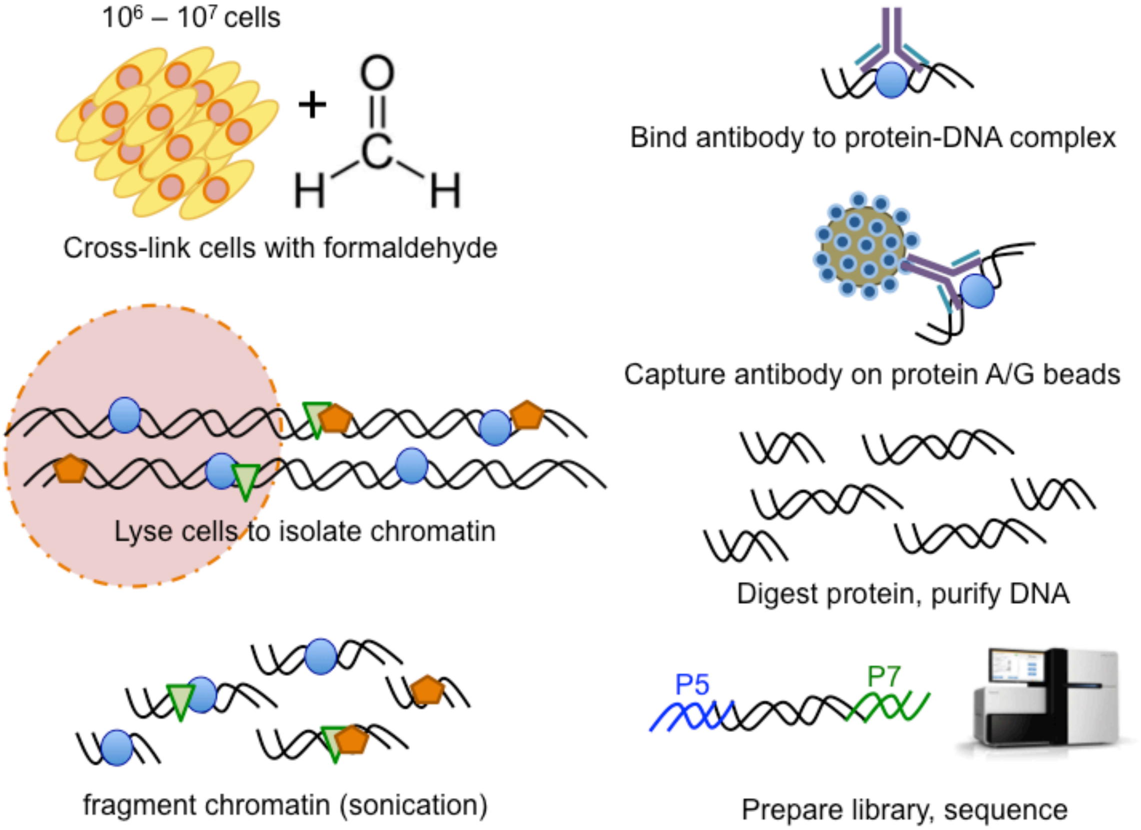

$10^6 - 10^7$ cells
+
O
C
H
H
Cross-link cells with formaldehyde
Bind antibody to protein-DNA complex
Capture antibody on protein A/G beads
Lyse cells to isolate chromatin
Digest protein, purify DNA
fragment chromatin (sonication)
P5
P7
Prepare library, sequence

Figure 1.3: Chromatin immunoprecipitation overview

Chromatin immunoprecipitation uses formaldehyde to induce covalent cross-linkages between DNA and any bound proteins. The cytoplasm (and any acellular components) is removed, and the nuclei are lysed using a gentle detergent and sonication. From this mixture of fragmented chromatin, a specific DNA-protein complex is pulled down using an antibody specific to that protein or chemical modification. The protein:DNA:antibody complexes are recovered using agarose beads coated with protein A (a bacterial protein that binds the constant region of an antibody). The bound DNA is purified by reversing the formaldehyde cross-linkage, then digesting the protein away using proteinase K. The purified DNA is then sequenced after library preparation. See the Methods section in Chapter 2 for a more detailed protocol.

Chromatin Conformation Capture and "C" based methods

To gain an understanding of the topology of chromatin in the nucleus, Chromatin Conformation Capture (3C, [43]) and related “C” techniques identify interactions between two loci that are topologically close (but may be linearly distant). The original

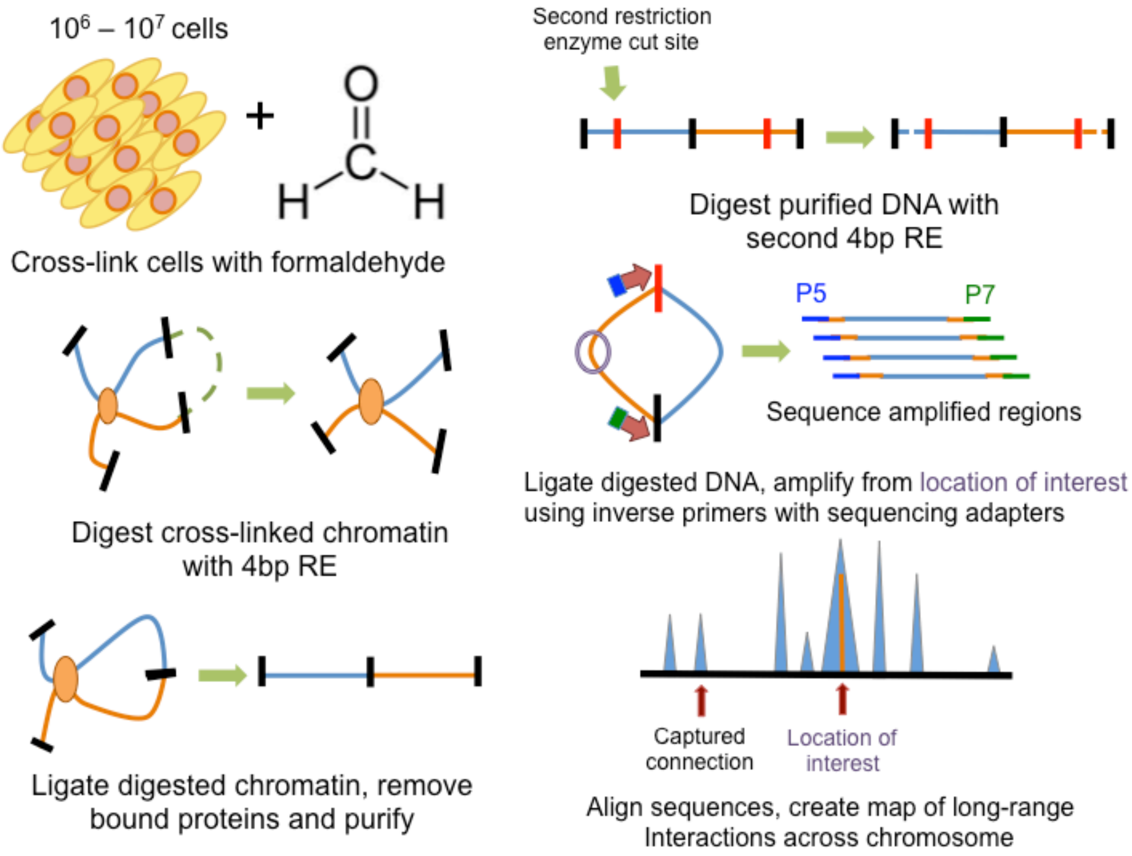

$10^6 - 10^7$ cells
Cross-link cells with formaldehyde
Digest cross-linked chromatin with 4bp RE
Ligate digested chromatin, remove bound proteins and purify
Second restriction enzyme cut site
Digest purified DNA with second 4bp RE
P5
P7
Sequence amplified regions
Ligate digested DNA, amplify from location of interest using inverse primers with sequencing adapters
Captured connection
Location of interest
Align sequences, create map of long-range Interactions across chromosome

Figure 1.4: Circular Chromatin Conformation Capture & sequencing (4C-seq)

4C-seq identifies long-range interactions between a single locus of interest and all other points on the same chromosome. After cross-linking with formaldehyde, the nucleus is lysed, and the chromatin is gently digested using sequential restriction digestion with a 4 bp restriction site, followed by ligation to create circles of DNA that are topologically close and may interact. Most of the interactions will occur within a Mb of the locus of interest, but some will be much further away. The statistical burden to ensure interactions occurring over a very long distance are non-random is high, which is why interactions between two chromosomes are rarely validated.

technique (3C) could validate a single long-range interaction (e.g. between a distal enhancer and promoter pair), while Hi-C interrogates all chromatin interactions across the genome on an “all by all” scale [44]. The large size of many mammalian genomes means for Hi-C sequencing data to be fine resolution, it must be sequenced very deeply, an expensive prospect. More tractable is 4C-seq (**Figure 1.4**, [45]), which looks at all long-range interactions originating from a single locus. Since the number of possible interactions interrogated is much smaller than for Hi-C, sub-kilobase resolution of long-range contacts can be established inexpensively. This technique, coupled with ChIP-seq, can elucidate plausible mechanisms of action for genetic polymorphisms in noncoding regions of the genome by identifying long-range interactions and the different types of regulatory complexes or proteins that can be brought into proximity by those interactions. Genomic context is important; the chromatin interactions that may dictate an effect in one tissue will not necessarily be present in all tissues in the body, and experiments must be designed carefully to compensate for these confounds.

Chapter 2: Bivalent chromatin domains in glioblastoma reveal a subtype-specific signature of glioma stem cells

Glioblastoma multiforme (GBM) can be clustered by gene expression into four main subtypes associated with prognosis and survival, but enhancers and other gene regulatory elements have not yet been identified in primary tumors. Here, we profiled six histone modifications and *CTCF* binding as well as gene expression in primary gliomas and identified chromatin states that define distinct regulatory elements across the tumor genome. Enhancers in the mesenchymal and classical tumor subtypes drive gene expression associated with cell migration and invasion, while enhancers in proneural tumors control genes associated with long-term survival in GBM. We identified for the first time in GBM, bivalent domains marked by activating and repressive chromatin modifications. Interestingly, the gene interaction network from common (subtype-independent) bivalent domains was highly enriched for homeobox genes and transcription factors, and dominated by the *SHH* and Wnt signaling pathways. This subtype-independent signature of early neural development may be indicative of poised de-differentiation capacity in glioblastoma, and could provide potential targets for therapy.

INTRODUCTION

Glioblastoma multiforme (GBM) is an aggressive primary brain tumor that accounts for 52% of all malignant primary brain neoplasias. The median time of survival with treatment is 14.6 months; only 5% of diagnosed individuals will survive five years from diagnosis [46, 47]. Given the dismal prognosis of GBM, many studies have focused on analysis of whole-genome/exome sequencing and gene expression data from primary GBM tumors to identify common gene mutations and expression profiles. These studies identified 4 molecular subtypes of GBM – classical, mesenchymal, neural, and proneural. These data have been invaluable in identifying genes and gene pathways that drive the development of GBM, and the identified subtypes predict some aspects of patient prognosis and response to treatment [48]. However, the underlying chromatin context that regulates gene expression programs in primary GBM tumors is largely unknown. Given that GBM lesions are developmentally plastic, and can change certain aspects of their cellular identity, understanding how they vary with regard to their chromatin structure will enable identification of key genes and regulatory motifs controlling differentiation capacity in GBM.

While several studies have quantified single histone modifications in GBM-derived cell lines, none of these studies have been performed in uncultured primary tissue, and few have looked at patterns derived from multiple histone modifications in the same cell line or tumor [49-53]. These studies established the general trend that repressive modifications (particularly polycomb silencing) are globally reduced in GBM, and active modifications (such as H3K4me3, H3K9ac, H3K27ac) are generally increased across the genome in GBM. This lack of data hinders efforts to conclusively identify and characterize the cell types that give rise to glioblastoma tumors. Indeed, a recent chromatin profiling study to identify enhancers revealed cell-type of origin in medulloblastoma, but no comparable dataset currently exists for GBM [54].

In this study, we sought to categorize regulatory regions of the genome in primary GBM tumors by profiling six post-translational modifications of histone H3, and binding of the multifunctional insulator binding protein *CTCF*, in conjunction with gene expression profiling of the same tumors. We used a HMM-based approach [55] and identified combinations of chromatin marks that defined distinct regulatory elements across the genome (**Figure 2.1a**). The resulting model encompassed 21 chromatin states that identified known regulatory elements such as enhancers and promoters, and identified bivalent regions in tumors for the first time. We were able to annotate any state in this model with matched expression data, generating a context-dependent view of gene expression that also identified regulatory regions that may control gene expression indirectly.

We were able to obtain nine glioblastoma multiforme tumors and two anaplastic astrocytoma tumors for this study. While the sample number was smaller than desired, each experiment performed (seven IPs per tumor, plus RNA sequencing) integrates data from millions of cells derived from the same homogenized tumor material. Thus, while the data are not as clean as cell lines, or single-cell data, each experiment interrogates the genome of millions of cells derived from the same tumor.

RESULTS

Gene expression in tumors recapitulates clinically distinct GBM subtypes

GBM tumors are highly heterogeneous [56], and tumors were homogenized before processing, so the data are representative of bulk tumor, as opposed to any specific population of cells. To ensure that the tumors we used for chromatin profiling

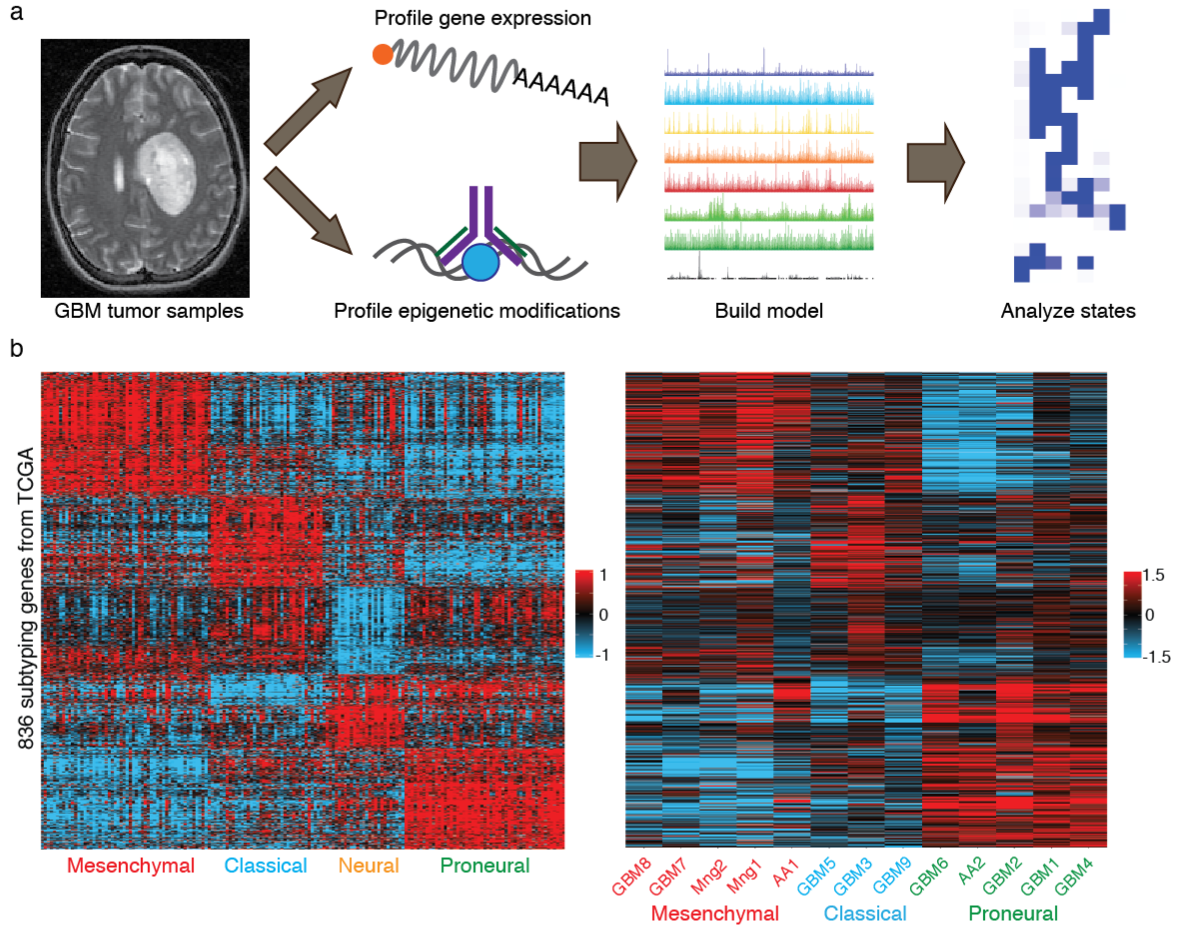


Figure 2.1: Bulk tumors represent the known molecular subtypes in GBM.

(a) Overview of our approach. We profiled histone modifications and used ChromHMM [55] to produce a model of chromatin states, and associated distinct chromatin states with gene expression profiles from the same tumors.
(b) The panel on the left displays microarray data used by TCGA to establish molecular subtypes in GBM [58]. Data were hierarchically clustered on both axes using Spearman's rho. The right hand panel displays RNA sequencing data generated in this study. The gene order is the same as on the left, but tumors were hierarchically clustered as above.

represented clinically valid GBM tumors, we analyzed gene expression profiles from the same tumors that were used for ChIP-seq. We performed RNA-seq from all primary tumors, as well as several commonly used GBM-derived cell lines and two independent lines of primary normal human astrocytes. After alignment, gene expression was quantified over protein coding genes using the GENCODE annotation for hg38 [57] (Methods). Using the 836 subtype classifier genes identified by TCGA [58], these data were plotted as a heatmap.

We found the same basic groups and expression patterns as in the TCGA tumors despite having many fewer tumors, and using RNA-seq to profile gene expression as opposed to the microarray analysis used previously by TCGA (**Figure 2.1b**). Thus, the tumor tumors that we used for chromatin profiling represent authentic and clinically relevant GBM subtypes. This comparison allows us to address heterogeneity in tumors as well. As we do not have access to pathology reports derived from the samples used in this study, we are unable to know the percentage of tumor (as opposed to stromal, or infiltrating) material in each sample. However, TCGA tumor samples were screened to have at least 80% tumor material present [58]. Given that our RNA-seq data cleanly replicates the TCGA subtype data, this indicates that the tumors analyzed in this study are similarly enriched for tumor material, and are not compared predominantly of stroma or infiltrating immune cells. While we could clearly detect the mesenchymal, classical and proneural subtypes, we did not detect any neural subtype tumors. The neural subtype is less common than the other 3 subtypes and there is some debate regarding whether the neural subtype is a distinct molecular subtype in GBM [48, 59, 60].

Two meningioma tumors clustered with the classical GBM subtype and were not used in subsequent analyses. GBM cell lines differed widely from the tumors in their gene expression patterns, and thus are not an accurate model of primary tumor lesions for genome-wide profiling studies. When we clustered all the tumors on the 836 subtyping genes identified by TCGA, three groups were evident: tumors, normal human astrocytes, and GBM-derived cell lines (**Figure 2.2**).

Global profiling of histone modification reveals biologically relevant patterns

To profile regulatory chromatin states in GBM, we developed a protocol to perform ChIP-seq in fresh-frozen primary GBM tumors (Methods) and concurrently sequenced the RNA derived from these tumors. We profiled four active histone modifications (H3K4me1, H3K4me3, H3K9ac, H3K27ac) and two repressive histone modifications (H3K27me3, H3K9me3) as well as the multifunctional insulator binding protein *CTCF*. Transcribed genes such as calmodulin (*CALM1*) displayed active histone modifications (**Figure 2.3a**), while transcriptionally silent genes such as keratin 72 (*KRT72*) showed broad repressive marks (**Figure 2.3b**). Genome-wide, active or repressive marks and *CTCF* binding clustered together, with tumors showing a given mark generally clustering together (**Figure 2.3c,d**). Thus, the biological state of the chromatin rather than tumor identity determined the clustering of datasets, indicating that our profiling data was reflective of the underlying chromatin state.

To systematically identify distinct chromatin states in the tumor genomes, we first called ChIP-seq peaks in each tumor, then used ChromHMM [55] to build a 21-state model of combinations of histone modifications across the genome (Methods). To focus on epigenetic states in GBM, we used only profiling data from GBM tumors to generate the model. Based on known associations of histone marks and *CTCF* binding with regulatory activities, we identified several functionally distinct chromatin states in the

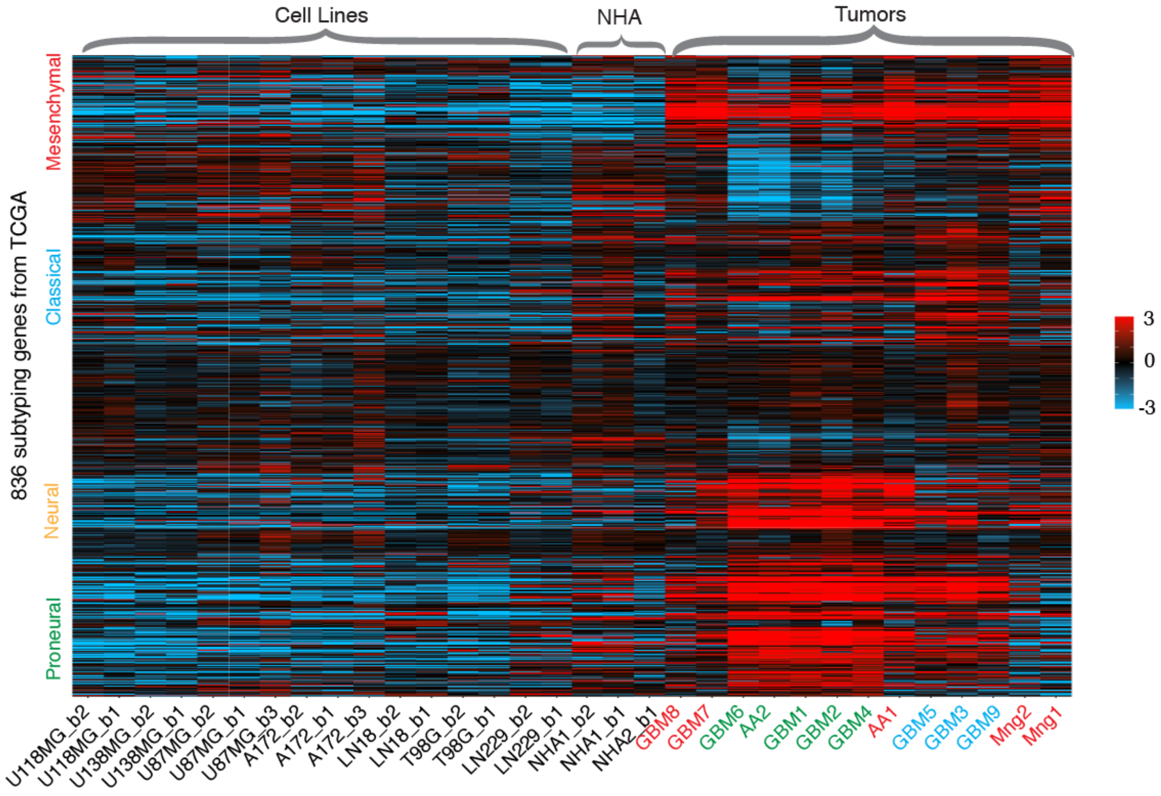


Figure 2.2: Expanded subtyping and clustering data, including GBM-derived cell lines.

Clustering of RNA-seq data over the 836 TCGA subtyping genes in **Figure 2.1** for all samples, both tumors and GBM-derived cell lines. Tumors were hierarchically clustered using Spearman's rho, with cuffnorm-derived FPKM values (Methods). Tumor color corresponds to subtype.

tumor genome [3, 28, 61]. There were several promoter and enhancer-like states, including an active enhancer state, polycomb and heterochromatin silenced states (**Table 2.1**). Interestingly, the 21-state model revealed the existence of a bivalent state marked by active H3K4me3 and repressive H3K27me3 modifications. Such bivalent states were first identified in embryonic stem cells (ESCs) [62], and have been identified in glioblastoma derived cell lines [63, 64] but to our knowledge, this is the first time they have been seen to exist in primary GBM tumors. A view of the ChIP-seq signal surrounding a given

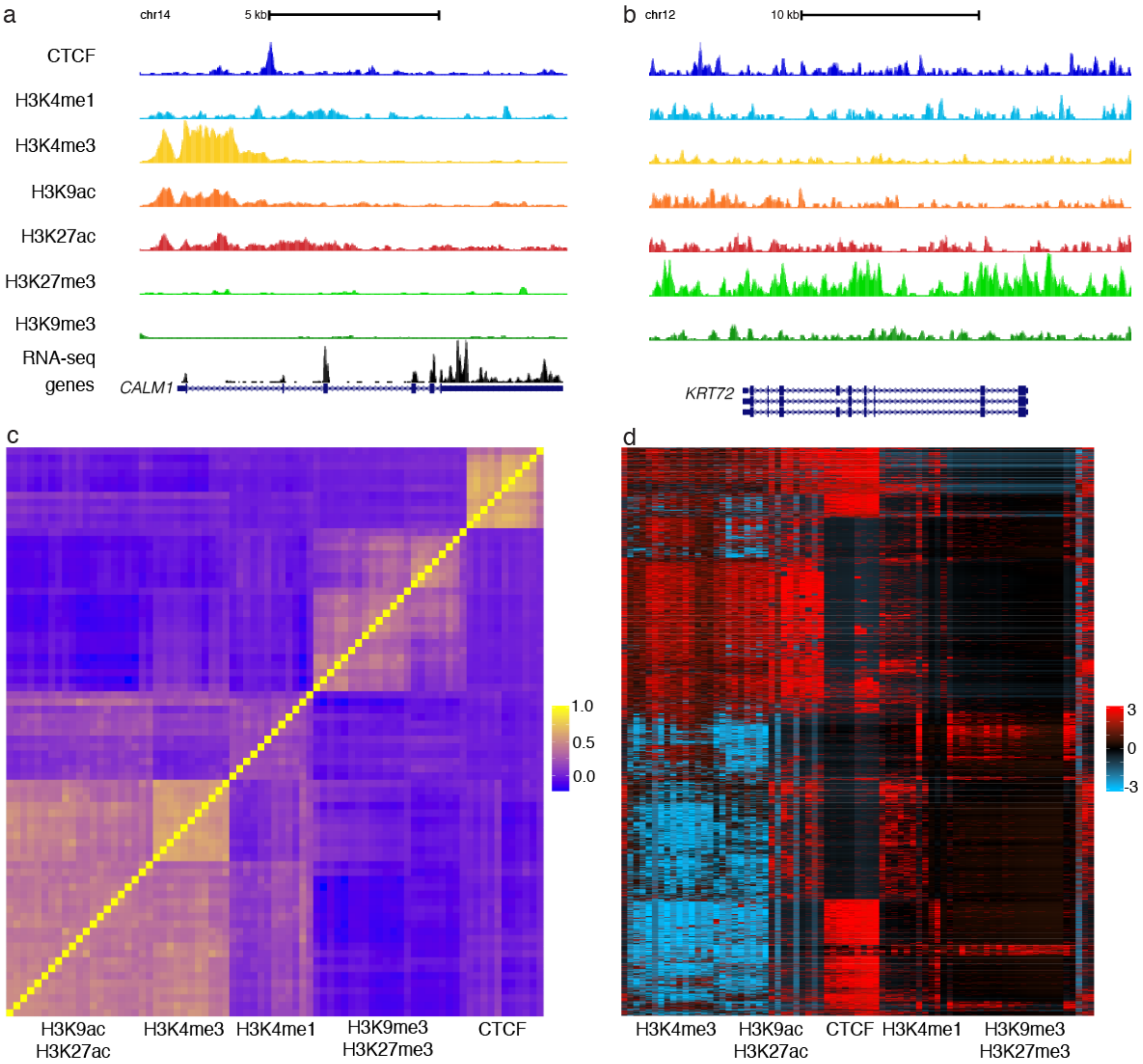


Figure 2.3: Epigenetic profiles in GBM tumors.

(a) Active chromatin over the promoter region for calmodulin (*CALM1*), which is highly expressed in the tumor GBM7. Region displayed: chr14:90,391,001-90,427,000 on the hg38 assembly. **(b)** A polycomb-repressed region, marked by H3K27me3, in the same tumor as **a** over the gene *KRT72*, which encodes a keratin protein. Region displayed: chr12:52,580,318-52,607,332 on the hg38 assembly. **(c)** A clustered correlation heatmap of all chromatin profiles generated in this study. Pairwise correlation coefficients across the 33,188 genomic loci with measurable signal in at least 15 experiments are shown, with clusters of chromatin marks indicated by the text. **(d)** Heatmap of normalized ChIP-seq signals in the same genomic loci shown in **c**. Both genomic loci (vertical) and experimental tumors (horizontal) were hierarchically clustered.

state	median cov	total cov	CTCF	K9ac	K27ac	K4me1	K4me3	K9me3	K27me3	class	
1	0.18%	3.33%	1.1%	0.1%	0.2%	0.8%	0.0%	99.4%	0.8%	heterochromatin	🟪
2	0.00%	0.43%	4.4%	2.0%	0.0%	11.6%	99.7%	99.2%	0.0%	repressed	🟪
3	0.54%	3.56%	2.6%	0.0%	0.0%	0.0%	99.2%	0.0%	0.0%	h3k4me3	🟨
4	0.36%	2.09%	0.0%	100.0%	0.0%	0.0%	100.0%	0.1%	0.0%	weak promoter	🟧
5	0.05%	0.44%	7.1%	99.3%	0.0%	99.1%	100.0%	0.1%	0.1%	weak promoter	🟧
6	0.09%	0.80%	8.7%	99.5%	99.9%	99.3%	99.9%	0.2%	0.7%	strong promoter	🟪
7	0.54%	1.80%	0.0%	99.7%	99.9%	0.0%	100.0%	0.2%	0.0%	promoter	🟥
8	0.14%	1.31%	4.0%	99.3%	99.0%	0.0%	0.0%	0.1%	0.0%	weak enhancer	🟦
9	0.07%	0.66%	4.0%	99.4%	99.2%	99.3%	0.0%	0.1%	0.1%	weak enhancer	🟦
10	0.06%	0.58%	4.3%	99.0%	0.0%	99.1%	0.0%	0.1%	0.0%	weak enhancer	🟦
11	0.26%	2.90%	1.4%	0.0%	0.0%	98.4%	0.0%	0.8%	0.1%	Inact. enhancer	⬛
12	0.07%	0.80%	1.7%	0.0%	99.0%	99.2%	0.0%	8.2%	0.2%	enhancer	🟦
13	0.51%	5.17%	1.2%	0.0%	97.9%	0.0%	0.0%	2.1%	0.0%	h3k27ac	🟦
14	0.06%	1.41%	3.3%	0.0%	99.4%	7.1%	99.7%	31.3%	0.0%	weak promoter	🟧
15	0.04%	0.56%	4.0%	0.0%	30.3%	99.1%	99.6%	0.1%	0.0%	weak enhancer	🟦
16	0.09%	0.85%	5.8%	37.3%	19.0%	10.7%	84.6%	25.9%	99.4%	bivalent	🟪
17	0.36%	4.76%	0.7%	0.0%	0.0%	0.2%	0.0%	4.1%	98.1%	polycomb	🟪
18	95.15%	98.97%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	no signal	⬜
19	0.47%	4.17%	3.9%	98.1%	0.0%	0.0%	0.0%	0.0%	0.0%	h3k9ac	🟩
20	0.08%	0.42%	100.0%	97.0%	63.3%	0.0%	99.3%	0.3%	0.2%	wk prmt + ctcf	🟫
21	0.24%	1.31%	99.8%	0.5%	0.6%	0.3%	2.7%	0.1%	0.0%	ctcf binding	🟩

Table 2.1: A 21-state model of chromatin states in the GBM genome.

The above table represents the output of ChromHMM for a model with 21 states. The "median cov" indicates the median percentage of the hg38 genome covered by that state for any tumor, while "total cov" indicates the maximum coverage of that state over all tumors. Columns 4-10 indicate the proportion of a given state marked by each chromatin mark.

chromatin state in a tumor showed that globally, the identified states faithfully reflected the underlying combinations of histone marks (**Figure 2.4**). At individual loci, the identified states captured the appropriate combination of marks corresponding to different types of functional elements such as promoters and enhancers, or silenced heterochromatin regions (**Figure 2.5**). Although expression across the states was variable, any state with a repressive mark was associated with a statistically significant reduction in expression compared with other states (**Figure 2.6a**).

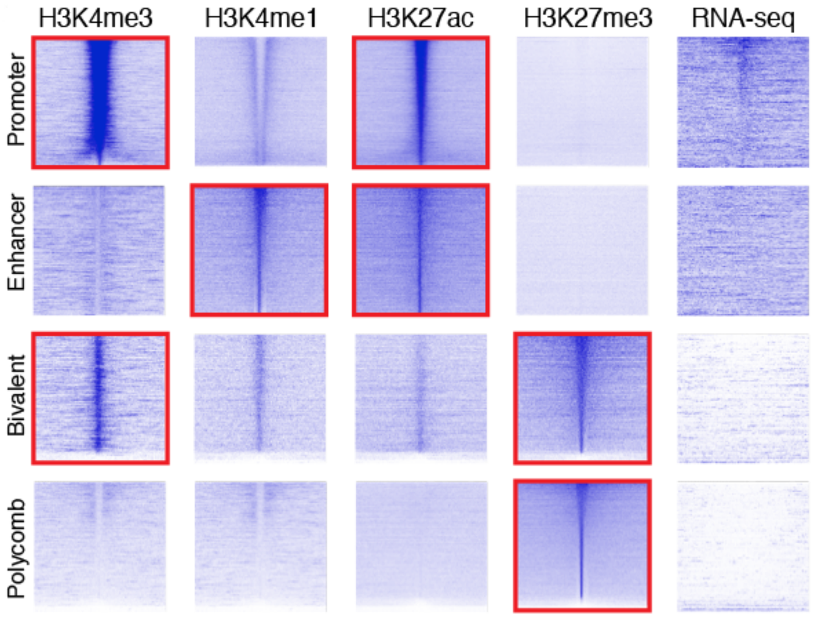


Figure 2.4: Epigenetic signal surrounding four chromatin states in GBM4.

The above image represents the sequencing signal (aligned reads) from four histone modifications as well as RNA sequencing surrounding four chromatin states. Each square is 20kb wide – 10kb upstream and 10kb downstream of the center of the chromatin state locus. Several thousand individual regions were surveyed for each chromatin state, and the y-axis sort order is the same across each row. The promoter state is sorted by the H3K4me3 signal, the enhancer state is sorted by the H3K4me1 signal, and the bivalent and polycomb states are sorted by the H3K27me3 signal. The aggregate number of reads across the center 2kb determines the sort order for each state.

We used whole genome bisulfite sequencing data from TCGA GBM tumors to examine DNA methylation levels corresponding to each state. The polycomb and heterochromatin silenced states were highly methylated, reflecting their transcriptional inactivity. Interestingly, although the genes nearest enhancers were highly expressed, the enhancers themselves were highly methylated (**Figure 2.6b**). While methylation is generally considered to be silencing [35], WGBS is unable to distinguish between methylation and

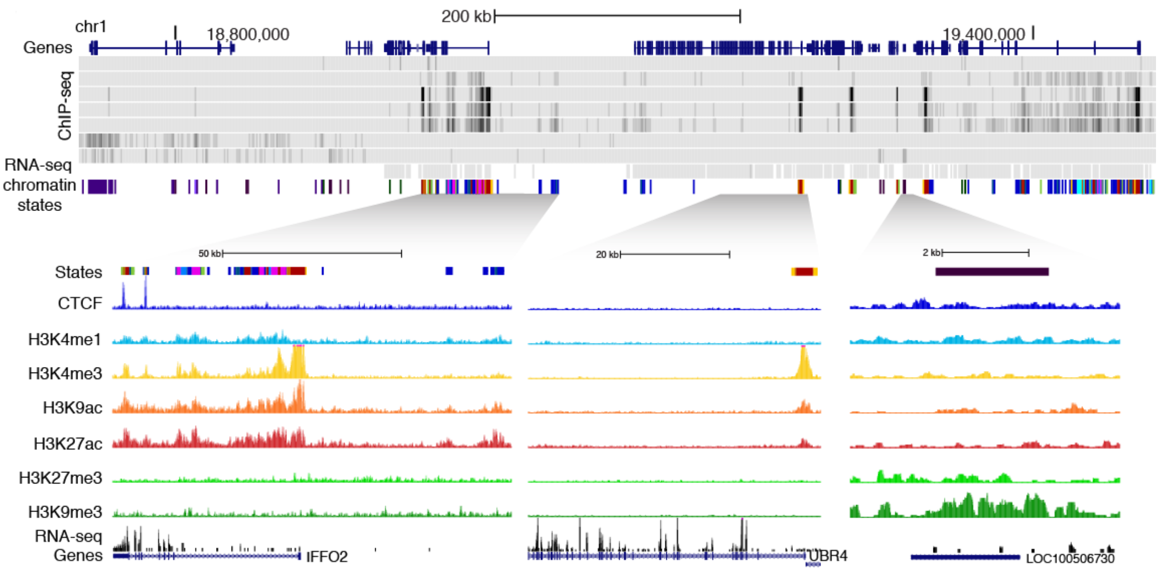


Figure 2.5: A view of 877 kb on chromosome 1 for the tumor GBM8

The figure above illustrates 7 ChIP-seq tracks, RNAseq, chromatin states and gene annotations. Below, a close-up view of chromatin states showing, from left, a weak enhancer, a promoter, and a heterochromatin-silenced region over a lncRNA of unknown function. Region displayed on the hg38 assembly: chr1:18,621,857-19,499,690.

5-hydroxymethylation (5hmC) of cytosine residues. Enhancers and gene bodies are specifically marked by 5hmC in ES cells [65, 66] and in GBM [67]. Using regions of high 5hmC from Johnson et al., [67] our enhancers are enriched for 5hmC compared to promoters (*t*-test; *P* = 6.212e-07), bivalent regions (*P* = 4.08e-09), polycomb silenced regions (*P* =1.866e-09), and are enriched compared to background levels in the genome (*P* = 2.778e-12) (**Figure 2.7a**). Genes corresponding to bivalent states were not expressed, and bivalent loci showed lower levels of methylation than polycomb and heterochromatin silenced regions. *CTCF* binding sites were associated with high levels of methylation, consistent with previous reports [68].

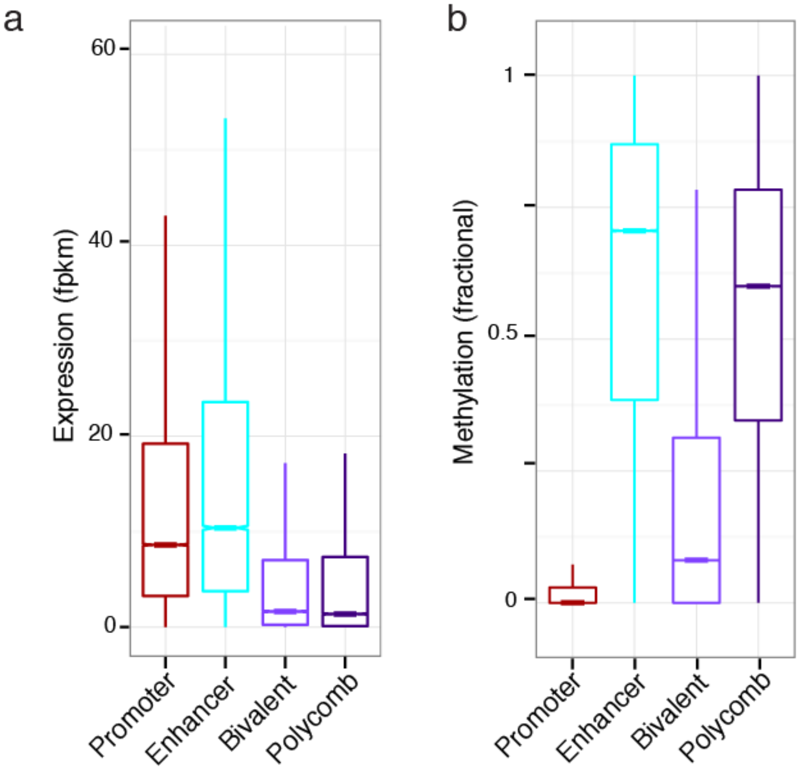


Figure 2.6: Expression and methylation across 4 chromatin states.

(a) Average expression of genes closest to each of the four states displayed in **Figure 2.4**. Normalized FPKM counts across all tumors were derived from cuffnorm (Methods).
(b) Average methylation across each of the states in **Figure 2.4**. Fractional methylation levels were derived from WGBS data in GBM and intersected with our data using bedtools intersect.

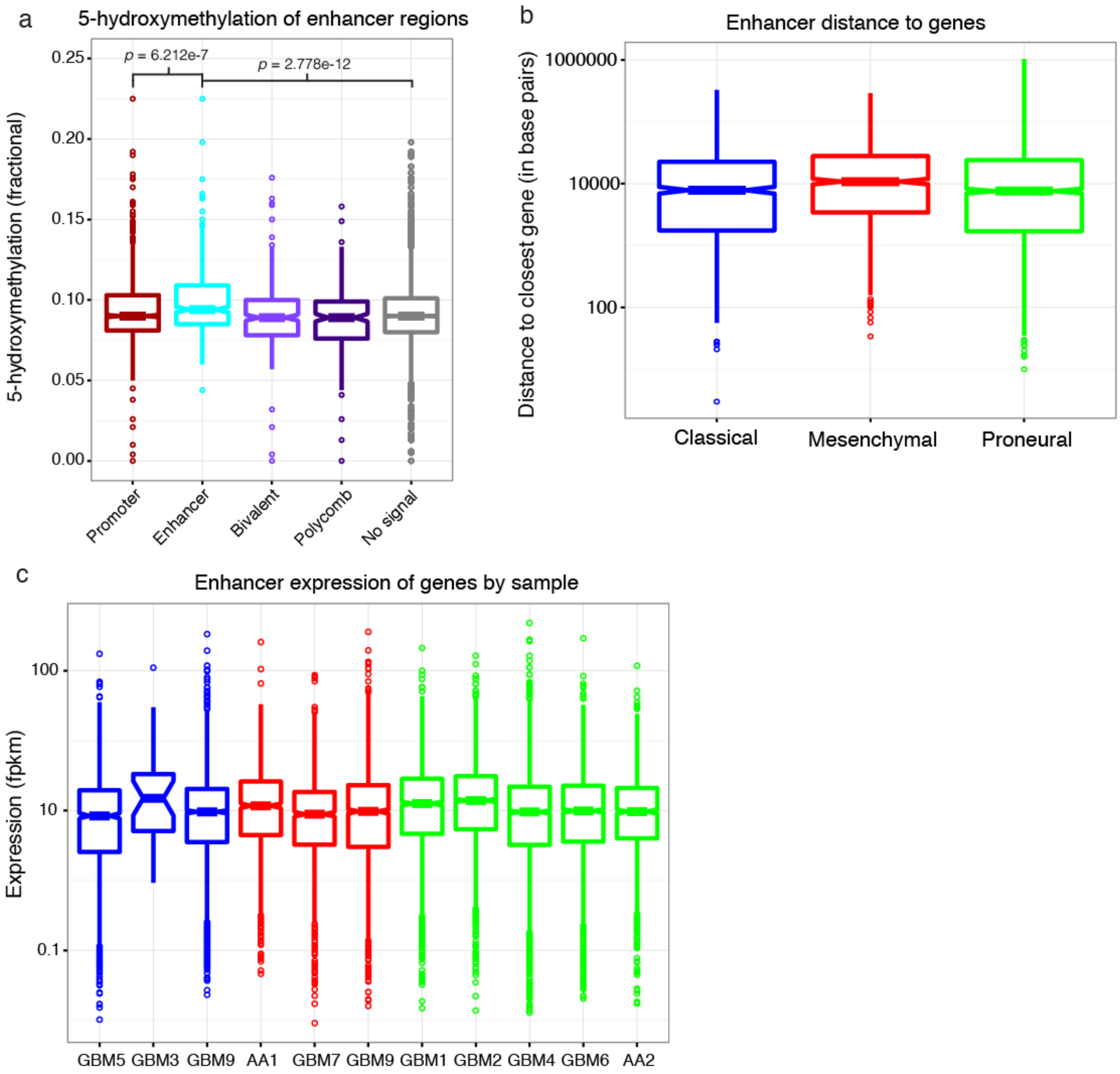


Figure 2.7: 5hmC signal over enhancers, enhancer distance to genes, average expression of enhancer-associated genes by tumor.

(a) 5-hydroxymethylation data from Johnson et al. [67] over five chromatin states from the model, promoter, enhancer, bivalent, polycomb and no signal (state 18, background).
(b) Boxplots by subtype displaying the data on distance from enhancers to the closest proximal gene; the distances displayed are only from enhancers that are not located within a gene body.
(c) Average expression of genes proximal to enhancers in each tumor. FPKM calculated as before, and tumors are colored by subtype.

GBM enhancers are located predominantly within introns and intergenic regions, and contain degenerate STAT/Klf/SP family motifs

	GBM1	GBM2	GBM3	GBM4	GBM5	GBM6	GBM7	GBM8	GBM9	AA1	AA2
Active enhancer (State 12)	3640	5658	33	8077	2191	5091	1783	4135	6401	1906	1908
Inactive enhancer (State 11)	11745	25828	1095	40921	9245	12332	17169	3260	14065	5951	9960
H3K27ac peaks	30979	62605	38262	44629	40024	49276	42779	60416	31341	34406	36431
H3K4me1 peaks	17437	45473	3324	59145	12890	22190	25253	9375	23910	13420	17183

Table 2.2: Enhancer counts for this dataset

Active enhancers varied widely in number among tumors, with a median number of 3,640 (**Table 2.2**). Though expression levels for genes associated with enhancers varied across tumors, there were no statistically significant changes in expression across tumors or subtypes or in the average distance to a gene (**Figure 2.7b,c**). Generally, enhancers localized within or upstream of gene bodies. The vast majority of enhancers in any given tumor were located in introns. The genomic distribution of enhancers that we identified in GBM tumors [54] and one from enhancers defined in cell lines by ENCODE [69] (**Figure 2.8a**).

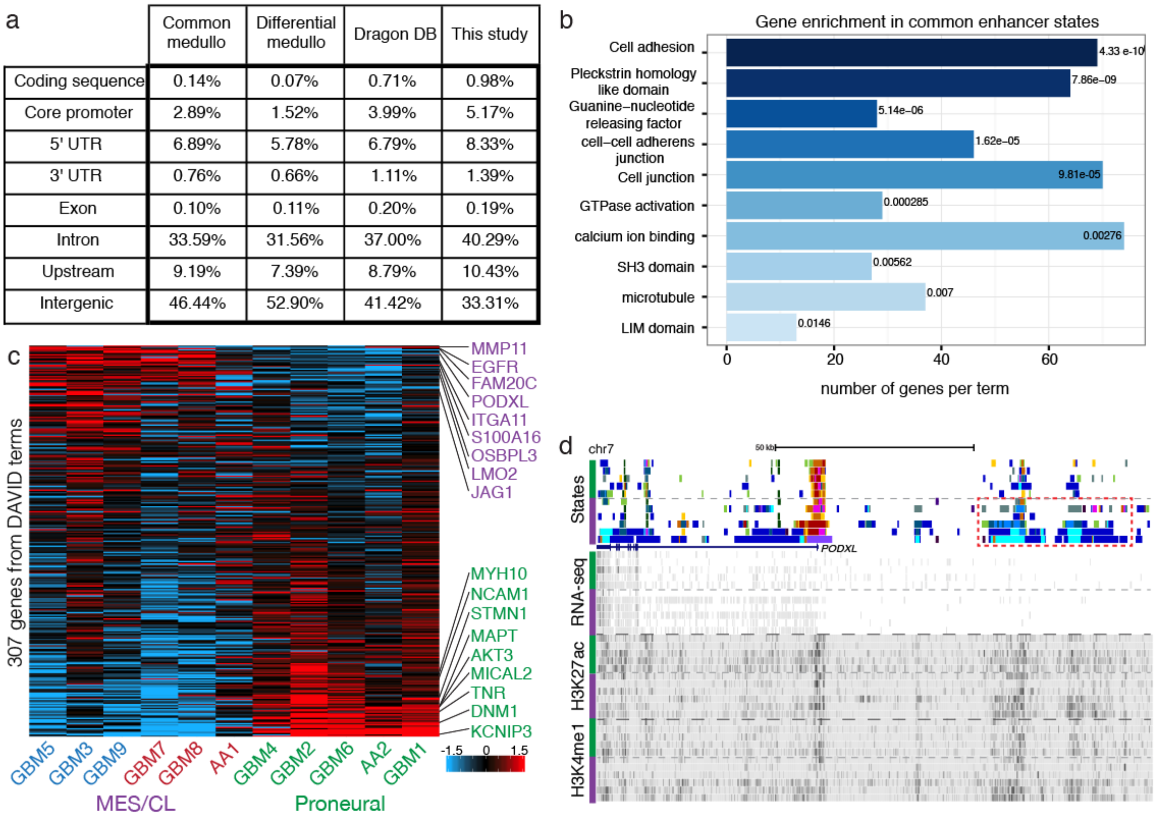
We used MEME-ChIP [70] to identify de novo motifs overrepresented in enhancers in each tumor. These motifs were largely degenerate, and bore resemblance to motifs for several families of transcription factors, such as a TTYCY short motif, with some similarity to KLF and STAT-like binding motifs (**Figure 2.9**). Other short motifs resembled portions of the *NFATC2* DNA-binding region, with representation by several

ETS-family transcription factors (*EHF*, *ETV2*). Several less degenerate 6-8 mers motifs strongly resembled primary and secondary binding sites for the *AP2* transcription factor family, as well as *TCF3*, *ASCL2* and *TCF5*. The observed motifs indicate that enhancers are enriched for binding of transcription factors controlling cellular proliferation and immune response.

Enhancers are subtype-specific and control genes involved in cell-cell contacts

The enhancer state in our model is defined by co-localization of H3K27ac and H3K4me1 (**Figure 2.4, Table 2.2**). This is a chromatin-based definition of enhancers, as opposed to the more classical definition of regions that can drive *lacZ* or *gal4* expression in an animal model. Properly, these regions are putative enhancers, as they are untested for driving expression of a reporter gene in an animal model, though analysis of matched expression data indicates genes affiliated with an enhancer do have higher expression than genes associated with a promoter alone (**Figure 2.6**).

Some transcriptional activity originated from these regions, but in a less defined manner than that from promoter regions (**Figure 2.4**). 1,817 enhancers, which covered 1,227 genes, were present in at least 3 out of 11 tumors and we defined this set as our "common enhancers". 307 of these genes were strongly enriched for pathways that mediate cell-cell interactions such as cell adhesion and cell-cell adherens junctions (**Figure 2.8b**). These enrichments were driven by many groups of genes, such as laminins (*LAMA5*, *LAMB3*), cadherins, (*CDH1*, *CDH4*), integrins (*ITGA11*, *ITGB5*), catenins (*CTNNA1*, *CTNND1*, *CTNND2*), cell adhesion molecules (*CADM1*, *HEPACAM*, *NCAM1*, *NRCAM*), myosins (*MYO1E*, *MYH9*, *MYH10*) and actinins (*ACTN1*, *ACTN4*).



	Common medullo	Differential medullo	Dragon DB	This study
Coding sequence	0.14%	0.07%	0.71%	0.98%
Core promoter	2.89%	1.52%	3.99%	5.17%
5' UTR	6.89%	5.78%	6.79%	8.33%
3' UTR	0.76%	0.66%	1.11%	1.39%
Exon	0.10%	0.11%	0.20%	0.19%
Intron	33.59%	31.56%	37.00%	40.29%
Upstream	9.19%	7.39%	8.79%	10.43%
Intergenic	46.44%	52.90%	41.42%	33.31%

Figure 2.8: GBM enhancers regulate gene expression in a cell-type specific manner.

(a) The distribution of enhancers identified in this study across defined elements in the genome (Methods), compared with two external enhancer datasets: Common and differentially regulated medulloblastoma enhancers [54] and Dragon DB enhancers defined by ENCODE in cell lines [69]. **(b)** Functional enrichment for genes associated with enhancers in at least 3 tumors. Benjamini adjusted *P*-values provided by DAVID [71] are shown to the side of each bar, with the shading proportional to the significance. **(c)** Heatmap of gene expression from the 307 enhancer associated genes with a functional annotation in **a**. **(d)** A 140 kb genome browser view of the region surrounding the gene PODXL. This enhancer is subtype-specific, and is much stronger in the MES/CL tumors (indicated by purple bars, with the enhancer region in MES/CL tumors outlined by a red dashed rectangle). The hg38 coordinates of the region shown are: chr7:131,500,691-131,640,269.

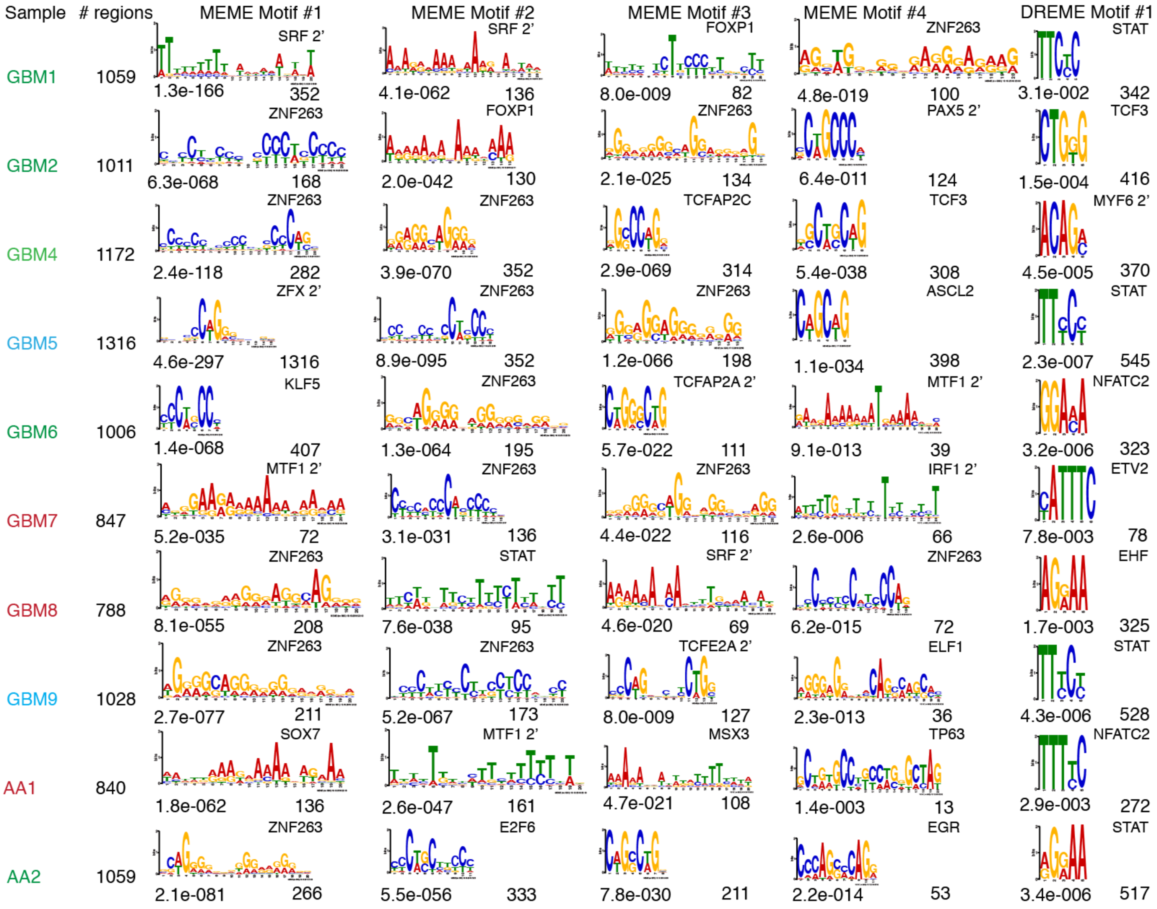


Figure 2.9: Sequence motifs in enhancers.

De-novo motifs identified by MEME-ChIP for each tumor. The first column indicates the tumor, followed by the number of enhancer regions that were analyzed. Within each motif entry, we list the P-value of the motif identified, then the number of occurrences of that motif, on the lower left and lower right of each panel, respectively. The upper right hand corner indicates the most similar motif identified by TOMTOM. The first four motifs identified by MEME are listed first, and the fifth column lists the most significant motif identified by DREME, a program focused on identifying shorter motifs de novo. DREME and TOMTOM are programs in the MEME suite [70], which was used for this analysis.

These 307 genes with a DAVID annotation from **Figure 2.8b** were split between proneural (PN) and mesenchymal/classical (MES/CL) tumors, with genes expressed in

PN tumors being minimally expressed in the MES/CL tumors and vice versa (**Figure 2.8c**). Of these 307 genes, 33 were on the TCGA list of subtype genes [48]. This includes five genes significantly upregulated in the proneural tumors (*EPHB1*, *MAPT*, *NCAM1*, *KIF21B*, *STMN1*), and two genes that are significantly up in the MES/CL tumors (*EGFR*, *OSBPL3*), in addition to 26 genes that are not statistically different across the two sets. In total, we identified 274 novel genes associated with GBM that are controlled by an enhancer in at least 3 tumors. The subtype specificity of enhancers was often visually evident, as with the gene *PODXL*, which showed strong enhancer signals and expression in the MES/CL tumors, but much weaker enhancer signals and expression in PN tumors (**Figure 2.8d**).

Bivalent regions are enriched for hedgehog signaling and developmental genes

	GBM1	GBM2	GBM3	GBM4	GBM5	GBM6	GBM7	GBM8	GBM9	AA1	AA2
Bivalent (State 16)	1719	1367	824	2638	1661	2202	4532	3615	10565	5514	1789
Polycomb (State 17)	4472	6672	19861	10841	7635	6791	43376	17833	43885	22159	4348
H3K4me3 peaks	54308	29580	38949	45845	23378	32594	29330	32866	41956	34173	33639
H3K27me3 peaks	4947	7975	20709	12068	8553	7538	48672	19630	48136	25331	5129

Table 2.3: Bivalent domain counts for this dataset

The model identified a bivalent state (state 16, **Table 2.3**) predominantly defined by co-localization of H3K4me3 and H3K27me3 (**Figure 2.4**). Transcription from this state was slightly higher than a polycomb silenced state, but lower than genes near active promoters or enhancers (**Figure 2.4**, **Figure 2.6**). There were 2,027 frequently bivalent regions that were present in at least 5 tumors, and these regions covered 1,510 distinct genes (Methods). 840 of these genes showed strong enrichment for pattern specification, regionalization, embryonic development and transcription factor activity (**Figure 2.10a**) and clustered with a PN versus MES/CL division in gene expression, similar to enhancers. Frequently bivalent regions were divided in expression by subtype and fell into 2 groups: Group 1 contained genes that were expressed in MES/CL tumors and were often bivalent in PN tumors, and Group 2 showed largely the reciprocal pattern (**Figure 2.10b**, **Figure 2.11**). For example, Group 1 contained many genes in the HOXB locus that were bivalent in PN tumors but expressed in MES/CL tumors (**Figure 2.10c**).

To identify regions that were bivalent in a subtype independent manner in GBM, we examined bivalent regions common to at least 8 out of 11 tumors, and identified 467 regions, which covered 381 unique genes. 68 of these 381 genes (17.8%) were homeobox genes, a highly significant enrichment given that only 1.25% of all genes are homeobox genes (*P* < 2.2 e-16, Fisher's Exact Test). Moreover, there were 127 transcription factors (33.3%) among the 381 commonly bivalent genes, an equally significant enrichment (*P* < 2.2 e-16, Fisher's Exact Test) compared to the 10% of all genes that are transcription factors. The occurrence of homeobox genes and transcription factors is thus likely to represent a functional attribute of bivalent chromatin domains in GBM. The bivalent regions were characterized by punctate H3K4me3 marks, with H3K27me3 more broadly distributed across the region (**Figure 2.10c**). The commonly bivalent genes were highly interconnected, with 192 of the genes connected through StringDb [72]. 30 genes were not connected to the main network, so the primary network comprises 162 nodes. Using HumanNet, 176 of the nodes were connected (AUC = 0.612; *P* = 1.54e-12) [73]. The

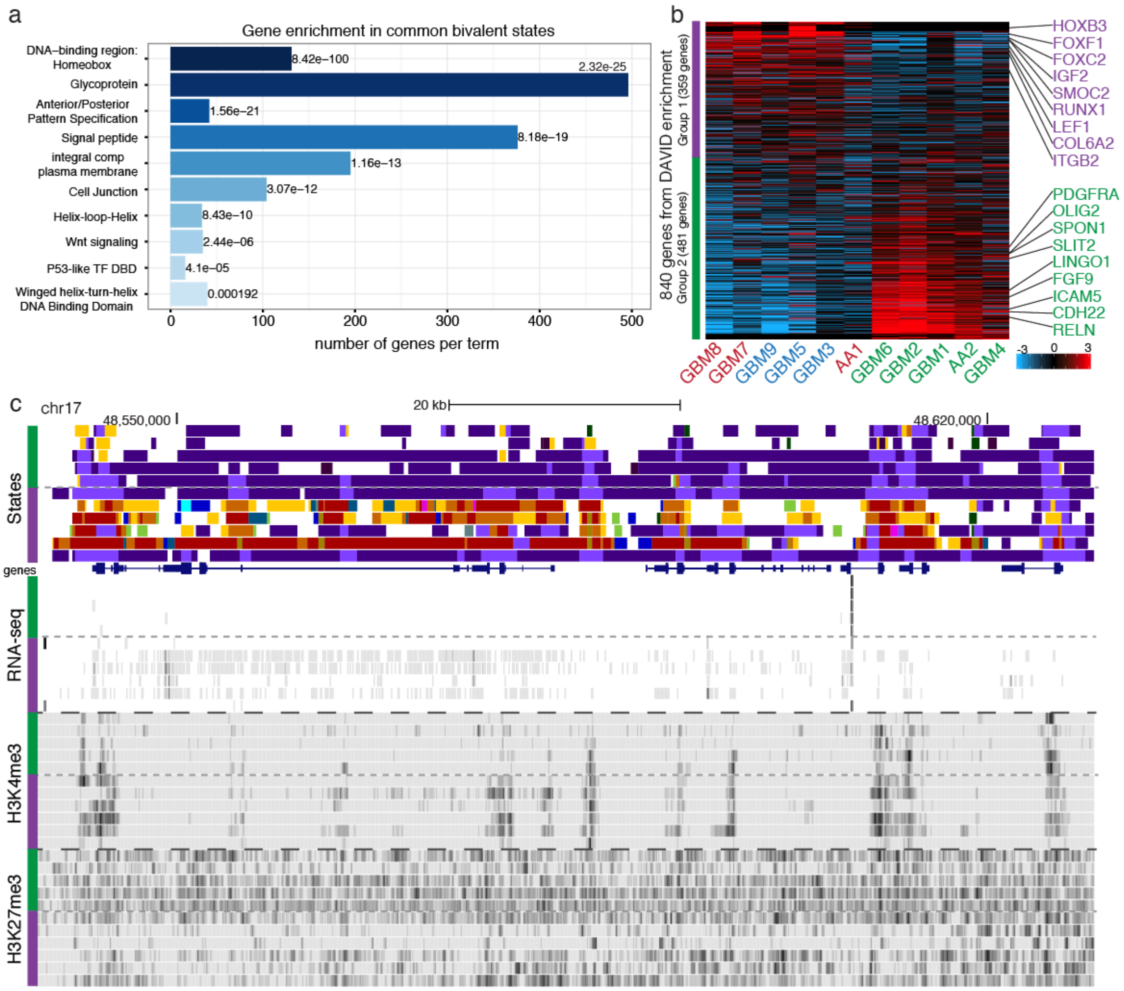


Figure 2.10: Bivalent regions underlie Wnt and SHH signaling in GBM.

(a) Functional enrichment for genes associated with bivalent domains in at least 5 tumors. FDR adjusted *P*-values provided by DAVID are shown above each column, with the shading proportional to the significance. **(b)** Heatmap of gene expression from the 840 genes with DAVID annotations in **a**, demonstrating clustering of PN and MES/CL tumors with regard to gene expression. The genes are sorted based on the ratio between average FPKM values across PN and MES/CL groups. **(c)** A genome browser view of a 90 kb region on chromosome 17, encompassing the HOXB cluster of genes. From top to bottom, tracks show chromatin states, genes, RNA-seq, H3K4me3 binding, and H3K27me3 binding. Proneural tumors (green bars) are bivalent, while 4 out of 6 MES/CL tumors (purple bars) show expression over this region. The hg38 coordinates of the region shown are chr17:48,539,612-48,628,935.

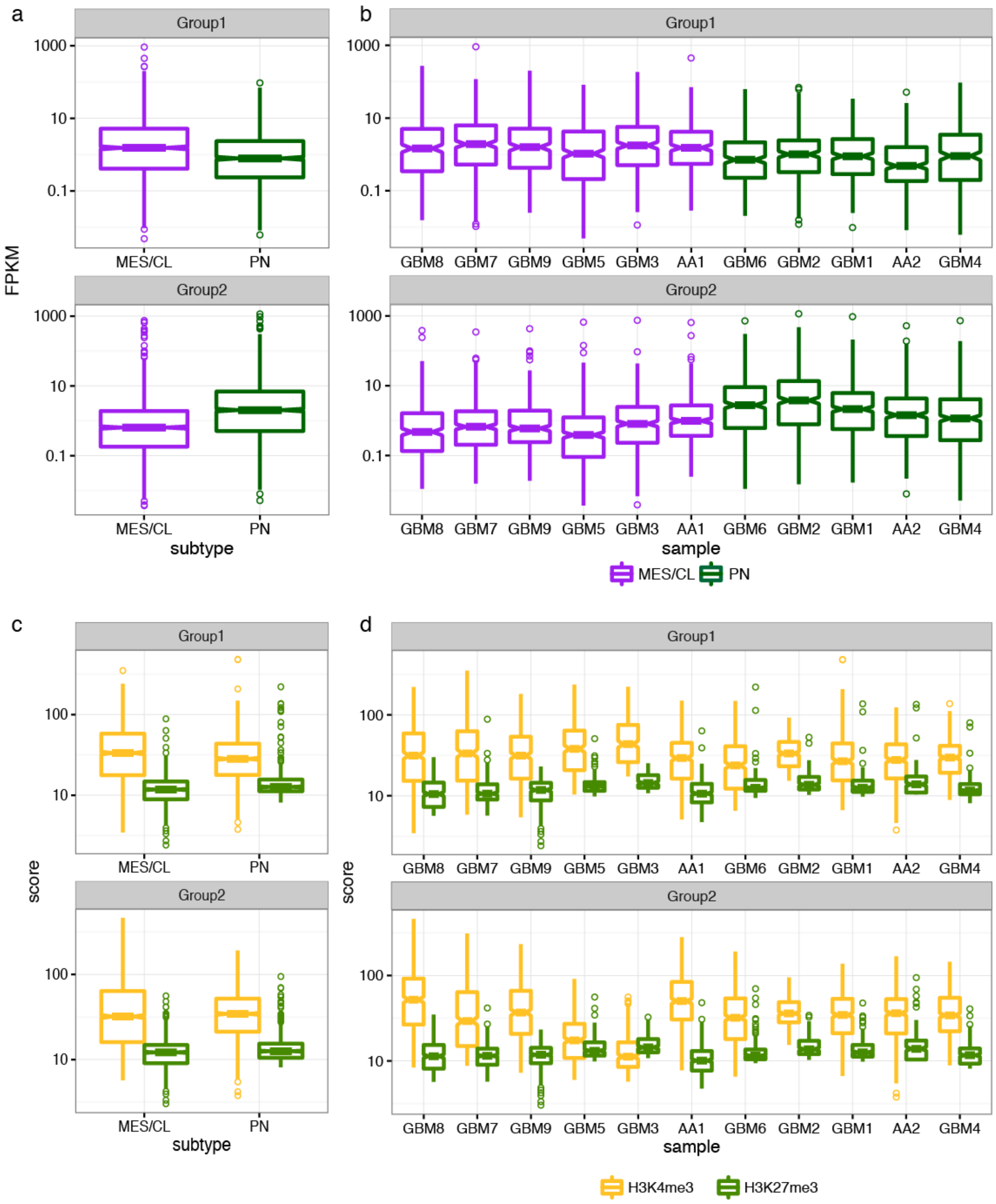
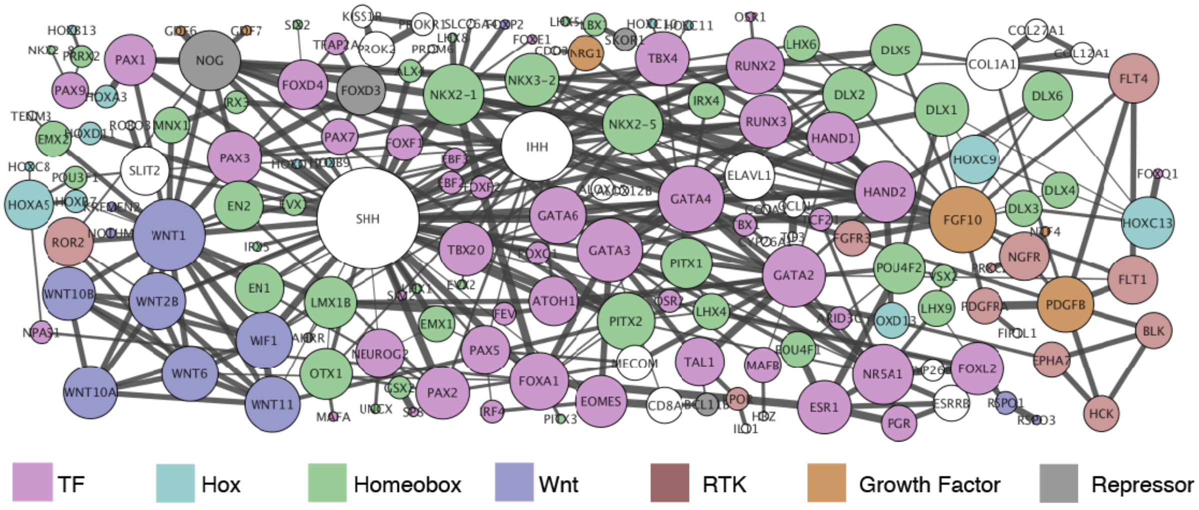


Figure 2.11 Expression and chromatin signal for bivalent domains and associated genes.

Figure 2.11 Expression and chromatin signal for bivalent domains and associated genes.

(a) Gene expression data in FPKM, separated by subtype and by bivalent group as shown in **Figure 2.10b**; Group 1 genes are expressed in MES/CL tumors and bivalent in PN tumors, Group 2 indicates genes that are expressed in PN tumors and bivalent in MES/CL tumors.
(b) As in **a**, but data have been broken out by tumor.
(c) Chromatin signal data for H3K4me3 and H3K27me3, separated by subtype and by bivalent group as shown in **Figure 2.10b**.
(d) As in **c**, but data have been broken out by tumor.



TF Hox Homeobox Wnt RTK Growth Factor Repressor

Figure 2.12: Interconnectivity between genes identified as bivalent in at least 8 tumors.

StringDb [72] was used to identify edges between nodes, and Cytoscape [74] was used to make the resulting plot. The number of connected edges scales with node size. The seven most common functional gene classes are colored, with a legend at the bottom of the panel; TF = transcription factor, RTK = receptor tyrosine kinase.

resulting network is dominated by *SHH* and *IHH*, with *WNT1*, GATA-family transcription factors (*GATA2*, *GATA3*, *GATA4*, *GATA6*), and the growth factor *FGF10* forming additional hubs (**Figure 2.12**). The presence of bivalent chromatin domains in cancer may indicate de-differentiation towards a more stem-cell like phenotype.

DISCUSSION

Although gene expression profiling of primary tumors suggests 4 molecular subtypes – classical, mesenchymal, neural and proneural – detailed phenotypic and molecular characterization of glioma stem cells (GSCs), which are thought to be the tumor initiating cells in glioblastoma, reveal two distinct subtypes of GSCs, corresponding to the mesenchymal and proneural types [75]. Strikingly, the genes targeted by the enhancers and bivalent chromatin states that we identified in primary tumors also separate them into two groups corresponding to the GSC-based classification of GBM. Moreover, genes targeted by enhancers appear to regulate pathways that are differentially active in the two GSC subtypes.

Many enhancer-associated genes that were significantly upregulated in MES/CL tumors promote cellular invasion and angiogenesis, a hallmark of mesenchymal GSCs [76, 77]. For example, *PODXL* (Podocalyxin-like) promotes cell migration, and its overexpression in GSCs is associated with a poor outcome [78]; *MMP11* (Matrix metalloprotease 11) cleaves the extracellular matrix, and promotes tumorigenesis and cellular invasion [79]; *S100A16* (a Ca^{++} binding protein) promotes the epithelial-to-mesenchymal transition (EMT) in breast cancer [80]; the protein kinase *FAM20C* is a marker of mesenchymal GSCs and promotes proliferation in triple-negative breast cancer [81]; *LMO2* (LIM Domain Only 2) promotes erythropoiesis and angiogenesis [82, 83], and is also a marker of GSCs [76, 77]. Integrin alpha 11 (*ITGA11*) is highly expressed in invasive triple negative breast cancer cells [84], and is involved in the tumorigenicity of non-small cell lung cancer [85].

Conversely, many enhancer-associated genes significantly upregulated in PN tumors were associated with increased survival. The *AKT3* isoform of *AKT* is inversely correlated with malignancy in GBM [86, 87], while dynamin-1 (*DNM1*) is associated with long-term survival in GBM [88]. Tenascin R (*TNR*) promotes the assembly of perineuronal nets, which stabilize synapses in the adult brain [89, 90]; in embryonic stem cells, it promotes differentiation into the neuronal lineage [91]. Neural cell adhesion molecule 1 (*NCAM1*) is involved in neuron-neuron interactions in the brain, and when repressed, the Wnt/β-catenin pathway is activated and cellular invasion increases *in vitro* [92]. *KCNIP3* interacts with potassium voltage gated ion channels, and is a negative regulator of N-cadherin processing [93], which may indicate an anti-cancer effect, given that N-cadherin expression is important for cancer cell metastasis [94].

The differentially expressed genes *MICAL2*, *STMN1* and *MYH10* are indicative of cellular proliferation and cytoskeletal destabilization in PN tumors. *MICAL2* destabilizes F-actin in both neural and non-neural cells [95] and is associated with EMT and actively metastasizing cancer cells; when knocked down *in vitro*, this effect is abrogated [96]. Stathmin-1 (*STMN1*) destabilizes microtubules, and its expression is increased in infiltrative astrocytomas [97], when *STMN1* is inhibited, cell proliferation is reduced, and cell cycle arrest is observed *in vitro* [98]. The unconventional non-muscle myosin *MYH10* is upregulated in PN tumors; unconventional myosins play a large role in control of cytoskeletal remodeling in support of lamellipodia spreading [99], as well as collagen synthesis [100]. Additionally, *MYH10* is important for moving the nucleus of a cell through tight spaces during migration to generate force [101]; this process is an important aspect of glioma invasion [102].

Many genes adjacent to frequently bivalent regions could be placed into two groups showing the same reciprocal relationship in expression between MES/CL and PN tumors as observed with enhancer-associated genes (**Figure 2.8c**, **Figure 2.10b**, **Figure 2.11**). Thus, many Group 1 genes active in MES/CL tumors, such as *COL6A2*, *SMOC2*, *ITGB2*, *FOXC2* and *HOXB3* have been associated with angiogenesis, cellular migration and invasive growth. In primary and metastatic brain tumors, *COL6A2* expression is associated with angiogenesis [103]. *SMOC2* is associated with endothelial cell proliferation and angiogenesis [104] and a Wnt-associated signature of stemness in intestinal crypt cells [105], which is required for progression of colon cancer [106]. *ITGB2* is involved in cellular migration, particularly of leukocytes along endothelial cell walls, which has implications for metastasis [107]; when mutated, it is associated with increased risk of glioma [108]. The Wnt signaling pathway transcription factor *LEF1* regulates stem cell renewal in GBM [109], and does this through activation of Wnt signaling [110]. The transcriptional activator *FOXC2* induces EMT [111, 112], enhances cell invasion and proliferation in GBM [113], and increases angiogenesis in a HUVEC model system [114]. The transcription factor *HOXB3* promotes invasiveness in prostate cancer [115], and when degraded, the cancer stem cell phenotype in ER+ breast cancer is inhibited [116]. In MES/CL tumors, the Hh target gene *GLI1* is highly expressed, and is an indicator of reduced survival in GBM [117]. The high expression of *GLI1* in the MES/CL group drives high expression of its target gene *MDM2*, a ubiquitin ligase which degrades *TP53* [118].

Some Group 2 genes active in PN tumors were protective. For example, *ICAM5* is an intracellular adhesion molecule that regulates interactions between neurons and microglia [119, 120], and it is often repressed in colon cancer [121]. Another Group 2 gene, *SLIT2*, provides axon guidance in the developing forebrain, and patients with *SLIT2* positive gliomas show better survival [122]. *LINGO1* is a negative regulator of myelination in oligodendrocytes [123, 124], and when inhibited in NSCs, the neuron lineage doesn't mature [125]; it promotes apoptosis after neural injury by inhibiting

WNK3 kinase [126].

However, other Group 2 genes highly expressed in PN tumors were strongly suggestive of GSCs. Notable among these was *OLIG2*, a lineage specific transcription factor for oligodendrocytes that is required for proliferation of GSCs [127, 128]. *OLIG2* regulates *PDGFRA* [129], which was also a Group 2 bivalent gene highly expressed in PN samples, and its expression is associated with an improved prognosis [130], though *PDGFRA* is also required for gliomagenesis [131]. Fibroblast growth factor 9 (*FGF9*) is a mitogen consistently expressed in human gliomas [132]; in ovarian cancer, *FGF9* expression is associated with activation of the Wnt and Hh signaling pathways [133].

Genes marked by bivalent chromatin in 70% of GBM (8/11 tumors) were highly interconnected and formed a network dominated by Wnt (*WNT1*, *WNT2B*, *WNT6*), and hedgehog (*SHH*, *IHH*) signaling, HOX and homeobox genes, and transcription factors. This bears strong similarity to signatures of bivalent chromatin both in embryonic stem cells, where the opposing active and polycomb repressed marks poise genes for developmental expression, as well as in cancer stem cells (CSCs) [62, 134, 135]. *WNT5B* was recently identified as vital for differentiation and cell growth in GSCs [139]. Expression of HOX gene loci in GBM is associated with a stem-cell signature of self-renewal and resistance to chemotherapy [136-138]. The enrichment for bivalent marks over HOX gene loci correlates with these findings, and indicates that bulk tumor may have some pre-existing activating histone modifications of GSC-relevant loci. However, true bivalent loci marked by H3K4me3 and H3K27me3 have not been identified previously in primary GBM tumors. Unexpectedly, the bivalent signature associated with

GSCs, which comprise a small fraction of the overall tumor, was instead observable in the bulk tumor.

Wnt and Hh signaling regulate EMT, invasion and proliferation in cancers, and certain components of these pathways were expressed in the GBM tumors profiled here. However, the master regulators *IHH*, *SHH*, and *WNT1* were nearly always silent, but poised for expression. The large number of transcription factors and homeobox genes that were commonly bivalent suggests a rapidly deployable program that allows for an Hh and Wnt-mediated transcriptional response that may drive the production of multipotent stem cells from more differentiated bulk tumor cells. Many of the genes expected to be signatures of GSCs specifically were highly expressed in unsorted tumor in a subtype specific manner, indicating that the genetic pathways necessary for GSC programming are present in any given GBM tumor cell. The Wnt and Hh pathways offer many opportunities for therapeutic intervention [140-142], and it is possible that activating the bivalent domains, perhaps using combinations of epigenetic modulators, could expose vulnerabilities in tumors that can then be targeted in a subtype specific manner.

MATERIALS AND METHODS

Chromatin immunoprecipitation in solid tumors and cell lines

All patients provided informed consent, and this study was approved by the Institutional Review Boards of St. David's Medical Center and of the University of Texas at Austin. Tumor tumors were collected during surgical resections as part of the standard of care, and only excess tissue that was not used for pathological analysis was used in this study. Tumors were immediately placed in a cryotube (Nalgene, Corning, NY) and flash frozen in liquid nitrogen after removal from the operating suite. Each tumor tumor was homogenized by crushing in a liquid nitrogen cooled Biopulverizer (BioSpec Products, Bartlesville, OK) mortar and pestle until particles were sub-millimeter size. A 10 mg sample of this powder from each tumor was reserved for RNA extraction and placed at -80°C. The homogenized tumor tissue was separated into aliquots by weight in sterile 15 mL conical tubes, and then suspended in PBS (Gibco, Life Technologies, Carlsbad CA) mixed with 10 μg/mL PMSF (Roche, Basel, Switzerland) in isopropyl alcohol, with 1% formaldehyde for cross-linking. Tumors were cross-linked for 15 minutes, rocking at room temperature, then washed with PBS + PMSF two times, centrifuging at 4°C, 500 g in between washes. Cross-linked tumors were flash frozen in liquid nitrogen and stored at -80°C until processing.

Samples were lysed in two steps to produce a crude preparation of nuclei. On ice, samples were resuspended in Farnham's lysis buffer (5 mM PIPES, 85 mM KCl, 0.5% NP-40), using 1 mL of buffer per IP. Tissue was dissociated into a single cell suspension using a 15 mL glass dounce (Wheaton), then incubated on ice for 10 minutes. Cells were centrifuged at 1200 g at 4 degrees for 10 minutes, and the pellet of mostly pure nuclei was gently resuspended in RIPA buffer (1x PBS, 1% NP-40, 0.5% Na Deoxycholate, 0.1% SDS) with protease inhibitors (COmplete EDTA-free tablets, Roche, Basel, Switzerland) to lyse for 10 minutes, again on ice, 500 μL RIPA per IP. The crude lysate was aliquoted (500 μl/tube) into polystyrene 15 mL conical tubes, and the tumors were sonicated in an ice bath, 30 seconds on, 60 seconds off, high intensity (Bioruptor, Diagenode, Denville, NJ). Sonication continued for 4 ten-minute cycles, with ice being replaced after each cycle. Tumors were centrifuged at 500 g at 4°C for 2 minutes to collect condensate, then transferred to microcentrifuge tubes and centrifuged at maximum speed (21000 g) for 15 minutes at 4°C.

Supernatant was transferred to fresh tubes, and volumes were brought up to 1 mL using RIPA buffer. For IP samples with rabbit antibodies, we used 30 μl of packed protein A beads per IP (Roche, Basel, Switzerland) as follows. Beads were washed three times in PBS plus 5 mg/mL BSA and protease inhibitors, with a 30 second centrifugation at 100 g in between washes. For each IP, 60 μl of resuspended beads in fresh wash solution were added, and samples plus beads were rocked for 30 minutes at 4°C. Samples were centrifuged at 100 g to pellet the beads, and the supernatant was transferred to a new tube. An input sample was removed at this stage (100 μl, 10% of total input) and antibodies were added for overnight incubation and rocking at 4°C. We performed the following IPs for each tumor: *CTCF* (EMD Millipore, Billerica, MA, USA, 07-729), H3K4me3 (EMD Millipore, 07-473), H3K4me1 (EMD Millipore, 07-436), H3K27me3 (EMD Millipore, 07-449), H3K9ac (EMD Millipore, 07-352), H3K9me3 (abcam, Cambridge, MA, USA, ab8898), H3K27ac (abcam, ab4729), using 10 μg of antibody per IP.

After the overnight incubation, we prepared protein A beads as before and added 60 μl of beads to each IP. Tumors plus beads rocked for an hour at 4°C, followed by six successive washes performed at 4°C. Each wash rocked for five minutes and beads were pelleted as described previously. Washes were performed in the following order: 2x low salt buffer (0.1% Na Deoxycholate, 1% Triton X-100, 1 mM EDTA, 50 mM HEPES (pH 7.5), 150 mM NaCl), 1x high salt buffer (0.1% Na Deoxycholate, 1% Triton X-100, 1mM EDTA, 50 mM HEPES (pH 7.5), 500 mM NaCl), 1x lithium chloride buffer (250 mM LiCl, 0.5% NP-40, 0.5% Na Deoxycholate, 1 mM EDTA, 10 mM TrisCl (pH 8.1)), 2x Buffer TE (10 mM TrisCl (pH 8.1), 1 mM EDTA).

To elute the antibody/DNA complexes from the beads, we resuspended the washed beads in 250 μl of freshly made 1% SDS and 0.1 M NaHCO3 buffer, and did the same with the frozen input sample. We rocked the samples for 15 minutes at 25°C, then collected the supernatant in a fresh tube and repeated the bead elution with an additional 250 μl of buffer. From here, we added 20 μl of 5M NaCl to each sample, and incubated the samples for at least 4 hours at 65°C in a water bath to reverse the formaldehyde cross-linking. We either froze the samples at -20°C or directly proceeded to the below steps for DNA extraction.

Samples were brought to 25°C and any residual RNA was removed by adding 5 μl 0.5 mg/mL RNase A to each tumor, then incubating for 30 minutes at 37°C. To digest proteins, we added 20 μl of 1 M Tris at pH 6.8, 10 μl 0.5 M EDTA, pH 8.0 and 3 μl 20 mg/mL Protease K to each tumor, then incubated them at 55°C for one hour. Samples were extracted with phenol-chloroform and precipitated with ethanol. We resuspended the DNA pellet in 15 μl sterile DNase and RNase free water (Ambion), and quantitated the DNA using a Qubit and Qubit HS DNA kit (Q32851, Life Technologies, Carlsbad CA, USA).

qPCR ChIP enrichment quantification and primers

To independently verify ChIP enrichments, we performed qPCR using positive and negative controls, as well as an input standard curve. We based the input standard curve on the percentage of input signal, and diluted the 10% input samples to produce data points at 1%, 0.1%, 0.01%, 0.001% and 0.001% (1:10, 1:100, 1:1000, 1:10000, 1:100000 dilutions, respectively). This was used to calculate a fold enrichment of the IP sample relative to input. Sequences for all primers used in this study are located in Table 2.4, below. We used Power SYBR Green PCR Master Mix (4367659, Applied Biosystems, Foster City, CA, USA) in 5 μl reactions and ran all plates on a ViiA7 RUO Real-Time PCR system located in the DNA Core facility at UT Austin.

Primer name	Sequence	Size (bp)	Amp. region hg38	Function
K4me3-1F	AGCATCAGGCCGTCAGCACA	78	chr8:144,791,510-144,791,587	active chromatin positive control
HK4me3-1R	TGCTGTGCTCGCAACTTCGC			
K4me3-2F	ACAGCATCCATGGCACCAACCC	150	chr1:52,404,778-52,404,927	active chromatin positive control
K4me3-2R	AAATGGGCCACAAGGGGGCT			
CTCF-F	TGGCAATGTTTTGAAAGCTG	61	chr13:98,840,224-98,840,284	CTCF positive control
CTCF-R	ACACCGCACTCCTTACTTGC			
H3K4me1-F	GAAGAACAGGGAATGGCAAA	77	chr8:119,982,737-119,982,813	H3K4me1 positive control
H3K4me1-R	GCCCTTCCCAGATAAAAAGC			
H3K27me3-F	TATGGTTGATTGCCTCGACA	68	chr1:37,631,891-37,631,958	H3K27me3 positive control
H3K27me3-R	AATGCTGCAATTAAAGGCAAA			
NegCtrl-F	GCAAGAGTCCTGGGTGAGAG	89	chr17:11,279,138-11,279,226	negative (no binding) control
NegCtrl-R	CGATGACAGTGCTTCTCTGG			

Table 2.4: qPCR primers used in this study.

Library preparation and sequencing

After quantitating the total DNA from each IP and assessing qPCR enrichment at positive control target sites, we prepared libraries according to the New England Biolabs NEBNext library prep kit (E6240L, New England Biolabs, Ipswich, MA, USA), with several changes to improve efficiency. We performed adapter ligation before size selection, we used Bioo adapters (514103, Bioo Scientific, Austin, TX, USA), and Ampure XP beads (A63881, Beckman-Coulter, Brea, CA, USA) in place of columns to do all reaction purifications and size selections. Generally, we also adjusted the number of cycles of PCR amplification to be as low as possible, often 6-8 cycles if feasible. Our sequencing was performed using standard Illumina chemistry on a HiSeq 2500 in either the Genome Sequencing facility at MD Anderson Cancer Center at Science Park (Smithville, TX) or the UT Austin Genome Sequencing and Analysis Facility (GSAF).

Alignment and peak calling in ChIP-seq data

Illumina paired-end sequencing of ChIP-seq libraries produced datasets for 7 marks: 6 histone marks (H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3) and the *CTCF* transcription factor, for 11 tumors. Sequencing of input libraries was also performed for each tumor. BWA (v0.7.12-r1039)[143] was used to align all ChIP-seq and input sequence data. All reads were hard-trimmed to 50 bases, and the aln and sampe commands from BWA were used to align the data to the hg38 (GRCh38) assembly from GENCODE version 24. We used the MarkDuplicates tool from the Picard suite (v1.123) (http://broadinstitute.github.io/picard) to flag duplicate sequences.

Given that ChIP-seq data can be variable from experiment to experiment with regard to signal depth, signal to noise ratio and read duplication levels, we used the Phantompeakqualtools v2.0 [144] tool from ENCODE. This tool reports a normalized strand cross-correlation coefficient (NSC), and we recorded this value while allowing 0, 1 or 2 duplicate reads. From this analysis, we determined that allowing 1 duplicate read resulted in the best tradeoff between adequate signal strength (indicated by an NSC value of at least 1.05) and minimizing signal covariance across each set of tumor experiments.

To identify enriched regions (peaks) in our ChIP-seq data, we utilized the MACS2 (v2.1.1.20160309)[145] callpeak function. We specified a relaxed *P*-value threshold of 0.01, and allowed one duplicate read per locus (keep-dup=2), with each ChIP-seq sample being paired with its input control library. We removed any peaks falling in genomic regions with high signal due to high copy number differences [146] using bedtools intersect with the –v option. Additionally, we removed low-complexity regions using the "Duke Excluded Regions" and "DAC Blacklisted regions" tracks from the UCSC genome browser (http://genome.ucsc.edu/cgibin/hgTrackUi?hgsid=334775099&c=chrX&g=wgEncodeMapability) and lifted these regions over to the hg38 genome assembly [147].

After initial peak calling, statistics were gathered; these included peak counts at a wide range of MACS2-reported *Q*-value and fold enrichment (FE) levels, and total coverage of the hg38 genome for each FE level. *P*-values and *Q*-values were strongly affected by sequencing depth for an experiment; we found that FE was more robust and less variable as a measure to define signal thresholds across each set of experiments. Based on peak counts and genome coverage at several FE thresholds, we selected three significance thresholds: high (FE 6+), target (FE 4.5+) and low (FE 3.5 or 4). The MACS2-generated “narrowPeak” format files were converted it a custom BED9+ format, with *P*-value and *Q*-value fields, FE, peak rank among all peaks in that experiment, and a significance level designator. All reported results in this study were generated using the peaks at the target level (FE 4.5+). We chose the above FE thresholds to balance between variance in FE across mark sets and genome coverage. At the target FE threshold, most *Q*-values were < 0.01, with all being < 0.05.

Finally, we used the MACS2 bdgcmp signal processing tool (https://github.com/taoliu/MACS/wiki/Build-Signal-Track) to generate genome-wide signal tracks. We used the linear-scale fold-enrichment method to compute per-base scores (-m FE). This process produced bigWig files which represent sequencing depth normalized fold enrichment of ChIP-seq peaks over input signal. These were loaded into the UCSC Genome Browser [148] and visualized on the hg38 genome assembly to produce figures. To prepare the data for building a model using ChromHMM, we collected the target level peaks (FE 4.5+) and bedtools merge was used on each individual bed file to combine any overlapping peak regions (v2.25.0)[149].

Chromatin states: systematic identification of histone co-localization

From the above merged consensus peaks for each experiment, we used ChromHMM v1.12 [55] to build our 21 state model, using only tumor data (11 tumors, 7 experiments per tumor, 77 total datasets used). We made a table of tumors and experiments and binarized our bed file peak data using the BinarizeBed functions, with the "-center" and "-peaks" options set. With our data binarized, we used the LearnModel function, with multithreading and iterations set to "-p 16 -r 200" with 21 states on the hg38 assembly. We assessed the genomic coverage of each state using bedtools coverage on hg38. Finally, we re-affiliated our model states with peak fold enrichment scores by using bedtools intersect for each tumor and model state. To understand which genes are regulated by which states, we used the GENCODE [57] annotation, version 24, and used bedtools closest to identify the distance between chromatin states and the most proximal protein coding genes. For enhancers, we analyzed the closest gene even if it was very far away (1 Mb or greater), but for bivalent regions, the closest gene (and thus the bivalent domain associated with it) was not analyzed unless the distance was less than 500 bp.

Assessing gene enrichments for common enhancer and bivalent states using DAVID

In order to assess enriched pathways and terms for our common enhancer and bivalent domains, we utilized DAVID 6.8 [71]. We identified enriched terms using functional annotation clustering. The DAVID terms in **Figure 2.8c** and **Figure 2.10a** were derived from functional annotation clustering as follows. For each functional annotation cluster, the term encompassing the largest number of genes while still having a significant *P*-value by the Benjamini correction was chosen, and an additional term could be chosen from any given cluster if the terms did not describe redundant features.

Annotation clusters were skipped if they appeared completely redundant to any previous cluster, or non-significant by Benjamini adjusted *P*-value of less than 0.05. This process was continued until at most 10 significant terms were identified.

RNA sequencing in solid tumors and cell lines

During the initial tissue processing, an approximately 10-15 mg sample of crushed tissue was set aside for RNA extraction. Tumors were removed from the -80° freezer, then 1 mL of TRIzol reagent (15596-026, Thermo Fisher Scientific, Waltham MA, USA) was added to each tumor, and the standard TRIzol protocol was followed for RNA isolation. RNA concentration was quantified using the NanoDrop ND-1000 (NanoDrop Products, Wilmington, DE, USA). The Ribo-Zero rRNA removal kit (MRZH116, Illumina, San Diego, CA, USA) was used to remove ribosomal RNAs, and the resulting RNA was used to prepare single-end or paired-end libraries with the NEBNext small RNA kit for Illumina (E7300S, New England Biolabs, Ipswich, MA, USA). The resulting libraries were run on an Illumina HiSeq 2500 as above. Nearly all cell line experiments were represented by two biological replicates, although some were represented by three replicates (NHA, U87MG, A172).

We began pre-alignment processing of RNA-seq data by removing any 3' Illumina adapters from fastq reads using cutadapt (v1.10)[150]; any sequences shorter than 36 bases after trimming were discarded. For experiments with fastq data from more than one sequencing lane, files were combined. Reads from rRNA and tRNA were removed by aligning to a reference containing human rRNA and tRNA sequences using

BWA (v0.7.12-r1039)[143], then retaining sequences that did not align. Tophat2 (v2.1.0)[151] was used to align the remaining sequences in a transcriptome-aware manner, aligning to the hg38 (GRCh38) assembly, using GENCODE release 24 to build a comprehensive gene annotation set. We used the MarkDuplicates tool from the Picard suite (v1.123) (http://broadinstitute.github.io/picard) to identify duplicate sequences in the resulting BAM files.

We utilized the Tuxedo suite [152] to perform transcript FPKM quantification on RNA-seq datasets. Each sample was quantified using cuffquant (cufflinks v2.2.1), using GENCODE version 24 annotations and specifying the –multi-read-correct and –frag-bias-correct options. The resulting cuffquant CBX files were gathered and processed together using cuffnorm (cufflinks v2.2.1), utilizing the geometric normalization method. We produced two sets of normalized data: one for tumors and cell lines, and another for only tumors. The resulting genes.fpkm_table files were used for subsequent FPKM expression analyses, including differential gene expression analysis.

Assignment of TCGA subtypes to tumors

To identify the TCGA molecular subtype of each tumor in this study, we extracted the ENSEMBL gene IDs from the 841 gene names from the TCGA subtyping study [58] (extracted from their supplementary file TCGA_unified_CORE_ClaNC840.txt). Against our cuffnorm data we identified 743 direct matches. We identified another 92 matches using HGNC name information, as some gene identifiers had changed between hg18 (the genome version used for the 2008 TCGA study) and hg38 (the version used in this study). The final result was 836 matches. FPKM values for the matching genes were extracted and ordered in the same manner as in the TCGA clustered data. In R, we processed the resulting matrices as follows: missing or 0 values were replaced by the smallest R double precision value, the resulting matrix was log2 transformed, rows were median centered and complete Spearman pairwise correlation of columns was performed to identify a matrix of differences. Finally, average-linkage hierarchial clustering was performed on the columns represented by the distance matrix – we assigned the TCGA subtypes mesenchymal, classical, and proneural to the tumors based on the three main branches of the resulting dendrogram.

Differential gene expression analysis

Based on the above assignment of TCGA subtypes to the samples used in this study, we performed differential gene expression analysis. Based on enhancer and bivalent expression patterns, we identified genes differentially expressed between mesenchymal/classical and proneural tumors using cuffdiff (cufflinks v2.2.1). We utilized the geometric normalization method, the pooled dispersion method, a minimum alignment count of 4, and the multi-read-correct and frag-bias-correct options. Group assignments were as follows: mesenchymal/classical: AA1, GBM3, GBM5, GBM7, GBM8, GBM9; proneural: AA2, GBM1, GBM2, GBM4, GBM6. Genes described as significantly differentially expressed between mesenchymal/classical versus proneural are those marked as significant at *Q*-value <= 0.05 by cuffdiff.

Chapter 3: Differential expression in GBM tumors and GBM-derived cell lines

Cultured cell lines derived from primary tumors have drastically increased mechanistic knowledge of cancer proliferation through their use as an *in vitro* model of many cancers. However, the act of culturing tumor-derived cells long term introduces widespread genetic and epigenetic changes in cellular identity [5, 6]. These changes are caused in part by the lack of a stromal microenvironment in the case of solid tumors, such as cancers of the breast, brain, and thyroid [153]. Tumors are removed from their complex three-dimensional, often necrotic and/or hypoxic environment into an environment with a surplus of growth factors, including glucose, and passaged under atmospheric, not physiological oxygen levels. We identified widespread gene expression changes in commercially available GBM-derived cell lines when compared to primary GBM tumors. Of the 4945 genes that were significantly upregulated in GBM primary tumors, 394 are consistently expressed in the GBM-derived cell lines examined and are enriched for genes governing cell motility, as well as chromatin remodeler and kinases.

INTRODUCTION

In glioblastoma multiforme (GBM), cultured GBM-derived cell lines have existed since 1969 [154, 155], however recent studies have established that cultured glioma stem cells (GSCs) may be a better model of GBM [156, 157]. Cultured GSCs recapitulate certain aspects of tumor subtype – GSCs cultured from mesenchymal tumors have a gene expression pattern that reflects this subtype, and the same is true for proneural lesions as well [81, 158-160]. Despite the existence of this improved *in vitro* model, establishing primary GSC culture from GBM tumors is not a trivial protocol [161-163]. As a result, many researchers continue to use decades old, commercially available GBM-derived cell lines. Their culture conditions are simple (defined media and fetal bovine serum), and as a result there are thousands of publications that rely on the seven cell lines used in this study (counts from Pubmed searches for the following terms on 4/3/17: U87MG = 1613; T98G = 1074; A172 = 642; LN229 = 194; LN18 = 91; U118MG = 62; U138MG = 58). Many of these studies were published after 2015, so this is not just reflective of an older model being superseded by a new model over time.

Despite the existence of superior *in vitro* models of GBM, much bench work is still reliant on a small group of immortalized cell lines. Given the widespread changes in cellular identity induced by cell culture [164], and the issues with reproducibility and relevance widespread in biomedical research [165], researchers need to be able to justify the use of a given cell line as a disease model. Quantification of the differences between primary tumor samples and immortalized cell lines will allow for the more judicious use of these lines. By understanding which genes and pathways in each cell line most closely resemble gene expression in primary tumors, experiments in these lines can be assessed for their clinical relevance in cultured GSCs or primary tumor samples.

RESULTS

There are 4945 genes that are significantly upregulated in tumors

As part of examining gene expression in cell lines to determine their suitability for subtyping (**Figure 2.2**), we examined gene expression across all protein-coding genes,

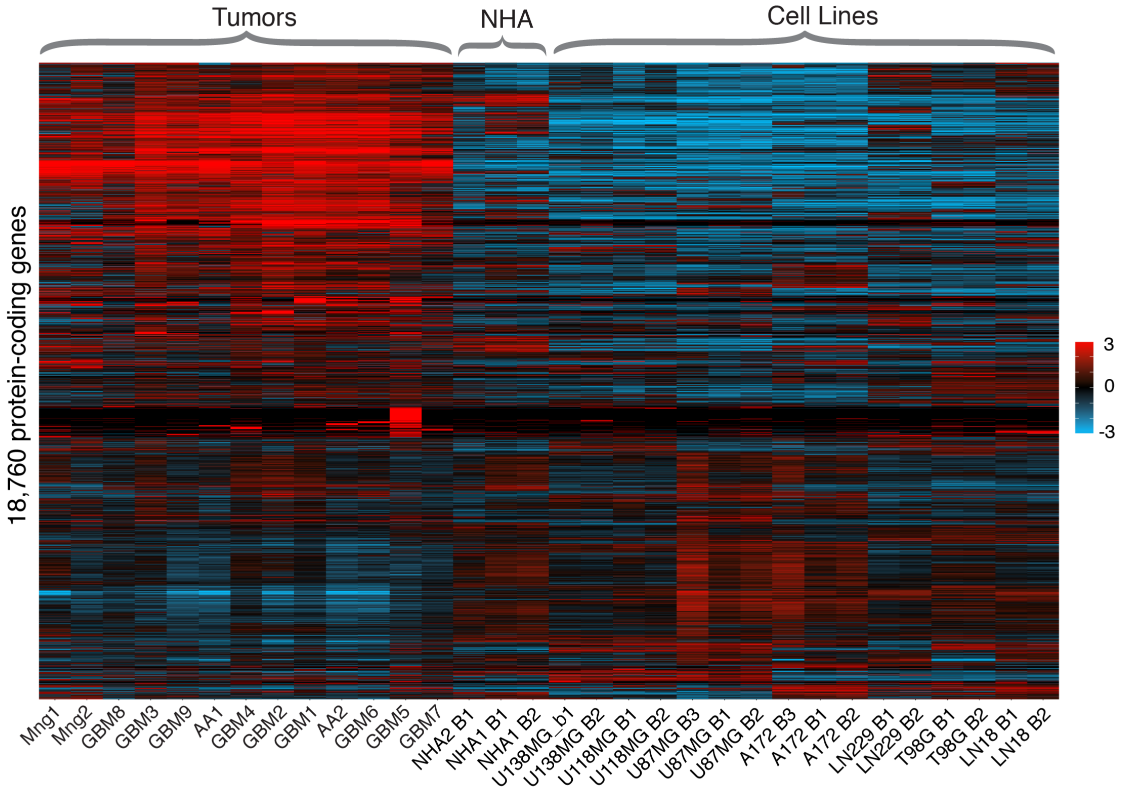


Figure 3.1: Tumor vs. cell line expression over all protein coding genes

Spearman's rho was used to cluster rows and columns over 32 RNA-seq datasets from 11 brain tumors, two meningiomas, normal human astrocytes (NHA) and 7 GBM-derived cell lines available from ATCC. Protein-coding genes were defined using GENCODE version 24. For the cell line and NHA data, at least two biological replicates were used.

in the 32 RNA-seq samples produced and described in Chapter 2. The clustered expression matrix revealed stark differences in gene expression between primary tumors, normal human astrocytes and immortalized GBM-derived cell lines (**Figure 3.1**). Despite these gross differences, there are groups of genes in each cell line that mimic the expression pattern of primary tumors.

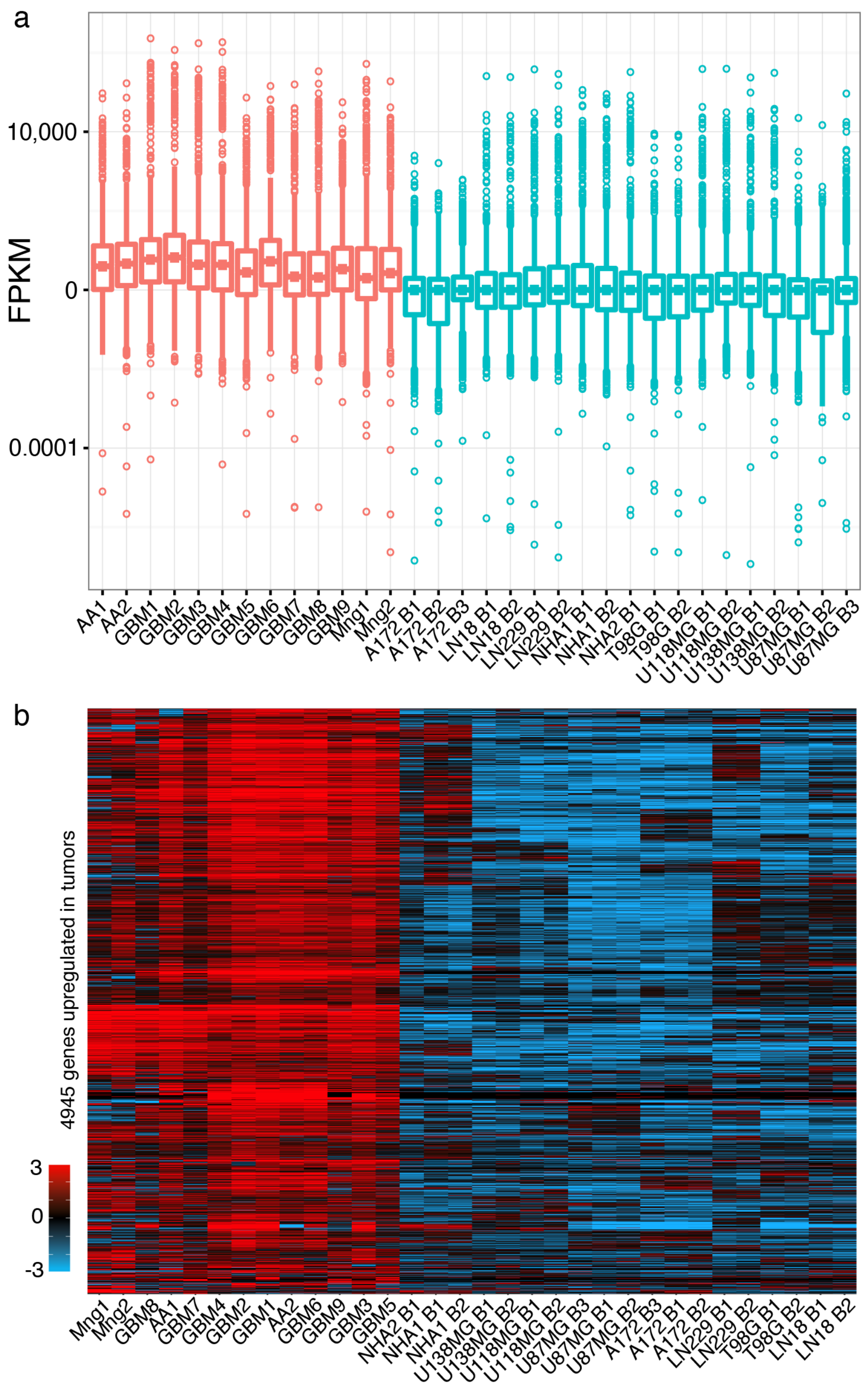


Figure 3.2: 4945 genes are upregulated in tumors compared to cell lines

Figure 3.2: 4945 genes are upregulated in tumors compared to cell lines

(a) Box plot of FPKM gene expression values over the 4945 genes that were upregulated in tissue for all RNA-seq experiments. Data have been normalized to remove the confounds of sequencing depth using cuffnorm, and significantly upregulated genes were identified using cuffdiff, with a *Q*-value threshold of less than 0.05.
(b) A heatmap of the log2 transformed data in a. An expression heatmap of 4945 genes significantly up in tumors, over all RNA-seq datasets. Data are clustered using Spearman's rho on rows and columns.

Given the visually obvious differences in expression between tumors and cell lines, we systematically identified differentially expressed genes across these groups using cuffdiff (Methods). Using tumor samples as one group, and cell lines as another group, we identified 4945 genes that are significantly more highly expressed in tumors, and 1416 genes that are significantly more highly expressed in cell lines. For the genes were significantly upregulated in tumors, the expression for the cell lines was much lower, but there was some overlap in the distribution (**Figure 3.2a**). This indicates that the majority of tumor-specific genes are minimally expressed, but each cell line contains a subset of genes where expression is more similar to the tumors. This trend is visible in the heatmap in **Figure 3.2b**.

To identify tumor specific genes that are expressed in the GBM-derived cell lines, we used the R package "reshape2" (https://github.com/hadley/reshape) to melt the cell line expression data into a "long" format, where rows represent all cell line-gene-FPKM combinations. Since the median FPKM for tumor-specific genes was 3.7 across all 4945 genes and 13 samples (**Figure 3.2a**), this cutoff was used to select cell line-gene-FPKM combinations. This threshold resulted in 2025 genes where cell line expression is at least at or above the median tumor level of expression, and 2920 genes where cell line expression is below the median tumor expression (**Figure 3.3**).

394 genes are highly expressed in all GBM-derived cell lines

Over the 2025 genes that are tumor-specific, but expressed in cell lines, the sets of genes specific to each cell line were identified. Interestingly, 394 genes passed the threshold in all seven cell lines (**Figure 3.4a,b**). These genes represent commonalities in

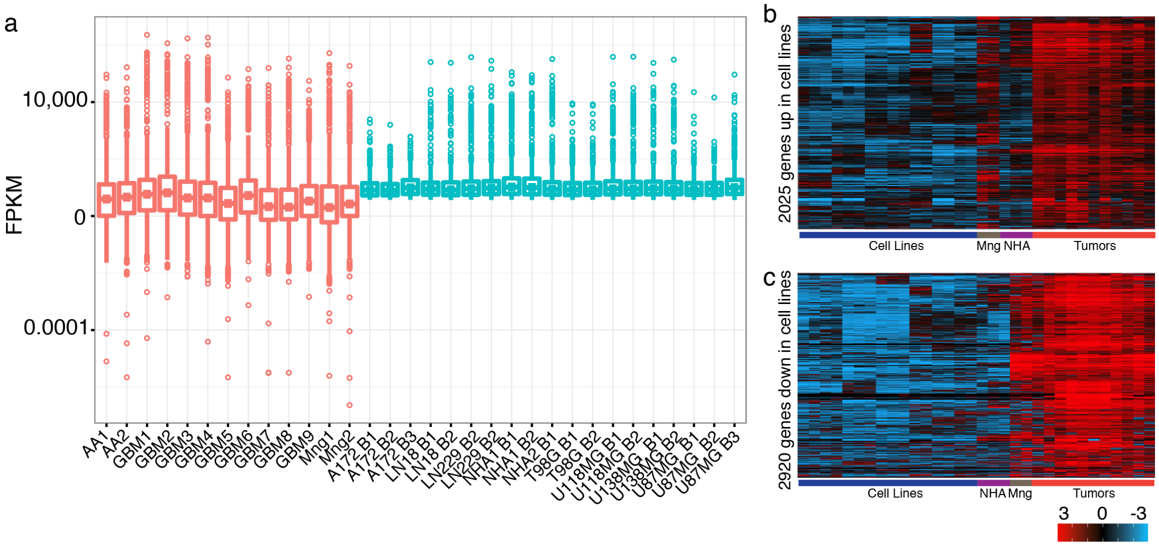


Figure 3.3: Identifying tumor specific genes expressed in GBM cell lines

RNA-seq data from cell lines was separated into genes expressed above and below the median tumor FPKM for upregulated genes (3.7).
(a) Box plot of FPKM values for tumors, and cell line data after filtering for gene-value pairs with an FPKM above 3.7.
(b) A heatmap representation of the 2025 tumor specific genes that passed the FPKM threshold for cell lines.
(c) The 2920 tumor-specific genes that didn't pass the FPKM threshold (cell line expression is below the median)

expression for all tumors and GBM-derived cell lines, so the genes and pathways may be the most reflective of primary tumor expression in studies that rely on these cell lines. The DAVID terms enriched in this set were diverse, and indicated common expression of GAP proteins, many of which are involved with cytoskeletal reorganization. The large number of proteins containing pleckstrin homology-like domains [166], as well as SH3 domains [167] and microtubule or cell junction interaction corroborates this signature (**Figure 3.4c**). Many proteins containing RING/FYVE/PHD-type zinc fingers were E3

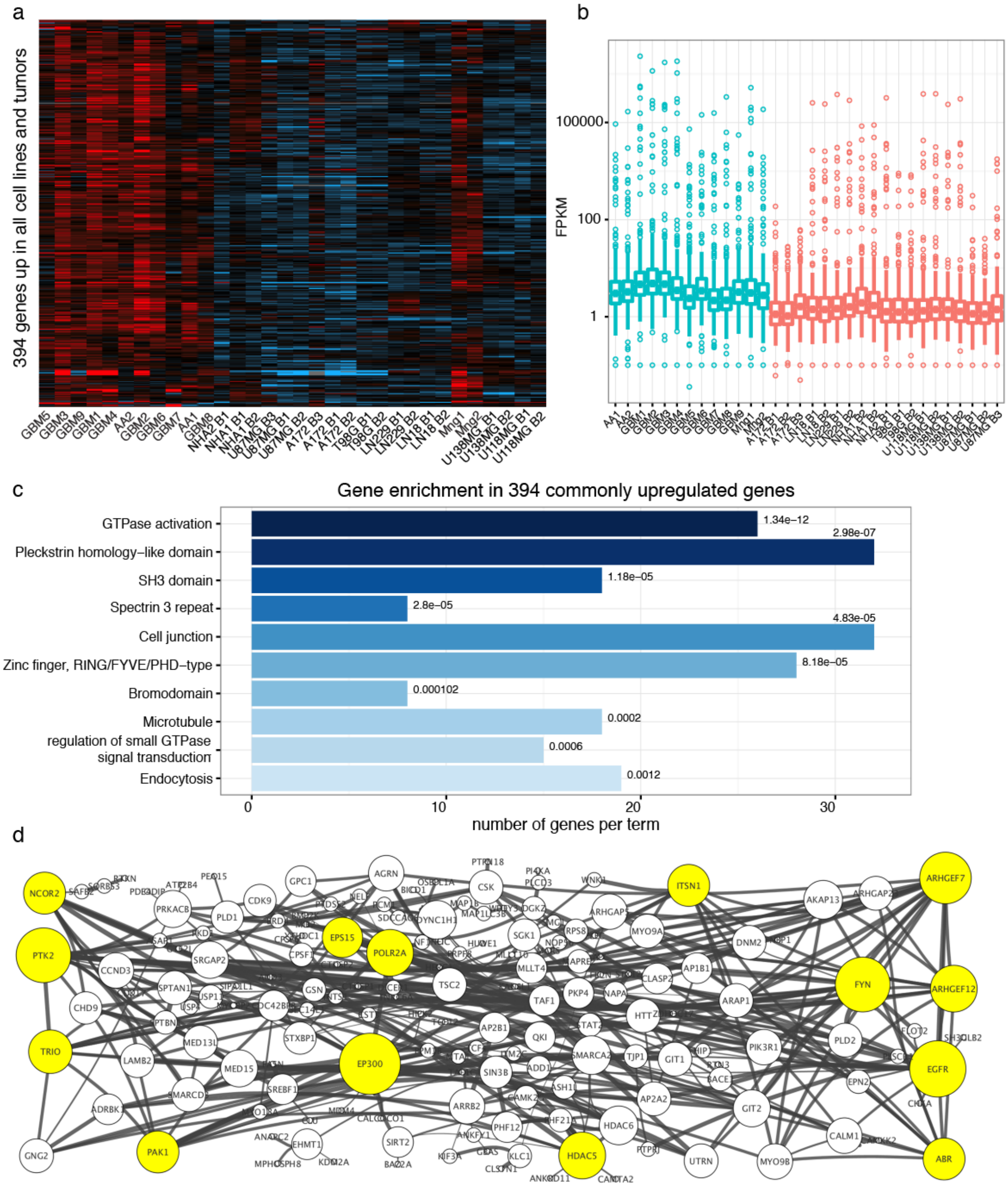


Figure 3.4: Identification of 394 genes expressed in all cell lines

Figure 3.4: Identification of 394 genes expressed in all cell lines

(a) A heatmap of the 394 genes that were commonly identified as being above the 3.7 FPKM threshold in **Figure 3.3**.
(b) A boxplot of the data represented by the heatmap in **a**.
(c) Enriched DAVID terms for the 394 genes represented in **a** and **b**.
(d) Interactions between the 394 genes, as identified through StringDb [72]. This network comprises 167 nodes and 337 edges. Node size is a function of the number of connected edges, and edge weight represents the confidence of the interaction between the two connected nodes. Nodes highlighted in yellow have at least 10 connections.

ubiquitin ligases (*RNF19A*, *DTX3*, *UBR3*, *MIB2*, *MARCH6*) or chromatin remodelers. The chromatin remodelers contained both bromodomain proteins (*BAZ2A*, *BAZ2B*, *BPTF*) and histone modifying enzymes (*ASH1L*, *PHF21A*, *KMT2C*, *KDM2A*, *KMT2E*, *HDAC6, EP300*).

Given the overlapping DAVID terms, these genes were assessed for their interconnectivity using interaction data from StringDb [72]. The resulting network was highly interconnected, with 167 genes being interconnected by 337 edges (**Figure 3.4d**). The connectivity was distributed, with 14 genes having at least 10 edges. These hubs formed three distinct groups: kinases and their effectors (*FYN*, *EGFR*, *PTK2*, *PAK1*, *EPS15*), chromatin remodelers/transcriptional regulators (*EP300*, *NCOR2*, *HDAC5*, *POLR2A*), and GTPase regulators and exchange factors (*ABR*, *ARHGEF7*, *ARHGEF12*, *TRIO, ITSN1*).

DAVID enrichment for each GBM-derived cell line

We identified lists of tumor-specific genes above the FPKM threshold in each cell line. From these lists we subtracted the 394 genes that are consistently up, described in the previous section, to avoid background from the genes and terms indicated in **Figure 3.4**. Each cell line presented the “cell junction” and “pleckstrin homology-like domains” DAVID terms, but other terms seemed more specific (**Figure 3.5**).

Normal human astrocytes (NHA), which are cultured but non-immortalized, were enriched for the immunoglobulin I-set and C2H2 zinc fingers. The microglial cell line U87MG is enriched for C2 domains and presynapse formation. The U138MG cell line (listed with U118MG as they are identical [168]) seems to have some specific enrichment

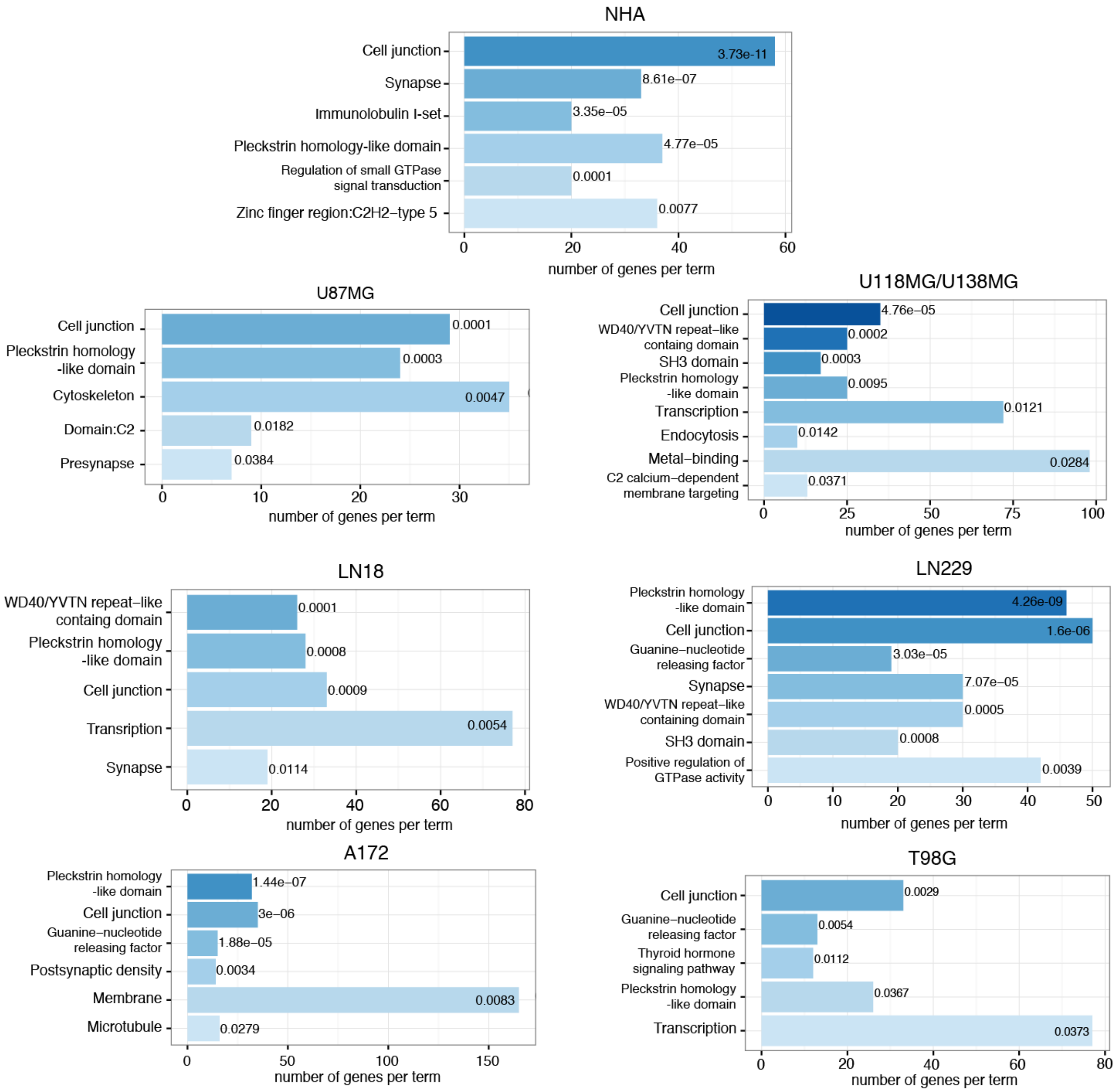


Figure 3.5: DAVID enrichment in each GBM-derived cell line.

The 394 genes common to expression in all cells line were removed from the gene set for each cell line analyzed, and the remaining gene IDs were examined using DAVID 6.8. The top terms from the Functional Annotation Clustering tool were aggregated as described in the Chapter 2 methods, with no more than 10 significant terms being accumulated for any given cell-line.

for endocytosis and metal binding. LN18 only has the generic term "transcription" as a unique feature, while LN229 has enrichment for SH3 domains (along with U138MG) and positive regulation of GTPase activity. A172 is enriched for membrane specific proteins and post-synaptic density, and the commonly used cell line T98G is enriched for the thyroid hormone signaling pathway. Three groups (U118/U138, LN229, LN18) contain the term "WD40/YVTN domain", which is a propeller shaped domain often used as a scaffold for large multiunit complexes [169].

DISCUSSION

Although GSCs are an effective *in vitro* model for GBM, their specialized culture conditions indicate they are currently less likely to be used than commercially available cell line models of GBM. These cell lines do not resemble primary tumors in their expression patterns (**Figure 3.1)**, however, given their widespread use as models of cancer, it is useful to identify which genes are commonly expressed in tumors and cell lines. Using the Tuxedo Suite [152], we identified 4945 genes significantly upregulated in primary GBM tumors, and examined which of these genes were expressed in each cell line (**Figure 3.2, Figure 3.3**).

We identified 394 genes that were highly expressed in tumors and cell lines (**Figure 3.4**) and found the resulting gene interaction network to be highly and diffusely interconnected, with 14 genes having at least 10 connections. These hub genes clustered into three groups: kinases, GTPase regulators and exchange factors, and chromatin remodelers. Many of these genes have a prior association with GBM.

The kinase hub genes (*FYN*, *EGFR*, *PTK2*, *FAK1*, *EPS15*) are all involved in cellular migration and adhesion in some fashion. *EGFR* is well known in GBM [170], so its presence is not surprising, however *EPS15* is involved in the recycling of *EGFR* [171], and expression of *EPS15* is associated with a favorable prognosis in breast cancer [172] and lung cancer [173]. The non-receptor tyrosine kinase *FYN* is necessary for cellular migration in GSCs [174], and is an effector of *EGFR* [175]. The tyrosine kinase *PTK2* is involved in cell migration and motility, its inhibition reduces the aggressiveness of GBM cells [176]. Finally the serine-threonine kinase *PAK1* (also known as *P21*) is involved in myriad aspects of cell proliferation and migration, and is also implicated in the invasiveness of GBM [177].

The genes annotated as chromatin remodelers (*EP300*, *NCOR2*, *HDAC5*, *POLR2A*) are a mixed group, with some being activators and others repressors. *EP300* is a histone acetylase, and drives cellular differentiation, and its expression in GBM is associated with an improved prognosis [178]. Nuclear co-repressor 2 (*NCOR2*) promotes chromatin condensation, and when targeted by an antagonistic microRNA (miR-100), cell proliferation is reduced [179]. In breast cancer, *NCOR2* expression is associated with resistance to anti-estrogen therapies, such as tamoxifen [180]. Expression of *HDAC5* is positively correlated with survival in GBM [181], but in other cancers, expression is associated with cellular proliferation and metastasis [182-185]. The presence of RNA polymerase II alpha subunit (*POLR2A*) is likely indicative of cells that are actively proliferating, but its ubiquity as part of the transcriptional machinery means that further associations are difficult to ascertain.

Of the GAP and GEF proteins (*ABR*, *ARHGEF7*, *ARHGEF12*, *TRIO, ITSN1*), two have a strong association with GBM. *TRIO* is a Rho GEF involved in actin-cytoskeleton reorganization, and mediates the invasive behavior of GBMs [186, 187]. *ITSN1* acts as a GEF for *CDC42*, and is essential for GBM proliferation [188]. *ABR* and *ARHGEF7* have no documented cancer associations, however *ARHGEF12* regulates cell morphology and invasion [189], particularly with regard to colorectal cancer [190]. These hub genes indicate cross talk between pathways mediating cytoskeletal reorganization, chromatin modification and cellular proliferation.

For genes specific to each cell line, the DAVID enrichments are relatively non-specific, but capture many genes. From a pathway perspective, cell junctions, pleckstrin-homology domains, and GTPase activity or GEF terms indicate that cellular migration and cytoskeleton remodeling are important forces governing growth in these lines. While the terms are the same, the genes associated with terms differ somewhat by cell line, thus this data should be valuable for determining which cell line to use in an experiment, given the genes or pathways of interest. The vast majority of DAVID terms are not unique to a single cell line – most occur in at least three datasets (**Figure 3.5**). While each cell line contains tumor specific genes that are also specific to expression in that cell line, the number of gene identifiers is low (less than 100) such that there is insufficient statistical power for any resulting DAVID term to be significant.

In summary, we have determined that the expression of primary GBM tumors and immortal cell lines derived from GBM tumors are widely divergent. A small subset of genes in each cell line align with expression of 4945 tumor-specific genes, and 394 of these genes are common to all seven cell lines examined in this study. To extend this analysis, obtaining and normalizing GSC and additional GBM expression data will further define changes in gene expression with regard to comparing primary tumors and cultured material. As certain immortal GBM-derived cell lines have been cultured as neurospheres [191, 192], comparison with these data would also indicate how using advanced culture conditions can improve the clinical relevance of commercially available cell lines.

Cultured cell lines provide a powerful and flexible system for validating the activity and effects of genes, mutations, and pathways identified in clinical cancer data. As tumor samples are generally precious and limited in supply, cell lines provide an easy way to interrogate cellular response when various genetic pathways are manipulated. Cultured cell lines also provide ample source material for validating the activity of enhancers and other genomic elements using techniques like luciferase assays. These methods are invaluable for validating initial results from clinical pathology and mutation analysis. However, given the widespread differences in expression evident when comparing clinical tumor samples to cultured cell lines, careful thought must be given to how cell lines recapitulate models of disease.

Glioma stem cells (GSCs) and other "sphere" based cultured methods use more complex cultured conditions than most immortalized lines. These conditions are more expensive and time-consuming to maintain, but also result in a much more biologically relevant model system, especially with regard to testing therapeutic options. However, immortalized cell lines, when cultured in standard conditions, are excellent workhorses for validating cell proliferation in the context of pharmacological inhibitions. Given this, a dual use model of *in vitro* cell line use may make the most sense. Use immortalized cell lines to rapidly screen for compounds of interest, but validate the most promising findings in a more complex model of GSCs (or the appropriate "sphere" culture model for the cancer or tissue under study).

Materials and Methods

RNA sequencing and analysis, differential gene expression analysis, and DAVID annotation of functional terms were performed as described in the Chapter 2 Methods section.

Cell lines and culture conditions

The following cell lines were purchased from ATCC (Manassas, VA, USA): U87MG (HTB-14), U138MG (HTB-16), U118MG (HTB-15), LN18 (CRL-2610), LN229 (CRL-2611), T98G (CRL-1690). Upon arrival in the lab, frozen vials were revived as specified by ATCC, and cells were cultured according to standard conditions specified by ATCC. Approximately 10 million cells were harvested at approximately 70% confluency for RNA extraction as described in Chapter 2 Methods.

Normal human astrocytes (NHA) were purchased and cultured as specified (CC-2565, Lonza Inc, Allendale, NJ, USA). RNA sequencing was performed when NHA1 was at a population doubling of 3.5 (NHA1 B1) and 7 (NHA1 B2), and for NHA2 the population doubling was 3.4.

Chapter 4: Novel association of polymorphic genetic variants with predictors of outcome of catheter ablation in atrial fibrillation: new directions from a prospective study (DECAF)

Genome-wide association studies (GWAS) have allowed the genetics of disease to move beyond the Mendelian "one-gene/one phenotype" model [193]. However, many of the SNPs associated with affected phenotypes are located in non-coding regions of the genome, making the mechanistic effect of such SNPs difficult to determine. Any single SNP discovered in this fashion may only confer a small amount of risk, thus necessitating an understanding of how such SNPs affect gene expression during development and in an adult to cause a disease phenotype in an individual.

The ENCODE project and similar consortia have vastly increased the amount of genomic data available in a number of widely used cell lines, some immortal, some primary, as well as healthy human tissues [4, 194]. These data provide a window into the non-sequence based human genome, and allow us to see regulatory functions of DNA. Identifying the regulatory regions proximal to a non-coding but risk associated SNP allows for some elucidation of which regions of the genome may function synergistically [195]. Identifying which genes or transcripts are located on the same topologically associated domain (TAD) as the SNP in question decreases the search space for mechanistic effects, and limits which genes may be affected by a given SNP [196].

This work was published in Mohanty S[1], Hall AW[1] et al., Novel association of polymorphic genetic variants with predictors of outcome of catheter ablation in atrial fibrillation: new directions from a prospective study (DECAF). *J Interv Card Electrophysiol*. 2016 Jan;45(1):7-17. VRI, AWH and MM conceived the study and designed the experiments. AWH performed the experiments, collected the data and analyzed the data. AWH and MM wrote the manuscript. For this chapter, AWH reanalyzed the data.

INTRODUCTION

Atrial fibrillation (AF) is an intriguing model for understanding the effects of risk-associated SNPs for several reasons: it is relatively common, with 2% of the US population affected (~6 million individuals, [197]), it is relatively easy to diagnose an AF episode using an EKG, and there are numerous GWAS SNPs associated with AF. While AF is comparatively easy to identify during an episode, treatment remains a more difficult prospect. Catheter ablation is a relatively recent noninvasive surgical option for treatment of AF that involves mapping the electrical conductivity of the heart using catheters that enter the body through the femoral vein. Areas of aberrant electrical activity are ablated using high frequency radio waves to prevent the arrhythmia from occuring in the future. Several drugs exist for controlling arrhythmias, but all have a substantial risk of side effects. As such, catheter ablation is emerging as a superior option for long-term treatment of AF [198, 199].

However, catheter ablation has an initial success rate of 67-77% for the first procedure, with higher success rates only occurring after multiple ablations [200]. There are several reasons for this variability in response to catheter ablation: advanced age, metabolic syndrome, preexisting scarring in the left atrium, and triggers originating from non-pulmonary vein sites (non-PV), all contribute to adverse outcomes following catheter ablation [201-203]. The common comorbidies indicate that individuals can be genetically predisposed to AF. The first GWAS for AF was published in 2007 [204], and identified the SNP rs2200733, located on chromosome 4q25, as a risk allele associated with AF. Since then, a number of additional GWAS for AF have concluded, and identified a number of additional SNPs associated with AF risk [205-208]. However, these studies are comparisons of cases and controls, and do not allow for examination, in a population of individuals with AF, what SNPs may be determinants of response to treatment. Therefore, we designed the prospective study: Determining the association of chromosomal variants with non-PV triggers and ablation-outcome in AF (DECAF). Here, we examined the effects of 16 GWAS SNPs associated with AF on individual response to catheter ablation therapy, coupled with metadata on patient phenotypes such as left atrial scar and non-PV triggers.

RESULTS

371 patients with AF were genotyped and assessed for AF characteristics

400 consenting patients scheduled to undergo cardiac ablation were enrolled in the study, and blood was collected as described in the methods section. Genomic DNA isolation and genotyping was attempted for all 400 samples and 29 samples were excluded due to a lack of consensus during genotyping or low quality DNA. 371 samples were successfully genotyped and included in this analysis, and the characteristics for this population are described in **Table 4.1**. We collected data on the location of the AF trigger sites in each patient, and for many (but not all) patients, the presence of left atrial scarring. Non-PV triggers were detected in 40 (27 %) patients with paroxysmal AF (PAF), 123 (70 %) persistent AF, and 46 (92 %) long-standing persistent (LSP) patients. Information on presence of LA scar at the time of ablation procedure was available for 276 patients; LA scar was observed in 40 of 99 (41 %) PAF patients, 87 of 134 (65 %) persistent AF, and 28 of 43 (65 %) LSP-AF patients.

Age, year	64±11
Male	249 (67%)
AF Type	
Paroxysmal	146 (39%)
Persistent	175 (47%)
LSP	50 (13%)
Body Mass Index	30±6.5
Diabetes	82 (22%)
Hypertension	245 (66%)
Coronary Artery Disease	64 (17%)
CVA/TIA	22 (5.9%)
LVEF, %	57±9
Left atrial size, cm	4.59±0.74

Table 4.1: Characteristics of the study population (n = 371).

CVA cerebrovascular accident, TIA transient ischemic attack, LVEF left ventricular ejection fraction, LSP long-standing persistent AF.

SNPs associations with non-PV trigger present and absent

Three SNPs predicted some aspects of non-PV trigger risk (**Table 4.2**). The presence of non-PV triggers was associated with rs2106261 as well as rs6843082 for the additive model. rs2106261 is located in an intron of the gene *ZFHX3*, a transcriptional regulator involved in myocyte differentiation [209]. rs6843082 is associated with a risk of ischemic stroke [210], and is 150 kb upstream of *PITX2*, near the non-coding RNA LINC01438. *PITX2* delimits sinoatrial node formation in the developing heart through microRNA regulation [211]. For the SNP rs1448817, the group without non-PV triggers is substantially more likely to be homozygous for the minor allele genotype with the recessive model. rs1448817 is 80 kb upstream of *PITX2*.

SNP	Non-PV trigger present (n=209)			Non-PV trigger absent (n=162)						
(Minor allele [a], major allele [A])	AA	Aa	aa	AA	Aa	aa	additive	dominant	recessive	χ^2
rs11047543 (A, G)	163	42	4	121	39	2	0.5491	0.4616	0.7000	0.5910
rs13376333 (T, C)	96	95	18	70	73	19	0.4790	0.6738	0.3831	0.5940
rs1448817 (G, A)	88	102	19	62	73	27	0.2596	0.5223	0.0381	0.0895
rs16997168 (T, C)	122	73	14	92	60	9	0.9189	0.8324	0.8286	0.8428
rs17042171 (A, C)	134	64	11	91	60	11	0.1017	0.1341	0.6586	0.2965
rs17375901 (T, C)	186	23	0	147	14	1	0.7365	0.6094	0.4367	0.4001
rs2106261 (T, C)	124	73	12	102	53	7	0.0011	0.5203	0.6722	0.7086
rs251253 (T,C)	29	108	71	24	80	58	0.8843	0.8814	0.6669	0.8884
rs3807989 (A, G)	86	89	32	66	79	17	0.9182	0.9153	0.2160	0.3064
rs3825214 (G, A)	135	69	5	111	48	3	0.4476	0.4404	1.0000	0.7163
rs6599230 (T, C)	127	69	13	90	59	13	0.2652	0.3397	0.5423	0.5615
rs6666258 (C, G)	96	95	18	70	73	19	0.4790	0.6738	0.3831	0.5937
rs6843082 (G, A)	105	85	18	67	72	23	0.0443	0.0929	0.0978	0.1091
rs7164883 (G, A)	142	60	6	104	53	5	0.4503	0.4381	1.0000	0.7096
rs7193343 (T, C)	117	80	12	103	52	6	0.0980	0.1353	0.4681	0.2661
rs8192284 (C, A)	70	103	35	65	73	23	0.1720	0.1926	0.5652	0.4016

Table 4.2: SNP frequencies in non-PV trigger present and absent

See Methods for the definitions of additive, dominant, and recessive models.
χ^2, chi squared test. Values in red are less than 0.05.

Non-PV trigger in persistent and non-persistent AF

We identified one SNP associated with an absence of non-PV triggers in persistent AF (**Table 4.3**). Persistent AF is non-episodic, and occurs constantly, or nearly so. In the population surveyed here, non-PV triggers are more likely to be absent in persistent AF, and studies of catheter ablation in patients with persistent AF recommend routine encircling of the pulmonary veins to improve the probability of success [212]. Both rs7193343 and rs2106261 are located in the same intron of *ZFHX3*, separated by a stretch of 23 kb.

SNP	Non-PV Trigger Present (n=40)			Non-PV Trigger Absent (n=106)						
(Minor allele [a], major allele [A])	AA	Aa	aa	AA	Aa	aa	additive	dominant	recessive	χ^2
rs11047543 (A, G)	34	5	1	77	28	1	0.2060	0.1338	0.4743	0.1665
rs13376333 (T, C)	18	18	4	50	48	8	0.8580	0.8541	0.7364	0.8858
rs1448817 (G, A)	18	17	5	39	53	14	0.3660	0.4472	1.0000	0.6528
rs16997168 (T, C)	20	16	4	63	38	4	0.2097	0.3481	0.2165	0.2684
rs17042171 (A, C)	27	9	4	58	45	3	0.3755	0.1904	0.0896	0.0294
rs17375901 (T, C)	36	4	0	97	9	0	0.7519	0.7519	1.0000	0.7752
rs2106261 (T, C)	19	17	4	66	34	6	0.1073	0.1328	0.4621	0.2479
rs251253 (T,C)	7	22	11	16	53	37	0.6201	0.7997	0.4319	0.6933
rs3807989 (A, G)	20	15	5	46	50	10	0.5964	0.5764	0.5555	0.5611
rs3825214 (G, A)	29	11	0	75	30	1	0.8408	1.0000	1.0000	0.8202
rs6599230 (T, C)	25	13	2	67	33	6	1.0000	1.0000	1.0000	0.9789
rs6666258 (C, G)	18	18	4	50	48	8	0.8580	0.8541	0.7364	0.8858
rs6843082 (G, A)	24	12	4	46	50	10	0.1090	0.0947	1.0000	0.1576
rs7164883 (G, A)	28	10	2	71	32	3	1.0000	0.8433	0.6148	0.7012
rs7193343 (T, C)	17	19	4	65	35	5	0.0314	0.0406	0.2608	0.0921
rs8192284 (C, A)	14	22	4	43	51	11	0.5832	0.5714	1.0000	0.7768

Table 4.3: Non-PV trigger present versus absent in persistent AF

Several SNPs were significantly associated with the presence of non-PV triggers in the non-persistent AF group. Four SNPs were associated with the presence of non-PV triggers using the recessive model: rs13376333, rs1448817, rs17042171, and rs6666258. rs13376333 and rs6666258 are 87 bases apart, located in an intron of the potassium calcium-activated ion channel *KCNN3*, involved in atrial repolarization [207]. rs1448817 is upstream of *PITX2*, rs17042171 is in an intron of *ZFHX3*.

SNP	**Non-PV Trigger Present (n=169)**			**Non-PV Trigger Absent (n=56)**						
(Minor allele [a], major allele [A])	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**	**additive**	**dominant**	**recessive**	**χ^2**
rs11047543 (A, G)	129	37	3	44	11	1	0.8593	0.8553	1.0000	0.9383
rs13376333 (T, C)	78	77	14	20	25	11	0.0795	0.2137	0.0267	0.0512
rs1448817 (G, A)	70	85	14	23	20	13	0.5585	1.0000	0.0074	0.0077
rs16997168 (T, C)	102	57	10	29	22	5	0.2342	0.2770	0.5357	0.4788
rs17042171 (A, C)	107	55	7	33	15	8	0.2366	0.6337	0.0137	0.0292
rs17375901 (T, C)	150	19	0	50	5	1	0.8132	1.0000	0.2489	0.1989
rs2106261 (T, C)	105	56	8	36	19	1	0.6434	0.8736	0.4570	0.6210
rs251253 (T,C)	22	86	60	8	27	21	0.8272	0.8226	0.8727	0.9253
rs3807989 (A, G)	66	74	27	20	29	7	0.8780	0.6379	0.6680	0.5961
rs3825214 (G, A)	106	58	5	36	18	2	1.0000	0.8742	1.0000	0.9389
rs6599230 (T, C)	102	56	11	23	26	7	0.0081	0.0134	0.1618	0.0344
rs6666258 (C, G)	78	77	14	20	25	11	0.0795	0.2137	0.0267	0.0512
rs6843082 (G, A)	81	73	14	21	22	13	0.0452	0.2150	0.0075	0.0114
rs7164883 (G, A)	114	50	4	33	21	2	0.2092	0.2563	0.6414	0.4663
rs7193343 (T, C)	100	61	8	38	17	1	0.2157	0.2711	0.4570	0.4004
rs8192284 (C, A)	56	81	31	22	22	12	0.5383	0.4229	0.6956	0.5093

Table 4.4: Non-PV trigger present versus absent in non-persistent AF

Two additional SNPs were associated with the presence of non-PV triggers in non-persistent AF. rs6599230 was associated with non-PV triggers in the additive and dominant models, while rs6843082 was associated with with non-PV triggers in the additive and recessive models. rs6599230 is a synonymous variant in an exon of *SCN5A* – the codon change from GCA to GCG keeps the residue an alanine in either case. Across the non-persistent AF group, three SNPs were also significantly different between the non-PV trigger present and non-PV trigger absent group for non-persistent AF, using the chi-squared test of independence (rs17042171, rs6599230, rs6843082).

SNP association with LA scar risk

A single SNP is associated with an increase in the presence of scarring on the left atrium: rs3807989. This SNP is located in an intron of caveolin-1 (*CAV1*). For AF, the G allele is associated with risk, however this SNP is also associated with differences in the PR interval (a measure of atrial and atrial ventricular node conduction [213]) and these traits are associated with the A allele [214, 215]. This SNP also tests with a low value for the chi squared test of independence, indicating that the incidence of this polymorphism between the LA scar present and LA scar absent groups is truly different.

Discussion

Although our population of patients with AF was small, several SNPs were associated with the overall presence of non-PV triggers in AF, as well as in persistent and non-persistent AF, and with left atrial scarring. This is impressive given the total population surveyed is only 371 individuals. However, when multiple testing corrections are applied to this data, only a single SNP retains significance: rs2106261 for an overall association with the presence of non-PV triggers. In larger studies of AF that retain metadata, it may be interesting to test for population specific associations of the above SNPs that did not survive multiple testing corrections.

From a mechanistic perspective, none of the associated SNPs caused amino acid changes in protein-coding genes, so the mechanism of the risk effect is likely subtle.

SNP	LA Scar Present (n=156)			No LA Scar (n=120)						
(Minor allele [a], major allele [A])	AA	Aa	aa	AA	Aa	aa	additive	dominant	recessive	χ^2
rs11047543 (A, G)	121	33	1	90	28	2	0.4825	0.5677	0.5823	0.6515
rs13376333 (T, C)	70	70	16	54	53	13	1.0000	1.0000	1.0000	0.9855
rs1448817 (G, A)	60	79	17	46	59	15	1.0000	1.0000	0.7076	0.9139
rs16997168 (T, C)	96	52	8	65	41	13	0.1273	0.2677	0.1069	0.1698
rs17042171 (A, C)	95	53	8	71	42	7	0.8113	0.8048	0.7959	0.9434
rs17375901 (T, C)	137	18	1	110	10	0	0.3294	0.3291	1.0000	0.4565
rs2106261 (T, C)	90	58	8	69	43	8	0.9060	1.0000	0.6122	0.8560
rs251253 (T, C)	26	73	56	17	67	36	0.7440	0.6174	0.3048	0.3556
rs3807989 (A, G)	68	60	27	46	64	10	0.8116	0.3887	0.0326	0.0197
rs3825214 (G, A)	98	52	6	75	43	2	0.8074	1.0000	0.4725	0.5391
rs6599230 (T, C)	97	48	11	72	43	5	1.0000	0.8033	0.4372	0.4598
rs6666258 (C, G)	70	70	16	54	53	13	1.0000	1.0000	1.0000	0.9855
rs6843082 (G, A)	73	66	16	54	51	15	0.6424	0.8074	0.5713	0.9176
rs7164883 (G, A)	100	51	5	82	33	4	0.4556	0.4414	1.0000	0.6755
rs7193343 (T, C)	86	62	8	68	44	7	0.8133	0.8065	0.7949	0.8809
rs8192284 (C, A)	57	78	21	42	58	19	0.8070	0.8993	0.2226	0.8429

Table 4.5: SNP associations for left atrial scar present versus absent

rs13376333 and rs1448817 are both upstream of the homeodomain gene *PITX2*. Increased expression of *PITX2* is observed in cardiac myocytes from chronic AF patients, and this increased expression results in a decrease of the voltage-gated Ca^{++} current and increase of the slow delayed inward rectifying potassium channel current [216]. *PITX2* is specifically expressed in the left atrium, and a change in expression causes increased susceptibility to AF rhythms [217]. An increase in *WNT8* expression is implicated as the cause in *in vitro* models [218]. Mouse models with a knockout of *PITX2* results in an impairment of the associated genes *ZFHX3* (rs17042171, rs2106261) and *KCNN3* (rs13376333 and rs6666258). A knockdown of *ZFHX3* is associated with arrhythmogenesis and disregulation of calcium homeostasis in atrial myocytes [219].

This data captured one association with an increase in left atrial scarring (rs3807989), however this SNP has risk associations with both alleles across multiple GWAS studies [208, 213-215]. This makes a mechanism difficult to determine, but may indicate this locus is associated with modulation of a profibrotic response [220]. Fibrosis is frequently identified in the left atrium of long-term AF patients, and this is associated with activation of the EMT [221, 222]. The widespread changes in gene expression brought on by the existence of AF (likely associated with changes in *PITX2* expression) may cause changes in tissue composition that change the electrical conductivity in the left atrium, disrupting the pacemaking activity of the sinoatrial node and causing arrhythmia.

MATERIALS AND METHODS

This prospective single-center pilot study enrolled 400 consecutive AF patients undergoing catheter ablation at Texas Cardiac Arrhythmia Institute, St. David's Medical

Center, Austin, TX, from December 20, 2012, to August 30, 2013. Patients with bleeding disorders and inability to provide written informed consent were excluded from the study. Echocardiograms were performed on all patients before ablation to obtain measurements on LA diameter and left ventricular ejection fraction (LVEF). Institutional protocol for standard mapping and ablation procedure was followed by all physicians as described in detail in our earlier publications [223]. The study was approved by our institutional review boards and registered at clinicaltrials.gov (NCT01751607). It was conducted in collaboration with department of Molecular Biosciences, University of Texas (UT) at Austin.

Statistical Analysis

We used R software for all statistical analysis (version 3.1.1). We stratified the results by population into three groups: Non PV triggers present versus Non PV triggers absent, persistent AF versus non-persistent AF, and left atrial scar versus no left atrial scar. For each SNP, we used three models of calculating the contribution of risk to each: recessive, dominant, and additive. The recessive model only counts a genotype of "aa" or two recessive risk alleles as a success when calculating contingency tables. The dominant model presumes any presence of the risk allele contributes to disease phenotype and counts "Aa" and "aa" genotypes as a success, values being held even for both. The additive model counts "Aa" and "aa" as a success, but since the recessive genotype "aa" contains two copies of the risk allele, it is counted twice. For these three models, for all data included in this study, we used Fisher's Exact Test for count data to calculate *P*-values, and also calculated the Chi-squared test for independence for each group, as a validation that the frequency of a given SNP is truly different across the two compared groups.

Whole blood collection and storage

For each patient, 3 mL of whole blood was collected in sodium-heparin tubes. Tubes were labeled with a unique anonymous identifier and stored in a −80 °C freezer in the St. David's Medical Center main laboratory. Frozen samples were batch collected weekly and transported on dry ice to the Iyer lab at the University of Texas at Austin. After transport, sample tubes were stored in a −80 °C freezer until processing. SNP analysis of all DNA specimens was conducted simultaneously at the end of the study. The researchers at UT Austin responsible for genotyping were blinded about the clinical characteristics and identification of the study participants.

DNA purification protocol

Genomic DNA was isolated from the whole blood using the QIAamp DNA Blood Mini kit (51106, Qiagen, Venlo, Netherlands). Heparin was removed from the purified genomic DNA using Bacteroides Heparinase I (P0735L, New England Biolabs, Ipswitch, MA). Briefly, 24 units of enzyme were added to the genomic DNA and digestion was run in a heat block for 2 h at 30 °C. After completion of the digestion, the reaction was phenol-chloroform extracted twice using phenol/chloroform/isoamyl alcohol (25:24:1, 15593-031, Life Technologies, Carlsbad, CA) in phase-lock tubes (2302830 5-Prime, Hilden, Germany). The aqueous layer was precipitated in 100 % ethanol at −80 °C for 20 min, and then centrifuged at 4 °C at maximum speed for 15 min in a refrigerated microfuge. DNA pellets were washed once with 70 % ethanol and centrifuged at maximum speed at room temperature for 5 min. Pellets were dried of ethanol for 15 min on the bench top and resuspended in 20 µl of diethylpyrocarbonate (DEPC)-treated water (AM9906, Life Technologies, Carlsbad, CA). The concentration of each sample was quantitated using a NanoDrop ND-1000 (ThermoFisher Scientific, Waltham, MA). These concentrations were used to make 50 ng/µl stocks of the genomic DNA in 96-well plates.

SNP genotyping assays

Genotyping of the samples was performed using a custom made OpenArray loaded with TaqMan SNP genotyping assays. The following 16 SNPs were genotyped: rs16997168, rs1448817, rs17042171, rs6843082, rs13376333, rs2106261, rs17375901, rs3807989, rs11047543, rs7193343, rs3825214, rs7164883, rs251253, rs8192284, rs6666258, and rs6599230. To run each OpenArray, 2.5 µl of genomic DNA was added to a 384 deep-well plate and mixed with 2.5 µl of TaqMan OpenArray Genotyping Master Mix (4404846, Life Technologies, Carlsbad, CA). The 384 deep-well plates were then used by the OpenArray AccuFill System (4457243, Life Technologies, Carlsbad, CA) to fill each OpenArray plate. Samples were run in triplicate on each plate (48 individuals per plate) and OpenArrays were filled 2 at a time. After filling and sealing the OpenArray, the plates were run using a QuantStudio 12K Flex system (4471090, Life Technologies, Carlsbad, CA), utilizing the OpenArray specifications for PCR and plate reading after completion of PCR in the DNA Core Facility at UT Austin. After the PCR reactions were complete, the fluorescence values for the plate were analyzed using a plate reading utility. This software reads the fluorescence values for each well of the PCR plate and determines what alleles are present in each individual for each SNP.

The raw genotyping data were initially analyzed for consistency across individuals and SNP genotypes using TaqMan Genotyper Software (Applied Biosystems). While we began our analysis with whole blood samples from 400 individuals, we removed 29 of those samples from our final analysis for several reasons. If we were unable to obtain a consensus genotype after running the SNP assays in triplicate two times, we discarded that sample. We also discarded samples that displayed very low amplification signals for our assays and samples where the replicates did not cluster together after running triplicate experiments twice. We were able to retain 371 of the 400 samples (92.75 %) for our final analysis.

Chapter 5: Future directions

The era of genomics is firmly upon us. As the cost of high-throughput sequencing drops, the amount of sequencing data is increasing at an exponential pace [224, 225]. As a whole, the field has noticed this massive influx of data, and the potential issues involved in reproducibility of complex experiments, and quality control of sequence data [226, 227]. More data is not necessarily better, and care should be taken to ensure what questions we ask are feasible to answer with a given experimental design [228]. From here, the challenge is to develop tractable, patient relevant, modern models of disease, and couple these with best practices in data and experiment reproducibility.

Chromatin modifiers as drugable targets; Wnt and Hh in cancer

Bulk GBM tumors contain a signature reminiscent of glioma stem cells that suggests Wnt and Hh signaling are important aspects in GBM stemness. The signatures identified in Chapter 2 indicate that Wnt and Hh signaling pathways control enhancer and bivalent domains in GBM tumors in a subtype specific manner. As the Wnt and Hh pathways are important therapeutic targets, there are many small molecule inhibitors currently in development [229-231]. However, this raises an important question: since the main effectors of the Wnt and Hh signaling pathways (*WNT1* and *SHH*, respectively) in GBM are bivalent, how can these pathways be targeted pharmacologically? Clearly the repressive H3K27me3 signal must be removed before these genes can be expressed (and targeted), so a multi-layered strategy is required.

First, application of a chromatin modifier (such as the H3K27 demethylases

JMJD3 [232] or *UTX* [233]) to remove the repressive H3K27me3 marks proximal to bivalent Wnt and Hh pathway effectors, followed by treatment with Wnt and Hh pathway inhibitors to prevent cell proliferation [234]. Chromatin modifiers are altering the treatment landscape in GBM, with multiple phase II clinical trials for HDAC inhibitors, so a complicated treatment regimen today is not necessarily impossible [235]. This type of combination therapy would be complex, and requires extensive preclinical validation in a model of GBM that recapitulates essential clinical features to ensure that the drugs involved have a robust response in GBM lesions, which are notoriously heterogeneous in nature.

Better *in vitro* models of disease

The above example indicates that the biomedical field has outgrown many old *in vitro* model systems. The era of relying on a small number of immortal and highly domesticated cell lines to define genetically complex diseases is drawing to a close. The field requires models that directly reflect essential characteristics of the disease in question. Emerging techniques in genetic engineering and cell culture offer compelling paths to this end. Models of stem cells for specific cancers (cancer stem cells, CSCs) are becoming more widespread. Chapter 3 brought up the concept of glioma stem cells (GSCs) as a model of the multipotent self-renewing cells present in GBM tumors. GSCs recapitulate the essential characteristics of GBM more effectively than established cell line models. However, improvements in cell culture techniques mean that existing cell lines (and the data derived from these lines) can still be useful [236].

Improved CSC culture is one aspect of building more relevant models of disease; induced pluripotent stem cells (iPSCs) are important paradigm as well. Subtle changes in somatic cells over time can result in a cancer, but this takes years. Taking a skin tissue sample from a patient with cancer or a genetic disease allows for an understanding of how an individual genetic background influences disease progression or response to therapy. Coupled with organoid culture, tissue-specific drug responses can be measured. This has been done to great effect in the intestinal organoids of cystic fibrosis patients [237], in cerebellar organoids to model CNS development and microcephaly [238], and in patient-derived cardiomyocytes [239].

The existence of novel tools doesn't mean that studying primary tissue is no longer necessary, only that it can be recapitulated by using less invasive methods. Examining gene expression and chromatin states in primary tissue provides the most accurate (if most complex) view of transcriptional regulation across the genome. Cell culture based models should be checked against primary tissue examples to ensure that the model matches the phenotype in uncultured material. Cell culture based models can assist in study of primary tissue as well – understanding which cell surface markers define tissue-specific stem cells (or other specific populations) can assist in their isolation from primary tissue. Methods such as flow cytometry can assist by purifying specific sub-populations of cells from primary tissue for further interrogation [240, 241]. Coupling flow sorting with chromatin immunoprecipitation and RNA sequencing protocols optimized for low numbers of input cells [242, 243] will allow for more specific expression, and chromatin based definitions of specific populations of cells from healthy and diseased tissue.

Noncoding genetic polymorphisms and chromatin topologies

The above examples deal with gross abnormalities in chromatin structure that drastically affect the underlying functionality of the cell(s) and create a clearly neoplastic or diseased state. Most genetic effects are orders of magnitude more subtle than cancer. Genetic variation among individuals generally works on a less severe level than megabase-scale losses or gains of genetic material [244]. The SNPs genotyped in Chapter 4 are subtle, but still cause a small detectable effect in phenotype in some subpopulations. Most human genetic variation works in this manner, and while GWAS studies establish genomic regions of interest, they do not put forward a probable model of effect.

Despite these difficult to detect outcomes, studies in chromatin topology are revealing potential mechanisms of action [245, 246]. In certain cases, the presence of a single nucleotide polymorphism can change a long-range chromosomal interaction [247]. The use of iPSCs to directly model disease in the genetic background of any specific patient is bringing clarity to aspects of disease phenotypes as diverse as macular degeneration, spinal muscular atrophy, cystic fibrosis, and cardiovascular disease [239, 248-250]. Coupling these specific models with directed genome editing using CRISPR-Cas9 type systems is the first step in targeted gene therapies to prevent disease states from occurring in genetically predisposed individuals [251-253].

References

1. Venter, J.C., et al., *The sequence of the human genome*. Science, 2001. **291**(5507): p. 1304-51.
2. Edwards, S.L., et al., *Beyond GWASs: illuminating the dark road from association to function*. Am J Hum Genet, 2013. **93**(5): p. 779-97.
3. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression*. Nature, 2009. **459**(7243): p. 108-12.
4. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. Nature, 2007. **447**(7146): p. 799-816.
5. Domcke, S., et al., *Evaluating cell lines as tumour models by comparison of genomic profiles*. Nat Commun, 2013. **4**: p. 2126.
6. Vincent, K.M., S.D. Findlay, and L.M. Postovit, *Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles*. Breast Cancer Res, 2015. **17**: p. 114.
7. Branco, M.R. and A. Pombo, *Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations*. PLoS Biol, 2006. **4**(5): p. e138.
8. Ugarte, F., et al., *Progressive Chromatin Condensation and H3K9 Methylation Regulate the Differentiation of Embryonic and Hematopoietic Stem Cells*. Stem Cell Reports, 2015. **5**(5): p. 728-40.
9. Azuara, V., et al., *Chromatin signatures of pluripotent cell lines*. Nat Cell Biol, 2006. **8**(5): p. 532-8.
10. Fussner, E., et al., *Constitutive heterochromatin reorganization during somatic cell reprogramming*. EMBO J, 2011. **30**(9): p. 1778-89.
11. Peters, A.H., et al., *Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability*. Cell, 2001. **107**(3): p. 323-37.
12. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome*. Cell, 2007. **129**(4): p. 823-37.
13. Dileep, V., et al., *Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program*. Genome Res, 2015. **25**(8): p. 1104-13.
14. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions*. Nature, 2012. **485**(7398): p. 376-80.
15. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping*. Cell, 2014. **159**(7): p. 1665-80.
16. Olins, D.E. and A.L. Olins, *Chromatin history: our view from the bridge*. Nat Rev Mol Cell Biol, 2003. **4**(10): p. 809-14.
17. Zheng, C. and J.J. Hayes, *Structures and interactions of the core histone tail domains*. Biopolymers, 2003. **68**(4): p. 539-46.

18. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications*. Cell Res, 2011. **21**(3): p. 381-95.

19. Hake, S.B. and C.D. Allis, *Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis"*. Proc Natl Acad Sci U S A, 2006. **103**(17): p. 6428-35.

20. Paull, T.T., et al., *A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage*. Curr Biol, 2000. **10**(15): p. 886-95.

21. Glaser, S., et al., *Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development*. Development, 2006. **133**(8): p. 1423-32.

22. Denissov, S., et al., *Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant*. Development, 2014. **141**(3): p. 526-37.

23. Hon, G.C., R.D. Hawkins, and B. Ren, *Predictive chromatin signatures in the mammalian genome*. Hum Mol Genet, 2009. **18**(R2): p. R195-201.

24. Wang, Y., X. Li, and H. Hu, *H3K4me2 reliably defines transcription factor binding regions in different cells*. Genomics, 2014. **103**(2-3): p. 222-8.

25. Pekowska, A., et al., *A unique H3K4me2 profile marks tissue-specific gene regulation*. Genome Res, 2010. **20**(11): p. 1493-502.

26. Lee, J.E., et al., *H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation*. Elife, 2013. **2**: p. e01503.

27. Yang, X.J. and E. Seto, *HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention*. Oncogene, 2007. **26**(37): p. 5310-8.

28. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans*. Nature, 2011. **470**(7333): p. 279-83.

29. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state*. Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.

30. Kratz, A., et al., *Core promoter structure and genomic context reflect histone 3 lysine 9 acetylation patterns*. BMC Genomics, 2010. **11**: p. 257.

31. Karmodiya, K., et al., *H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells*. BMC Genomics, 2012. **13**: p. 424.

32. Grant, P.A., et al., *Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex*. Genes Dev, 1997. **11**(13): p. 1640-50.

33. Yang, X.J. and E. Seto, *The Rpd3/Hda1 family of lysine deacetylases: from bacteria and yeast to mice and men*. Nat Rev Mol Cell Biol, 2008. **9**(3): p. 206-18.

34. Fuks, F., et al., *The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase*. Nucleic Acids Res, 2003. **31**(9): p. 2305-12.

35. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences*. Nature, 2009. **462**(7271): p. 315-22.

36. Wu, X., J.V. Johansen, and K. Helin, *Fbxl10/Kdm2b recruits polycomb repressive complex 1 to CpG islands and regulates H2A ubiquitylation*. Mol Cell, 2013. **49**(6): p. 1134-46.

37. Blackledge, N.P., et al., *Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation*. Cell, 2014. **157**(6): p. 1445-59.

38. Chaturvedi, C.P., et al., *Maintenance of gene silencing by the coordinate action of the H3K9 methyltransferase G9a/KMT1C and the H3K4 demethylase Jarid1a/KDM5A*. Proc Natl Acad Sci U S A, 2012. **109**(46): p. 18845-50.

39. Shi, L., et al., *Histone demethylase JMJD2B coordinates H3K4/H3K9 methylation and promotes hormonally responsive breast carcinogenesis*. Proc Natl Acad Sci U S A, 2011. **108**(18): p. 7541-6.

40. Kim, J.Y., et al., *KDM3B is the H3K9 demethylase involved in transcriptional activation of lmo2 in leukemia*. Mol Cell Biol, 2012. **32**(14): p. 2917-33.

41. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing*. Nat Methods, 2007. **4**(8): p. 651-7.

42. Park, D., et al., *Widespread misinterpretable ChIP-seq bias in yeast*. PLoS One, 2013. **8**(12): p. e83506.

43. Carter, D., et al., *Long-range chromatin regulatory interactions in vivo*. Nat Genet, 2002. **32**(4): p. 623-6.

44. van Berkum, N.L., et al., *Hi-C: a method to study the three-dimensional architecture of genomes*. J Vis Exp, 2010(39).

45. van de Werken, H.J., et al., *4C technology: protocols and data analysis*. Methods Enzymol, 2012. **513**: p. 89-112.

46. Ostrom, Q.T., et al., *CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012*. Neuro Oncol, 2015. **17 Suppl 4**: p. iv1-iv62.

47. Delgado-Lopez, P.D. and E.M. Corrales-Garcia, *Survival in glioblastoma: a review on the impact of treatment modalities*. Clin Transl Oncol, 2016. **18**(11): p. 1062-1071.

48. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1*. Cancer Cell, 2010. **17**(1): p. 98-110.

49. Nagarajan, R.P., et al., *Recurrent epimutations activate gene body promoters in primary glioblastoma*. Genome Res, 2014. **24**(5): p. 761-74.

50. Lewis, P.W., et al., *Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma*. Science, 2013. **340**(6134): p. 857-61.

51. Chan, K.M., et al., *The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression*. Genes Dev, 2013. **27**(9): p. 985-90.

52. Cenci, T., et al., *Prognostic relevance of c-Myc and BMI1 expression in patients with glioblastoma*. Am J Clin Pathol, 2012. **138**(3): p. 390-6.

53. Lucio-Eterovic, A.K., et al., *Differential expression of 12 histone deacetylase (HDAC) genes in astrocytomas and normal brain tissue: class II and IV are hypoexpressed in glioblastomas*. BMC Cancer, 2008. **8**: p. 243.
54. Lin, C.Y., et al., *Active medulloblastoma enhancers reveal subgroup-specific cellular origins*. Nature, 2016. **530**(7588): p. 57-62.
55. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization*. Nat Methods, 2012. **9**(3): p. 215-6.
56. Sottoriva, A., et al., *Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics*. Proc Natl Acad Sci U S A, 2013. **110**(10): p. 4009-14.
57. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project*. Genome Res, 2012. **22**(9): p. 1760-74.
58. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways*. Nature, 2008. **455**(7216): p. 1061-8.
59. Huse, J.T., H.S. Phillips, and C.W. Brennan, *Molecular subclassification of diffuse gliomas: seeing order in the chaos*. Glia, 2011. **59**(8): p. 1190-9.
60. Goodenberger, M.L. and R.B. Jenkins, *Genetics of adult glioma*. Cancer Genet, 2012. **205**(12): p. 613-21.
61. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome*. Nat Genet, 2007. **39**(3): p. 311-8.
62. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells*. Cell, 2006. **125**(2): p. 315-26.
63. Lin, B., et al., *Global analysis of H3K4me3 and H3K27me3 profiles in glioblastoma stem cells and identification of SLC17A7 as a bivalent tumor suppressor gene*. Oncotarget, 2015. **6**(7): p. 5369-81.
64. Yoo, S. and M.C. Bieda, *Differences among brain tumor stem cell types and fetal neural stem cells in focal regions of histone modifications and DNA methylation, broad regions of modifications, and bivalent promoters*. BMC Genomics, 2014. **15**: p. 724.
65. Stroud, H., et al., *5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells*. Genome Biol, 2011. **12**(6): p. R54.
66. Szulwach, K.E., et al., *Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells*. PLoS Genet, 2011. **7**(6): p. e1002154.
67. Johnson, K.C., et al., *5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients*. Nat Commun, 2016. **7**: p. 13177.
68. Wang, H., et al., *Widespread plasticity in CTCF occupancy linked to DNA methylation*. Genome Res, 2012. **22**(9): p. 1680-8.
69. Ashoor, H., et al., *DENdb: database of integrated human enhancers*. Database (Oxford), 2015. **2015**.

70. Bailey, T.L., et al., *The MEME Suite*. Nucleic Acids Res, 2015. **43**(W1): p. W39-49.
71. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nat Protoc, 2009. **4**(1): p. 44-57.
72. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life*. Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
73. Lee, I., et al., *Prioritizing candidate disease genes by network-based boosting of genome-wide association data*. Genome Res, 2011. **21**(7): p. 1109-21.
74. Cline, M.S., et al., *Integration of biological networks and gene expression data using Cytoscape*. Nat Protoc, 2007. **2**(10): p. 2366-82.
75. Nakano, I., *Stem cell signature in glioblastoma: therapeutic development for a moving target*. J Neurosurg, 2015. **122**(2): p. 324-30.
76. Cheng, L., et al., *Glioblastoma stem cells generate vascular pericytes to support vessel function and tumor growth*. Cell, 2013. **153**(1): p. 139-52.
77. Kim, S.H., et al., *The LIM-only transcription factor LMO2 determines tumorigenic and angiogenic traits in glioma stem cells*. Cell Death Differ, 2015. **22**(9): p. 1517-25.
78. Binder, Z.A., et al., *Podocalyxin-like protein is expressed in glioblastoma multiforme stem-like cells and is associated with poor outcome*. PLoS One, 2013. **8**(10): p. e75945.
79. Kou, Y.B., et al., *Knockdown of MMP11 inhibits proliferation and invasion of gastric cancer cells*. Int J Immunopathol Pharmacol, 2013. **26**(2): p. 361-70.
80. Zhou, W., et al., *Up-regulation of S100A16 expression promotes epithelial-mesenchymal transition via Notch1 pathway in breast cancer*. J Biomed Sci, 2014. **21**: p. 97.
81. Chandran, U.R., et al., *Gene expression profiling distinguishes proneural glioma stem cells from mesenchymal glioma stem cells*. Genom Data, 2015. **5**: p. 333-336.
82. Warren, A.J., et al., *The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development*. Cell, 1994. **78**(1): p. 45-57.
83. Yamada, Y., et al., *The oncogenic LIM-only transcription factor Lmo2 regulates angiogenesis but not vasculogenesis in mice*. Proc Natl Acad Sci U S A, 2000. **97**(1): p. 320-4.
84. Westcott, J.M., et al., *An epigenetically distinct breast cancer cell subpopulation promotes collective invasion*. J Clin Invest, 2015. **125**(5): p. 1927-43.
85. Zhu, C.Q., et al., *Integrin alpha 11 regulates IGF2 expression in fibroblasts to enhance tumorigenicity of human non-small-cell lung cancer cells*. Proc Natl Acad Sci U S A, 2007. **104**(28): p. 11754-9.
86. Mure, H., et al., *Akt2 and Akt3 play a pivotal role in malignant gliomas*. Neuro Oncol, 2010. **12**(3): p. 221-32.

87. Joy, A., et al., *The role of AKT isoforms in glioblastoma: AKT3 delays tumor progression*. J Neurooncol, 2016. **130**(1): p. 43-52.

88. Patel, V.N., et al., *Network signatures of survival in glioblastoma multiforme*. PLoS Comput Biol, 2013. **9**(9): p. e1003237.

89. Brauer, K., L. Werner, and L. Leibnitz, *Perineuronal nets of glia*. J Hirnforsch, 1982. **23**(6): p. 701-8.

90. Chiquet-Ehrismann, R. and R.P. Tucker, *Tenascins and the importance of adhesion modulation*. Cold Spring Harb Perspect Biol, 2011. **3**(5).

91. Hargus, G., et al., *Tenascin-R promotes neuronal differentiation of embryonic stem cells and recruitment of host-derived neural precursor cells after excitotoxic lesion of the mouse striatum*. Stem Cells, 2008. **26**(8): p. 1973-84.

92. Wang, Z., et al., *MiR-30a-5p is induced by Wnt/beta-catenin pathway and promotes glioma cell invasion by repressing NCAM*. Biochem Biophys Res Commun, 2015. **465**(3): p. 374-80.

93. Jang, C., et al., *Calsenilin regulates presenilin 1/gamma-secretase-mediated N-cadherin epsilon-cleavage and beta-catenin signaling*. FASEB J, 2011. **25**(12): p. 4174-83.

94. Ramis-Conde, I., et al., *Multi-scale modelling of cancer cell intravasation: the role of cadherins in metastasis*. Phys Biol, 2009. **6**(1): p. 016008.

95. Giridharan, S.S., et al., *Differential regulation of actin microfilaments by human MICAL proteins*. J Cell Sci, 2012. **125**(Pt 3): p. 614-24.

96. Mariotti, S., et al., *MICAL2 is a novel human cancer gene controlling mesenchymal to epithelial transition involved in cancer growth and invasion*. Oncotarget, 2016. **7**(2): p. 1808-25.

97. Marie, S.K., et al., *Stathmin involvement in the maternal embryonic leucine zipper kinase pathway in glioblastoma*. Proteome Sci, 2016. **14**: p. 6.

98. Wang, R., et al., *LRRC4 inhibits the proliferation of human glioma cells by modulating the expression of STMN1 and microtubule polymerization*. J Cell Biochem, 2011. **112**(12): p. 3621-9.

99. Betapudi, V., *Myosin II motor proteins with different functions determine the fate of lamellipodia extension during cell spreading*. PLoS One, 2010. **5**(1): p. e8560.

100. Cai, L., et al., *Nonmuscle myosin-dependent synthesis of type I collagen*. J Mol Biol, 2010. **401**(4): p. 564-78.

101. Thomas, D.G., et al., *Non-muscle myosin IIB is critical for nuclear translocation during 3D invasion*. J Cell Biol, 2015. **210**(4): p. 583-94.

102. Cuddapah, V.A., et al., *A neurocentric perspective on glioma invasion*. Nat Rev Neurosci, 2014. **15**(7): p. 455-65.

103. Liu, Y., et al., *Vascular gene expression patterns are conserved in primary and metastatic brain tumors*. J Neurooncol, 2010. **99**(1): p. 13-24.

104. Rocnik, E.F., et al., *The novel SPARC family member SMOC-2 potentiates angiogenic growth factor activity*. J Biol Chem, 2006. **281**(32): p. 22855-64.

105. Guezguez, A., et al., *Modulation of stemness in a human normal intestinal epithelial crypt cell line by activation of the WNT signaling pathway*. Exp Cell Res, 2014. **322**(2): p. 355-64.

106. Shvab, A., et al., *Induction of the intestinal stem cell signature gene SMOC-2 is required for L1-mediated colon cancer progression*. Oncogene, 2016. **35**(5): p. 549-57.

107. Ostermann, G., et al., *JAM-1 is a ligand of the beta(2) integrin LFA-1 involved in transendothelial migration of leukocytes*. Nat Immunol, 2002. **3**(2): p. 151-8.

108. Rajaraman, P., et al., *Common variation in genes related to innate immunity and risk of adult glioma*. Cancer Epidemiol Biomarkers Prev, 2009. **18**(5): p. 1651-8.

109. Gao, X., et al., *LEF1 regulates glioblastoma cell proliferation, migration, invasion, and cancer stem-like cell self-renewal*. Tumour Biol, 2014. **35**(11): p. 11505-11.

110. Rheinbay, E., et al., *An aberrant transcription factor network essential for Wnt signaling and stem cell maintenance in glioblastoma*. Cell Rep, 2013. **3**(5): p. 1567-79.

111. Liu, B., et al., *Overexpressed FOXC2 in ovarian cancer enhances the epithelial-to-mesenchymal transition and invasion of ovarian cancer cells*. Oncol Rep, 2014. **31**(6): p. 2545-54.

112. Cai, J., et al., *FOXF2 suppresses the FOXC2-mediated epithelial-mesenchymal transition and multidrug resistance of basal-like breast cancer*. Cancer Lett, 2015. **367**(2): p. 129-37.

113. Li, W., et al., *FOXC2 often overexpressed in glioblastoma enhances proliferation and invasion in glioblastoma cells*. Oncol Res, 2013. **21**(2): p. 111-20.

114. Lin, Y., et al., *Inorganic phosphate induces cancer cell mediated angiogenesis dependent on forkhead box protein C2 (FOXC2) regulated osteopontin expression*. Mol Carcinog, 2015. **54**(9): p. 926-34.

115. Chen, J., et al., *HoxB3 promotes prostate cancer cell progression by transactivating CDCA3*. Cancer Lett, 2013. **330**(2): p. 217-24.

116. Fu, H., et al., *miR-375 inhibits cancer stem cell phenotype and tamoxifen resistance by degrading HOXB3 in human ER-positive breast cancer*. Oncol Rep, 2017. **37**(2): p. 1093-1099.

117. Rossi, M., et al., *beta-catenin and Gli1 are prognostic markers in glioblastoma*. Cancer Biol Ther, 2011. **11**(8): p. 753-61.

118. Haupt, Y., et al., *Mdm2 promotes the rapid degradation of p53*. Nature, 1997. **387**(6630): p. 296-9.

119. Mizuno, T., et al., *Neuronal adhesion molecule telencephalin induces rapid cell spreading of microglia*. Brain Res, 1999. **849**(1-2): p. 58-66.

120. Tian, L., et al., *Binding of T lymphocytes to hippocampal neurons through ICAM-5 (telencephalin) and characterization of its interaction with the leukocyte integrin CD11a/CD18*. Eur J Immunol, 2000. **30**(3): p. 810-8.

121. Ashktorab, H., et al., *Toward a comprehensive and systematic methylome signature in colorectal cancers*. Epigenetics, 2013. **8**(8): p. 807-15.

122. Liu, L., et al., *Slit2 and Robo1 expression as biomarkers for assessing prognosis in brain glioma patients*. Surg Oncol, 2016. **25**(4): p. 405-410.

123. Mi, S., et al., *LINGO-1 negatively regulates myelination by oligodendrocytes*. Nat Neurosci, 2005. **8**(6): p. 745-51.

124. Yin, W. and B. Hu, *Knockdown of Lingo1b protein promotes myelination and oligodendrocyte differentiation in zebrafish*. Exp Neurol, 2014. **251**: p. 72-83.

125. Loov, C., et al., *Neutralization of LINGO-1 during in vitro differentiation of neural stem cells results in proliferation of immature neurons*. PLoS One, 2012. **7**(1): p. e29771.

126. Zhang, Z., et al., *LINGO-1 receptor promotes neuronal apoptosis by inhibiting WNK3 kinase activity*. J Biol Chem, 2013. **288**(17): p. 12152-60.

127. Ligon, K.L., et al., *Olig2-regulated lineage-restricted pathway controls replication competence in neural stem cells and malignant glioma*. Neuron, 2007. **53**(4): p. 503-17.

128. Kupp, R., et al., *Lineage-Restricted OLIG2-RTK Signaling Governs the Molecular Subtype of Glioma Stem-like Cells*. Cell Rep, 2016. **16**(11): p. 2838-45.

129. Lu, F., et al., *Olig2-Dependent Reciprocal Shift in PDGF and EGF Receptor Signaling Regulates Tumor Phenotype and Mitotic Growth in Malignant Glioma*. Cancer Cell, 2016. **29**(5): p. 669-83.

130. Chen, D., et al., *Better prognosis of patients with glioma expressing FGF2-dependent PDGFRA irrespective of morphological diagnosis*. PLoS One, 2013. **8**(4): p. e61556.

131. Liu, K.W., et al., *SHP-2/PTPN11 mediates gliomagenesis driven by PDGFRA and INK4A/ARF aberrations in mice and humans*. J Clin Invest, 2011. **121**(3): p. 905-17.

132. Todo, T., et al., *Expression and growth stimulatory effect of fibroblast growth factor 9 in human brain tumors*. Neurosurgery, 1998. **43**(2): p. 337-46.

133. Schmid, S., et al., *Wnt and hedgehog gene pathway expression in serous ovarian cancer*. Int J Gynecol Cancer, 2011. **21**(6): p. 975-80.

134. Nicolis, S.K., *Cancer stem cells and "stemness" genes in neuro-oncology*. Neurobiol Dis, 2007. **25**(2): p. 217-29.

135. Feinberg, A.P., M.A. Koldobskiy, and A. Gondor, *Epigenetic modulators, modifiers and mediators in cancer aetiology and progression*. Nat Rev Genet, 2016. **17**(5): p. 284-99.

136. Murat, A., et al., *Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma*. J Clin Oncol, 2008. **26**(18): p. 3015-24.

137. Tabuse, M., et al., *Functional analysis of HOXD9 in human gliomas and glioma cancer stem cells*. Mol Cancer, 2011. **10**: p. 60.

138. Kurscheid, S., et al., *Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma*. Genome Biol, 2015. **16**: p. 16.

139. Hu, B., et al., *Epigenetic Activation of WNT5A Drives Glioblastoma Stem Cell Differentiation and Invasive Growth*. Cell, 2016. **167**(5): p. 1281-1295 e18.
140. Takebe, N., et al., *Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update*. Nat Rev Clin Oncol, 2015. **12**(8): p. 445-64.
141. Suwala, A.K., et al., *Clipping the Wings of Glioblastoma: Modulation of WNT as a Novel Therapeutic Strategy*. J Neuropathol Exp Neurol, 2016. **75**(5): p. 388-96.
142. Zhang, J., X.J. Tian, and J. Xing, *Signal Transduction Pathways of EMT Induced by TGF-beta, SHH, and WNT and Their Crosstalks*. J Clin Med, 2016. **5**(4).
143. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-60.
144. Marinov, G.K., et al., *Large-scale quality analysis of published ChIP-seq data*. G3 (Bethesda), 2014. **4**(2): p. 209-23.
145. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS)*. Genome Biol, 2008. **9**(9): p. R137.
146. Boyle, A.P., et al., *High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells*. Genome Res, 2011. **21**(3): p. 456-64.
147. Rosenbloom, K.R., et al., *The UCSC Genome Browser database: 2015 update*. Nucleic Acids Res, 2015. **43**(Database issue): p. D670-81.
148. Kent, W.J., et al., *The human genome browser at UCSC*. Genome Res, 2002. **12**(6): p. 996-1006.
149. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-2.
150. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. EMBnet.journal, 2011. **17**(1): p. 10-12.
151. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions*. Genome Biol, 2013. **14**(4): p. R36.
152. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq*. Nat Biotechnol, 2013. **31**(1): p. 46-53.
153. van Staveren, W.C., et al., *Human cancer cell lines: Experimental models for cancer cells in situ? For cancer stem cells?* Biochim Biophys Acta, 2009. **1795**(2): p. 92-103.
154. Westermark, B., J. Ponten, and R. Hugosson, *Determinants for the establishment of permanent tissue culture lines from human gliomas*. Acta Pathol Microbiol Scand A, 1973. **81**(6): p. 791-805.
155. Ponten, J., B. Westermark, and R. Hugosson, *Regulation of proliferation and movement of human glialike cells in culture*. Exp Cell Res, 1969. **58**(2): p. 393-400.
156. Lee, J., et al., *Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines*. Cancer Cell, 2006. **9**(5): p. 391-403.
157. Kang, S.K., J.B. Park, and S.H. Cha, *Multipotent, dedifferentiated cancer stem-like cells from brain gliomas*. Stem Cells Dev, 2006. **15**(3): p. 423-35.

158. Mao, P., et al., *Mesenchymal glioma stem cells are maintained by activated glycolytic metabolism involving aldehyde dehydrogenase 1A3*. Proc Natl Acad Sci U S A, 2013. **110**(21): p. 8644-9.
159. Bhat, K.P., et al., *Mesenchymal differentiation mediated by NF-kappaB promotes radiation resistance in glioblastoma*. Cancer Cell, 2013. **24**(3): p. 331-46.
160. Hossain, A., et al., *Mesenchymal Stem Cells Isolated From Human Gliomas Increase Proliferation and Maintain Stemness of Glioma Stem Cells Through the IL-6/gp130/STAT3 Pathway*. Stem Cells, 2015. **33**(8): p. 2400-15.
161. Jensen, J.B. and M. Parmar, *Strengths and limitations of the neurosphere culture system*. Mol Neurobiol, 2006. **34**(3): p. 153-61.
162. Azari, H., et al., *Isolation and expansion of human glioblastoma multiforme tumor cells using the neurosphere assay*. J Vis Exp, 2011(56): p. e3633.
163. Rahman, M., et al., *Neurosphere and adherent culture conditions are equivalent for malignant glioma stem cell lines*. Anat Cell Biol, 2015. **48**(1): p. 25-35.
164. Schnabel, M., et al., *Dedifferentiation-associated changes in morphology and gene expression in primary human articular chondrocytes in cell culture*. Osteoarthritis Cartilage, 2002. **10**(1): p. 62-70.
165. Collins, F.S. and L.A. Tabak, *Policy: NIH plans to enhance reproducibility*. Nature, 2014. **505**(7485): p. 612-3.
166. Saraste, M. and M. Hyvonen, *Pleckstrin homology domains: a fact file*. Curr Opin Struct Biol, 1995. **5**(3): p. 403-8.
167. Mayer, B.J., *SH3 domains: complexity in moderation*. J Cell Sci, 2001. **114**(Pt 7): p. 1253-63.
168. Bady, P., et al., *DNA fingerprinting of glioma cell lines and considerations on similarity measurements*. Neuro Oncol, 2012. **14**(6): p. 701-11.
169. Li, D. and R. Roberts, *WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases*. Cell Mol Life Sci, 2001. **58**(14): p. 2085-97.
170. Liu, F., et al., *EGFR Mutation Promotes Glioblastoma through Epigenome and Transcription Factor Network Remodeling*. Mol Cell, 2015. **60**(2): p. 307-18.
171. Amessou, M., et al., *Spatio-temporal regulation of EGFR signaling by the Eps15 homology domain-containing protein 3 (EHD3)*. Oncotarget, 2016. **7**(48): p. 79203-79216.
172. Dai, X., Z. Liu, and S. Zhang, *Over-expression of EPS15 is a favorable prognostic factor in breast cancer*. Mol Biosyst, 2015. **11**(11): p. 2978-85.
173. Li, M.Y., et al., *Protein tyrosine phosphatase PTPN3 inhibits lung cancer cell proliferation and migration by promoting EGFR endocytic degradation*. Oncogene, 2015. **34**(29): p. 3791-803.
174. Han, X., et al., *The role of Src family kinases in growth and migration of glioma stem cells*. Int J Oncol, 2014. **45**(1): p. 302-10.
175. Lu, K.V., et al., *Fyn and SRC are effectors of oncogenic epidermal growth factor receptor signaling in glioblastoma patients*. Cancer Res, 2009. **69**(17): p. 6889-98.

176. Zhou, P., et al., *CD151-alpha3beta1 integrin complexes are prognostic markers of glioblastoma and cooperate with EGFR to drive tumor cell motility and invasion*. Oncotarget, 2015. **6**(30): p. 29675-93.
177. Aoki, H., et al., *Phosphorylated Pak1 level in the cytoplasm correlates with shorter survival time in patients with glioblastoma*. Clin Cancer Res, 2007. **13**(22 Pt 1): p. 6603-9.
178. Panicker, S.P., et al., *p300- and Myc-mediated regulation of glioblastoma multiforme cell differentiation*. Oncotarget, 2010. **1**(4): p. 289-303.
179. Alrfaei, B.M., R. Vemuganti, and J.S. Kuo, *microRNA-100 targets SMRT/NCOR2, reduces proliferation, and improves survival in glioblastoma animal models*. PLoS One, 2013. **8**(11): p. e80865.
180. van Agthoven, T., et al., *CITED2 and NCOR2 in anti-oestrogen resistance and progression of breast cancer*. Br J Cancer, 2009. **101**(11): p. 1824-32.
181. Dali-Youcef, N., et al., *Gene expression mapping of histone deacetylases and co-factors, and correlation with survival time and 1H-HRMAS metabolomic profile in human gliomas*. Sci Rep, 2015. **5**: p. 9087.
182. Chen, J., et al., *HDAC5 promotes osteosarcoma progression by upregulation of Twist 1 expression*. Tumour Biol, 2014. **35**(2): p. 1383-7.
183. He, P., et al., *HDAC5 promotes colorectal cancer cell proliferation by up-regulating DLL4 expression*. Int J Clin Exp Med, 2015. **8**(4): p. 6510-6.
184. Liu, J., et al., *Both HDAC5 and HDAC6 are required for the proliferation and metastasis of melanoma cells*. J Transl Med, 2016. **14**: p. 7.
185. Li, A., et al., *HDAC5, a potential therapeutic target and prognostic biomarker, promotes proliferation, invasion and migration in human breast cancer*. Oncotarget, 2016. **7**(25): p. 37966-37978.
186. Salhia, B., et al., *The guanine nucleotide exchange factors trio, Ect2, and Vav3 mediate the invasive behavior of glioblastoma*. Am J Pathol, 2008. **173**(6): p. 1828-38.
187. Kwiatkowska, A., et al., *The small GTPase RhoG mediates glioblastoma cell invasion*. Mol Cancer, 2012. **11**: p. 65.
188. Gu, F., et al., *Intersectin1-S, a multidomain adapter protein, is essential for malignant glioma proliferation*. Glia, 2015. **63**(9): p. 1595-605.
189. Kitzing, T.M., et al., *Positive feedback between Dia1, LARG, and RhoA regulates cell morphology and invasion*. Genes Dev, 2007. **21**(12): p. 1478-83.
190. Zeng, Y., et al., *Formin-like2 regulates Rho/ROCK pathway to promote actin assembly and cell invasion of colorectal cancer*. Cancer Sci, 2015. **106**(10): p. 1385-93.
191. Annabi, B., et al., *A MT1-MMP/NF-kappaB signaling axis as a checkpoint controller of COX-2 expression in CD133+ U87 glioblastoma cells*. J Neuroinflammation, 2009. **6**: p. 8.
192. Xia, H., et al., *Loss of brain-enriched miR-124 microRNA enhances stem-like traits and invasiveness of glioma cells*. J Biol Chem, 2012. **287**(13): p. 9962-71.

193. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations*. Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.
194. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium*. Nat Biotechnol, 2010. **28**(10): p. 1045-8.
195. Faye, L.L., et al., *Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification*. PLoS Genet, 2013. **9**(8): p. e1003609.
196. Duggal, G., H. Wang, and C. Kingsford, *Higher-order chromatin domains link eQTLs with the expression of far-away genes*. Nucleic Acids Res, 2014. **42**(1): p. 87-96.
197. Colilla, S., et al., *Estimates of current and future incidence and prevalence of atrial fibrillation in the U.S. adult population*. Am J Cardiol, 2013. **112**(8): p. 1142-7.
198. Jais, P., et al., *Catheter ablation versus antiarrhythmic drugs for atrial fibrillation: the A4 study*. Circulation, 2008. **118**(24): p. 2498-505.
199. Mont, L., et al., *Catheter ablation vs. antiarrhythmic drug treatment of persistent atrial fibrillation: a multicentre, randomized, controlled trial (SARA study)*. Eur Heart J, 2014. **35**(8): p. 501-7.
200. Bhargava, M., et al., *Impact of type of atrial fibrillation and repeat catheter ablation on long-term freedom from atrial fibrillation: results from a multicenter study*. Heart Rhythm, 2009. **6**(10): p. 1403-12.
201. Verma, A., et al., *Pre-existent left atrial scarring in patients undergoing pulmonary vein antrum isolation: an independent predictor of procedural failure*. J Am Coll Cardiol, 2005. **45**(2): p. 285-92.
202. Di Biase, L., et al., *Left atrial appendage: an underrecognized trigger site of atrial fibrillation*. Circulation, 2010. **122**(2): p. 109-18.
203. Mohanty, S., et al., *Impact of metabolic syndrome on procedural outcomes in patients with atrial fibrillation undergoing catheter ablation*. J Am Coll Cardiol, 2012. **59**(14): p. 1295-301.
204. Gudbjartsson, D.F., et al., *Variants conferring risk of atrial fibrillation on chromosome 4q25*. Nature, 2007. **448**(7151): p. 353-7.
205. Gudbjartsson, D.F., et al., *A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke*. Nat Genet, 2009. **41**(8): p. 876-8.
206. Benjamin, E.J., et al., *Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry*. Nat Genet, 2009. **41**(8): p. 879-81.
207. Ellinor, P.T., et al., *Common variants in KCNN3 are associated with lone atrial fibrillation*. Nat Genet, 2010. **42**(3): p. 240-4.
208. Ellinor, P.T., et al., *Meta-analysis identifies six new susceptibility loci for atrial fibrillation*. Nat Genet, 2012. **44**(6): p. 670-5.
209. Berry, F.B., et al., *Positive and negative regulation of myogenic differentiation of C2C12 cells by isoforms of the multiple homeodomain zinc finger transcription factor ATBF1*. J Biol Chem, 2001. **276**(27): p. 25057-65.

210. Traylor, M., et al., *Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies*. Lancet Neurol, 2012. **11**(11): p. 951-62.

211. Wang, J., et al., *Pitx2-microRNA pathway that delimits sinoatrial node development and inhibits predisposition to atrial fibrillation*. Proc Natl Acad Sci U S A, 2014. **111**(25): p. 9181-6.

212. Wynn, G.J., et al., *Efficacy of catheter ablation for persistent atrial fibrillation: a systematic review and meta-analysis of evidence from randomized and nonrandomized controlled trials*. Circ Arrhythm Electrophysiol, 2014. **7**(5): p. 841-52.

213. Pfeufer, A., et al., *Genome-wide association study of PR interval*. Nat Genet, 2010. **42**(2): p. 153-9.

214. Verweij, N., et al., *Genetic determinants of P wave duration and PR segment*. Circ Cardiovasc Genet, 2014. **7**(4): p. 475-81.

215. Sano, M., et al., *Genome-wide association study of electrocardiographic parameters identifies a new association for PR interval and confirms previously reported associations*. Hum Mol Genet, 2014. **23**(24): p. 6668-76.

216. Perez-Hernandez, M., et al., *Pitx2c increases in atrial myocytes from chronic atrial fibrillation patients enhancing IKs and decreasing ICa,L*. Cardiovasc Res, 2016. **109**(3): p. 431-41.

217. Kirchhof, P., et al., *PITX2c is expressed in the adult left atrium, and reducing Pitx2c expression promotes atrial fibrillation inducibility and complex changes in gene expression*. Circ Cardiovasc Genet, 2011. **4**(2): p. 123-33.

218. Lozano-Velasco, E., et al., *Pitx2 impairs calcium handling in a dose-dependent manner by modulating Wnt signalling*. Cardiovasc Res, 2016. **109**(1): p. 55-66.

219. Kao, Y.H., et al., *ZFHX3 knockdown increases arrhythmogenesis and dysregulates calcium homeostasis in HL-1 atrial myocytes*. Int J Cardiol, 2016. **210**: p. 85-92.

220. Jalife, J., *Mechanisms of persistent atrial fibrillation*. Curr Opin Cardiol, 2014. **29**(1): p. 20-7.

221. Morgan, R., et al., *Slow Conduction in the Border Zones of Patchy Fibrosis Stabilizes the Drivers for Atrial Fibrillation: Insights from Multi-Scale Human Atrial Modeling*. Front Physiol, 2016. **7**: p. 474.

222. Kato, T., et al., *Endothelial-mesenchymal transition in human atrial fibrillation*. J Cardiol, 2017. **69**(5): p. 706-711.

223. Di Biase, L., et al., *Periprocedural stroke and bleeding complications in patients undergoing catheter ablation of atrial fibrillation with different anticoagulation management: results from the Role of Coumadin in Preventing Thromboembolism in Atrial Fibrillation (AF) Patients Undergoing Catheter Ablation (COMPARE) randomized trial*. Circulation, 2014. **129**(25): p. 2638-44.

224. Wu, Z., Y. Hu, and P.E. Melton, *Longitudinal data analysis for genetic studies in the whole-genome sequencing era*. Genet Epidemiol, 2014. **38 Suppl 1**: p. S74-80.

225. Kim, Y.J., et al., *A new strategy for enhancing imputation quality of rare variants from next-generation sequencing data via combining SNP and exome chip data*. BMC Genomics, 2015. **16**: p. 1109.

226. Sandve, G.K., et al., *Ten simple rules for reproducible computational research*. PLoS Comput Biol, 2013. **9**(10): p. e1003285.

227. Freedman, L.P. and J. Inglese, *The increasing urgency for standards in basic biologic research*. Cancer Res, 2014. **74**(15): p. 4024-9.

228. Guo, Y., et al., *Exome sequencing generates high quality data in non-target regions*. BMC Genomics, 2012. **13**: p. 194.

229. An, S.M., Q.P. Ding, and L.S. Li, *Stem cell signaling as a target for novel drug discovery: recent progress in the WNT and Hedgehog pathways*. Acta Pharmacol Sin, 2013. **34**(6): p. 777-83.

230. Brechbiel, J., K. Miller-Moslin, and A.A. Adjei, *Crosstalk between hedgehog and other signaling pathways as a basis for combination therapies in cancer*. Cancer Treat Rev, 2014. **40**(6): p. 750-9.

231. Tai, D., et al., *Targeting the WNT Signaling Pathway in Cancer Therapeutics*. Oncologist, 2015. **20**(10): p. 1189-98.

232. Park, D.H., et al., *Activation of neuronal gene expression by the JMJD3 demethylase is required for postnatal and adult brain neurogenesis*. Cell Rep, 2014. **8**(5): p. 1290-9.

233. Guo, X., et al., *Nicotine induces alteration of H3K27 demethylase UTX in kidney cancer cell*. Hum Exp Toxicol, 2014. **33**(3): p. 264-9.

234. Kahlert, U.D., et al., *Pharmacologic Wnt Inhibition Reduces Proliferation, Survival, and Clonogenicity of Glioblastoma Cells*. J Neuropathol Exp Neurol, 2015. **74**(9): p. 889-900.

235. Bezecny, P., *Histone deacetylase inhibitors in glioblastoma: pre-clinical and clinical experience*. Med Oncol, 2014. **31**(6): p. 985.

236. Godoy, P.R., A.P. Montaldi, and E.T. Sakamoto-Hojo, *HEB silencing induces anti-proliferative effects on U87MG cells cultured as neurospheres and monolayers*. Mol Med Rep, 2016. **14**(6): p. 5253-5260.

237. Schwank, G., et al., *Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients*. Cell Stem Cell, 2013. **13**(6): p. 653-8.

238. Lancaster, M.A., et al., *Cerebral organoids model human brain development and microcephaly*. Nature, 2013. **501**(7467): p. 373-9.

239. Egashira, T., et al., *Patient-Specific Induced Pluripotent Stem Cell Models: Characterization of iPS Cell-Derived Cardiomyocytes*. Methods Mol Biol, 2016. **1353**: p. 343-53.

240. Chang, Q. and D. Hedley, *Emerging applications of flow cytometry in solid tumor biology*. Methods, 2012. **57**(3): p. 359-67.

241. Denes, V., et al., *Metastasis blood test by flow cytometry: in vivo cancer spheroids and the role of hypoxia*. Int J Cancer, 2015. **136**(7): p. 1528-36.

242. Gilfillan, G.D., et al., *Limitations and possibilities of low cell number ChIP-seq*. BMC Genomics, 2012. **13**: p. 645.

243. Saliba, A.E., et al., *Single-cell RNA-seq: advances and future challenges*. Nucleic Acids Res, 2014. **42**(14): p. 8845-60.

244. Sulem, P., et al., *Identification of a large set of rare complete human knockouts*. Nat Genet, 2015. **47**(5): p. 448-52.

245. Cai, M., et al., *4C-seq revealed long-range interactions of a functional enhancer at the 8q24 prostate cancer risk locus*. Sci Rep, 2016. **6**: p. 22462.

246. Meddens, C.A., et al., *Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease*. Genome Biol, 2016. **17**(1): p. 247.

247. Wang, S., et al., *An enhancer element harboring variants associated with systemic lupus erythematosus engages the TNFAIP3 promoter to influence A20 expression*. PLoS Genet, 2013. **9**(9): p. e1003750.

248. Wiley, L.A., et al., *Patient-specific induced pluripotent stem cells (iPSCs) for the study and treatment of retinal degenerative diseases*. Prog Retin Eye Res, 2015. **44**: p. 15-35.

249. Yoshida, M., et al., *Modeling the early phenotype at the neuromuscular junction of spinal muscular atrophy using patient-derived iPSCs*. Stem Cell Reports, 2015. **4**(4): p. 561-8.

250. Novak, A., et al., *Functional abnormalities in iPSC-derived cardiomyocytes generated from CPVT1 and CPVT2 patients carrying ryanodine or calsequestrin mutations*. J Cell Mol Med, 2015. **19**(8): p. 2006-18.

251. Xie, F., et al., *Seamless gene correction of beta-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac*. Genome Res, 2014. **24**(9): p. 1526-33.

252. Li, H.L., et al., *Precise correction of the dystrophin gene in duchenne muscular dystrophy patient induced pluripotent stem cells by TALEN and CRISPR-Cas9*. Stem Cell Reports, 2015. **4**(1): p. 143-54.

253. Huang, X., et al., *Production of Gene-Corrected Adult Beta Globin Protein in Human Erythrocytes Differentiated from Patient iPSCs After Genome Editing of the Sickle Point Mutation*. Stem Cells, 2015. **33**(5): p. 1470-9.

Vita

Amelia Weber Hall was born in Austin, TX, December 1984. After a childhood in New England spent transplanting large hods of moss from one rock to another rock, she enrolled at the University of Rochester in August 2003. There, in the "Flour City", she spent 4 years studying science, religion, living in an ecologically minded vegetarian commune, and dating a burly ginger math enthusiast. In May of 2006 she got her start as a molecular biologist-for-hire in the laboratory of Dr. Vera Gorbunova and graduated with BS in Molecular Genetics (Distinction in Research) in May of 2007.

In August of 2007, Amelia and her burly ginger set out for Austin, Texas. This was 6 to 12 months before Austin officially became the destination for "cool millennials", and they unwittingly set off a trend. From August 2007 until May 2010 Amelia worked as a laboratory technician for Dr. Richard Aldrich at UT Austin, and learned how to be slightly less socially awkward through the power of friendship. In August of 2010, she began graduate school at the University of Texas at Austin and joined the laboratory of Dr. Vishy Iyer in May of 2011.

Permanent email: ameliahall [at] utexas [dot] edu

This dissertation was typed by Amelia Weber Hall.