CHARACTERIZATION OF CHROMATIN DYSREGULATION IN CANCER THROUGH ANALYSIS OF FRESH AND ARCHIVAL HUMAN SAMPLES

JEREMY M. SIMON

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2013

Approved by:

Ian J. Davis

Corbin D. Jones

Charles M. Perou

Terrence S. Furey

Timothy C. Elston

ABSTRACT

Jeremy M. Simon: Characterization of chromatin dysregulation in cancer through analysis of fresh and archival human samples.
(Under the direction of Ian J. Davis.)

In the past several years, advances in high-throughput DNA sequencing have enabled massive tumor genome sequencing studies to identify recurrent mutations across many different cancer types and thousands of patients. These mutational surveys have reinforced the paradigm that cancer is a "disease of the genome" through the identification of inactivating mutations in well studied tumor- suppressors and activating mutations in well studied oncogenes, particularly in adult cancers. A new class of recurrent mutations has emerged as well, inactivating genes that encode chromatin regulators; these are disproportionately prevalent in pediatric and hematological malignancies. The molecular consequences of chromatin regulator mutations on a genome-wide scale, and moreover, how other genetic insults drive chromatin dysregulation and potentially enhance tumorigenesis, were until now completely unknown. In the chapters that follow, we show that a translocation-derived transcription factor chimera in Ewing Sarcoma acquired chromatin modifying activity such that it acts as a pioneer factor, altering chromatin configuration and inducing transcriptional dysregulation. We also demonstrate how alterations in chromatin link aberrancies in transcript processing with histone methyltransferase loss in clear cell Renal Cell Carcinoma through functional studies of chromatin accessibility and RNA processing in primary human tumors. Lastly, we describe how simple modifications to our experimental assay of chromatin accessibility permit the usage of archival (FFPE) human specimens. Together, in addition to contributing a greater understanding of chromatin biology and dysregulation in human cancers, this work will enable large-scale studies of the causes and roles of chromatin dysregulation in other models of human disease. It has also led to the initiation of high-throughput screens for compounds that affect chromatin accessibility, and subsequently tumor cell proliferation. Future work will utilize chromatin accessibility information as a novel clinical diagnostic and prognostic to guide and enhance patient treatment.

ACKNOWLEDGMENTS

This work would not have been possible without the support of my advisors, Ian Davis and Jason Lieb. Their guidance over the last five years has been unrelenting. My development as a scientist, a scientific writer, and my independence are all a direct consequence of your mentorship. You also have each set up an environment in which I could obtain invaluable feedback on my work. Our labs have seen a lot of change over the years, particularly growth, but each member of the Davis and Lieb labs have fostered my development. I would like to specifically acknowledge Paul Giresi, Kohta Ikegami, Dan McKay, Morten Jensen, Colin Lickwar, Mukund Patel, and Nick Gomez for their training and assistance getting my projects started. For their continued support, in combination with that of every other Davis and Lieb lab member, I am truly grateful.

The environment in which I conducted this and other work has also been highly collaborative. In my first two years, I worked very closely with Kyle Gaulton, who also contributed greatly to my success. More recently, close collaboration with Kim Rathmell and Kate Hacker, as well as Brian Strahl and Billy Kim, has proven to be not only be highly successful, but also hugely helpful to my development. Jeff Roach also assisted greatly in setting up and maintaining various computational needs through UNC ITS.

I would also like to thank the rest of my thesis committee: Corbin Jones, Chuck Perou, Terry Furey, and Tim Elston. You have each provided incredibly helpful feedback, constructive criticism, and guidance both within the confines of official committee meetings or through quick responses by email. Corbin, you have always been like a scientific uncle to me, you have provided endless recommendations and enlightenment, much of which occurring over oft-needed beer.

My fellow students in the Bioinformatics curriculum as well as the Genetics curriculum have also played an important role in my development. The monthly BCB socials, weekly

trivia nights, or the countless other scientific or non-scientific gatherings have been truly helpful in balancing out our otherwise stressful lives. In particular, I'd like to specifically thank Matt Berginski, Stephen Bush, John Didion, Tess Jeffers, Will Jeck, Michael Iglesia, Damien Croteau-Chonka, and Toby Clarke for the many unforgettable and enjoyable times we've shared.

Finally, I am truly indebted to my family who have supported me throughout my graduate career. Each of you have maintained your belief in me and supported my academic goals, helping me through the difficult times and helping celebrate the victories. Thank you for everything you have provided, I love you all!

TABLE OF CONTENTS

LIST OF FIGURES . . . . **xi**

LIST OF ABBREVIATIONS . . . . **xiv**

1 INTRODUCTION **1**

1.1 Initial discoveries from tumor genome sequencing 2

1.2 Chromatin structure, histone modifications, and regulation of gene expression 3

1.3 Epigenetics of cancer 4

1.4 Isolation of active regulatory elements from human chromatin 5

1.5 Epigenetic therapies in cancer 7

1.6 Thesis contributions 7

2 TUMOR-SPECIFIC RETARGETING OF AN ONCOGENIC TRANSCRIPTION FACTOR CHIMERA RESULTS IN DYSREGULATION OF CHROMATIN AND TRANSCRIPTION **9**

2.1 Introduction 9

2.2 Results 11

2.2.1 Chimerism and cell lineage influence genomic targeting 11

2.2.2 EWS-FLI and FLI regulate divergent gene programs 16

2.2.3 Differentially targeted regions are marked by DNA sequence and regulatory variation 19

2.2.4 Chimerism retargets EWS-FLI to tandem tetranucleotide repeats . . . 24

2.2.5 Combinatorial DNA binding motifs distinguish endothelial cell targeted sites 30

2.2.6 Epigenetic factors distinguish microsatellite repeats in Ewing Sarcoma 32

2.2.7 EWS-FLI targets enhancer like elements altering and maintaining the local chromatin environment 36

2.3 Discussion 38

2.4 Methods 43

2.4.1 Cell culture 43

2.4.2 Lentiviral knockdown-expression 44

2.4.3 Chromatin Immunoprecipitation (ChIP) and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) 44

2.4.4 Quality Control and Reference Genome Alignment 44

2.4.5 Peak calling and permutation 45

2.4.6 Hierarchical Clustering and Motif Identification 45

2.4.7 Flow Cytometry 46

3 VARIATION IN CHROMATIN ACCESSIBILITY IN HUMAN KIDNEY CANCER LINKS H3K36 METHYLTRANSFERASE LOSS WITH WIDESPREAD RNA PROCESSING DEFECTS 47

3.1 Introduction 47

3.2 Results 49

3.2.1 Differences in chromatin accessibility between tumors and normal kidney tissue corroborate the underlying role of HIF in ccRCC 49

3.2.2 *SETD2* mutations link H3K36me3 loss with changes in chromatin accessibility 50

3.2.3 *SETD2* mutation is associated with DNA hypomethylation proximal to sites of nucleosome depletion 60

3.2.4 Intron retention and splicing defects affect a large fraction of genes with altered chromatin accessibility in *SETD2*-mutant tumors 60

3.2.5 Widespread RNA processing defects linked with *SETD2* mutations persist in the mature RNA pool and are marked by altered chromatin accessibility 62

3.3 Discussion 67

3.4 Methods 76

3.4.1 Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) and hierarchical clustering of differentially open chromatin 76

3.4.2 Re-processing of HIF1A, HIF2A, and ARNT ChIP-seq data 76

3.4.3 Ontologies associated with differentially open chromatin 77

3.4.4 Motif analysis 77

3.4.5 SureSelect custom capture and mutation calling 77

3.4.6 Histone methylation ChIP-seq and data processing 77

3.4.7 Feature overlap permutations 78

3.4.8 Tissue Microarray Construction and Immunohistochemistry 78

3.4.9 Intron retention estimates by RNA-seq 79

3.4.10 Quantitative RT-PCR 79

3.4.11 Differential splicing analysis 80

4 ISOLATION OF REGULATORY ELEMENTS FROM ARCHIVAL HUMAN SPECIMENS USING FAIRE 81

4.1 Introduction 81

4.2 Results 83

4.2.1 FFPE-FAIRE shows high concordance with frozen tissue and cultured cells in controlled xenograft system 83

4.2.2 FFPE-FAIRE is robust to as few as 1 x 10^6 cells 85

4.2.3 Ongoing work 85

4.3 Discussion 87

4.4 Methods 88

4.4.1 Xenograft model and estimation of tumor nuclei density 88

4.4.2 FFPE-FAIRE 88

4.4.3 Reanalysis of EWS-FLI binding sites and GGAA microsatellites . . . 89

4.4.4 ChIN score calculation 89

5 CONCLUSIONS AND FUTURE DIRECTIONS 91

BIBLIOGRAPHY 97

LIST OF FIGURES

1.1 Disproportionate frequency of chromatin regulator mutations in pediatric and hematological cancers 6

2.1 Experimental schema for lineage-specific transcription factor silencing and expression 13

2.2 Cell cycle profile of Ewing Sarcoma cells 14

2.3 Chimerism alters ETS-mediated targeting 15

2.4 The range of factor- and cell-type-specific binding; *NR0B1* 17

2.5 The range of factor- and cell-type-specific binding; *EPHA2* 18

2.6 Differentially expressed genes in endothelial and Ewing cells 20

2.7 EWS-FLI and FLI1 differentially expressed genes have distinct biological functions 21

2.8 Hierarchical clustering identifies cell- and transcription factor-specific variation in genomic targeting 24

2.9 EWS-ETS fusions target GGAA-containing microsatellite repeats 25

2.10 Annotation of clusters of binding sites using GREAT shows biologically relevant associations with ontologies 27

2.11 Upregulated genes are closer to EWS-FLI binding sites 28

2.12 FLI1 binding sites are closer to FLI1 differentially expressed genes 29

2.13 EWS-FLI-bound tetranucleotide repeats demonstrate repeat length polymorphism 31

2.14 EWS-FLI and FLI1 occupied similar sites in a normal cellular context 33

2.15 Permutation abolished EWS-FLI and FLI1 signal enrichment at all identified sites 34

2.16 Deregulation of repetitive elements in Ewing Sarcoma 36

2.17 UCSC Genome Browser snapshots of epigenetic patterns in EWS502 (black), HUVEC (orange), and H1hESC (green) 37

2.18 UCSC Genome Browser snapshots of epigenetic patterns in EWS502 (black), HUVEC (orange), and H1hESC (green) 38

2.19 Epigenetics and chromatin accessibility of GGAA microsatellites 39

2.20 Predicted nucleosome occupancy of EWS-FLI and FLI1 binding sites 40

2.21 EWS-FLI is capable of epigenetic reprogramming 41

3.1 Schematic representation of dataset integration and genomic site identification 52

3.2 Regions of tumor-specific nucleosome eviction identify the underlying role of HIF in ccRCC 54

3.3 Gene ontology associations with sites in Clusters 1–3 55

3.4 SETD2 mutations link H3K36me3 loss with changes in chromatin accessibility 57

3.5 META-gene plot of H3K36me3 ChIP-seq signal from normal kidney 58

3.6 Decreased *SETD2* expression in *SETD2*-normal tumors results in H3K36me3 deficiency 60

3.7 Nucleosome-depleted regions in *SETD2*-mutant tumors display localized DNA hypomethylation 61

3.8 H3K36me3 deficiency is associated with intron retention 63

3.9 H3K36me3-deficient tumors display increased intron retention compared to H3K36me3-normal tumors 64

3.10 *GAPDH* exhibits low intron retention in H3K36me3-deficient tumors 65

3.11 Enriched ontologies among genes with increased intron retention 66

3.12 Widespread RNA processing defects linked with *SETD2* mutations persist in the mature RNA pool and are marked by altered chromatin accessibility 69

3.13 Intron retention in *SETD2*-mutant tumors persists into mature, polyadenylated RNA 70

3.14 *SETD2*-mutant tumors display widespread changes in RNA processing 71

3.15 Aberrant splicing is preferentially detected in highly transcribed genes 72

3.16 Aberrant splicing and intron retention is preferentially detected in long genes and genes with more exons 73

3.17 Enriched ontologies among misspliced genes 75

4.1 FFPE-FAIRE shows high concordance with frozen tissue and cultured cells in controlled xenograft system 85

4.2 FFPE-FAIRE is robust to as few as 1 x 10^6 cells 86

LIST OF ABBREVIATIONS

ChIN	Chromatin Integrity Number
ChIP	Chromatin Immunoprecipitation
DNase	Deoxyribonuclease
ENCODE	Encyclopedia of DNA Elements
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FFPE	Formalin-Fixed Paraffin Embedded
HIF	Hypoxia Inducible Factors
HUVEC	Human Umbilical Vein Endothelial Cells
IHC	Immunohistochemistry
IRS	Intron Retention Score
MNase	Micrococcal nuclease
PCA	Principal Components Analysis
RCC	Renal Cell Carcinoma
RIN	RNA Integrity Number
RPKM	Reads Per Kilobase exon per Million mapped reads
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TMA	Tissue Microarray
TSS	Transcriptional Start Sites
UTR	Untranslated Region

CHAPTER 1

INTRODUCTION

Over the course of the last decade, since the completion of the sequencing of the human genome, significant technological advances in DNA sequencing have fueled a revolution in genomics with widespread implications for studies of human health and disease. High-throughput DNA sequencing was introduced in 2005–2007 [1] [2], and since then, the cost of DNA sequencing per megabase has fallen by over four orders of magnitude, far exceeding the rate predicted by Moore's Law. Coupled with both previously known and novel molecular biological techniques, such as chromatin immunoprecipitation (ChIP) [3] [4] [5] [6], DNase hypersensitivity [7] [8], micrococcal nuclease (MNase) digestion [9], Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) [10], bisulfite (BiS) conversion [11], RNA immunoprecipitation (RIP) [12], and others [13], high-throughput sequencing has since been used extensively to study gene expression and its many regulatory processes in human cell lines and tissues. Large consortia such as ENCODE, The Cancer Genome Atlas (TCGA), and the Epigenome Roadmap have generated an enormous wealth of sequencing-based data, much of which is now available to the scientific community. Despite their goals being rooted in basic research, these consortia, as well as various groups utilizing their data, have begun to examine the processes that govern human health and those that underlie disease (e.g. [14] [15]. The entire field of medicine, particularly oncology, is thus on the verge of a metamorphosis in which genomic technologies guide patient diagnostics, prognostics, and therapy.

1.1 Initial discoveries from tumor genome sequencing

The Cancer Genome Atlas, the International Cancer Genome Consortium, and others have now cataloged the frequencies and types of mutations in many different cancer types and subtypes across thousands of patients [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26]. Many of the results have been predictable though nonetheless useful. One such finding is that cancers with a clear link to mutagenic exposure such as melanoma, lung squamous cell carcinoma, and lung adenocarcinoma often have genomes with tens of thousands of mutations in coding and non-coding space, large aberrancies in copy number, and catastrophic large- or small-scale genetic insults such as chromothripsis or kataegis, respectively [25] [26]. This high mutation rate observed is often coupled with DNA replication or repair defects [25] [26]. Another such finding is that there are recurrent activating mutations in well-studied oncogenes (e.g. Ras, Raf, *SRC*, *EGFR*, *MYC*, etc) and inactivating mutations in well-studied tumor suppressors (e.g. *TP53*, *CDKN2A*, *PTEN*, *RB1*, etc) [25]. These mutations were discovered in the 1970s and 1980s because they are not only prevalent among many cancer types, but also can directly cause oncogenic transformation of normal cells (reviewed in [27]).

The derangement of cellular signaling, for example through oncogene activation, also has prominent effects at the level of transcription (e.g. [28]). Differences in transcript abundance can be exploited to sub-classify tumors into groups that demonstrate differences in patient outcome or therapeutic response. Breast carcinomas, for example, can be subclassified into subtypes based on their transcriptional profiles [29] [18]. Similar analysis of glioblastomas revealed four tumor subtypes, each of which have transcriptional differences likely driven largely by mutations in *TP53*, *PDGFRA*, *NF1*, or *PTEN* [30]. Together, oncogenic activation and/or tumor suppressor silencing and the observable defects in cell signaling, transcription, and cell proliferation have reinforced the paradigm that cancer is a “disease of the genome”.

Tumor genome sequencing has also yielded a number of surprising results. Most notably, a new class of mutations have been found in genes that encode proteins involved with chromatin remodeling or modification of histones or DNA (e.g. *PBRM1*, *SMARCB1*, *ARID1A*, *SETD2*, *MLL2*, *DNMT3A*, *UTX*, etc), or encode histone core proteins themselves or their variants (e.g. *H3F3A*) [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] [42] [43]. The advent of high-throughput sequencing allowed for the discovery of mutations in a tumor genome in two weeks or less, but prior to the studies described here, no comprehensive analysis of the molecular consequences of chromatin regulator mutations has been performed in human tumors.

1.2 Chromatin structure, histone modifications, and regulation of gene expression

The genomes of eukaryotic organisms are packaged into a structure known as chromatin. This packaging allows for the approximately 2 meter long DNA polymer to fit within a 2-micron nucleus. This packaging is achieved by wrapping the DNA double helix approximately twice (147 bp) around an octamer of histone proteins, consisting of two copies each of histone proteins H2A, H2B, H3, and H4 [44]. This repeating functional unit of chromatin, in which DNA is complexed with histone proteins, is known as the nucleosome core particle. Nucleosomes are further coiled and compacted to form the 30 nm fiber, a solenoidal structure that is further compacted to form the mitotic chromosome [45].

The regulation of chromatin structure and DNA-templated processes such as transcription is tightly orchestrated. The binding of sequence-specific regulatory proteins including transcription factors is affected by chromatin organization. Displacement, destabilization, or repositioning of nucleosomes is, in many cases, a necessary precursor to the binding of such regulatory factors (e.g. [46]), and may influence the modulation of gene expression. In addition to positioning of nucleosomes themselves, modification of the tails of histones H3 and H4 can influence transcription factor binding and activity. Histone tails can be methylated, acetylated, phosphorylated, ubiquitylated, and/or otherwise decorated by many post-translational modifications [47] [48]. Nucleosomes may also carry histone variants, which together with post-translational modifications can lead to changes in nucleosome stability [49] [50] [51].

Histone modifications and variants are often heritable across generations, therefore are frequently classified as “epigenetic” alterations, although this designation is controversial [52]. Each modification and variation alters nucleosome stability or functions to roughly divide the genome into active, recently active, poised, or repressed states (e.g. [53]). Many such regions of the genome are demarcated differentially by histone modifications or exhibit altered nucleosome occupancy across cell types in a way that corresponds to cell-type-specific gene expression [53] [14] [15] [54] [55]. Each of these dynamic processes are carefully controlled by histone methyltransferases and demethylases, histone acetylases and deacetylases, and chromatin remodelers (reviewed in [55]).

1.3 Epigenetics of cancer

Many forms of cancer are now known to contain mutations in chromatin regulators. It has been suggested that alterations at the level of chromatin resulting from these mutations may play a prominent role in oncogenesis [32], may confer differences in patient survival, or may associate with more advanced disease [56]. To investigate whether the prevalence of chromatin regulator mutations was uniform across cancer types, we mined data from numerous large-scale tumor-sequencing projects, as well as other primary publications and databases for 22 different cancer types (rhabdoid, neuroblastoma, medulloblastoma, Acute Lymphoblastic Leukemia, Ewing Sarcoma, glioblastoma multiforme, clear cell Renal Cell Carcinoma, melanoma, lung adenocarcinoma, Acute Myeloid Leukemia, esophageal adenocarcinoma, lung squamous cell carcinoma, Chronic Lymphocytic Leukemia, as well as tumors of the cervix, thyroid, breast, ovary, prostate, colon/rectum, stomach, pancreas, and bladder) [16] [18] [17] [19] [20] [21] [22] [23] [24] [25] [26] [57] [58] [59] [60] [61], SEER. We compared the median age of cancer onset and the median number of coding mutations per megabase of exome (“mutational load”) in the cancers. We also computed what we refer to as the “epigenetic load”, which is the frequency at which a chromatin regulator mutation appeared in the five most abundant mutations for a particular cancer, normalized by the total mutational load for that cancer. These data illustrate that a new paradigm is emerging: pediatric and hematological cancers carry a surprisingly low mutational load, and those mutations that do occur are disproportionately prevalent in chromatin regulators. (Figure 1.1). The presumed effects on chromatin therefore suggest that these tumors can be categorized by being a disease of the genome, and also more specifically a disease of the epigenome. There was one exception to the group of pediatric tumors: Ewing Sarcoma. This is likely due to the unique mechanism of carcinogenesis in that cancer, which is caused by expression of a translocation product (EWS-FLI; discussed in Chapter 2 [57]). Other cancers carrying translocations that form chimeric DNA-binding factors, including prostate cancer and various forms of leukemia, may thus also exhibit alterations in chromatin structure.

Mutations in chromatin regulators, though prevalent in pediatric malignancies, are also present in many if not all of the adult cancers but typically at lower frequency; one notable exception is clear cell Renal Cell Carcinoma. It is not yet well understood which class of mutation occurs first in these contexts. It has been suggested that epigenetic alterations, such as the loss of DNA methylation, may lead to additional mutations due to genomic instability [62] or induced plasticity in cellular differentiation [63]. Therefore, one possible theory may be that mutations in chromatin regulators arise at an early age. Sometimes these mutations result in pediatric malignancy if they occurred in a particularly susceptible cellular niche. Other times, however, these mutations may initially contribute only subtly to changes in cellular processes but create an environment in which many more mutations can accumulate over the course of many years. This theory could offer an explanation for the higher mutational load in adult cancers, particularly those without known associations with mutagenic exposure.

1.4 Isolation of active regulatory elements from human chromatin

Elucidating the functional consequences of mutations in genes encoding chromatin regulatory proteins in human tumor specimens requires the application of techniques initially developed for cultured cells. Chromatin accessibility is characterized by the displacement or

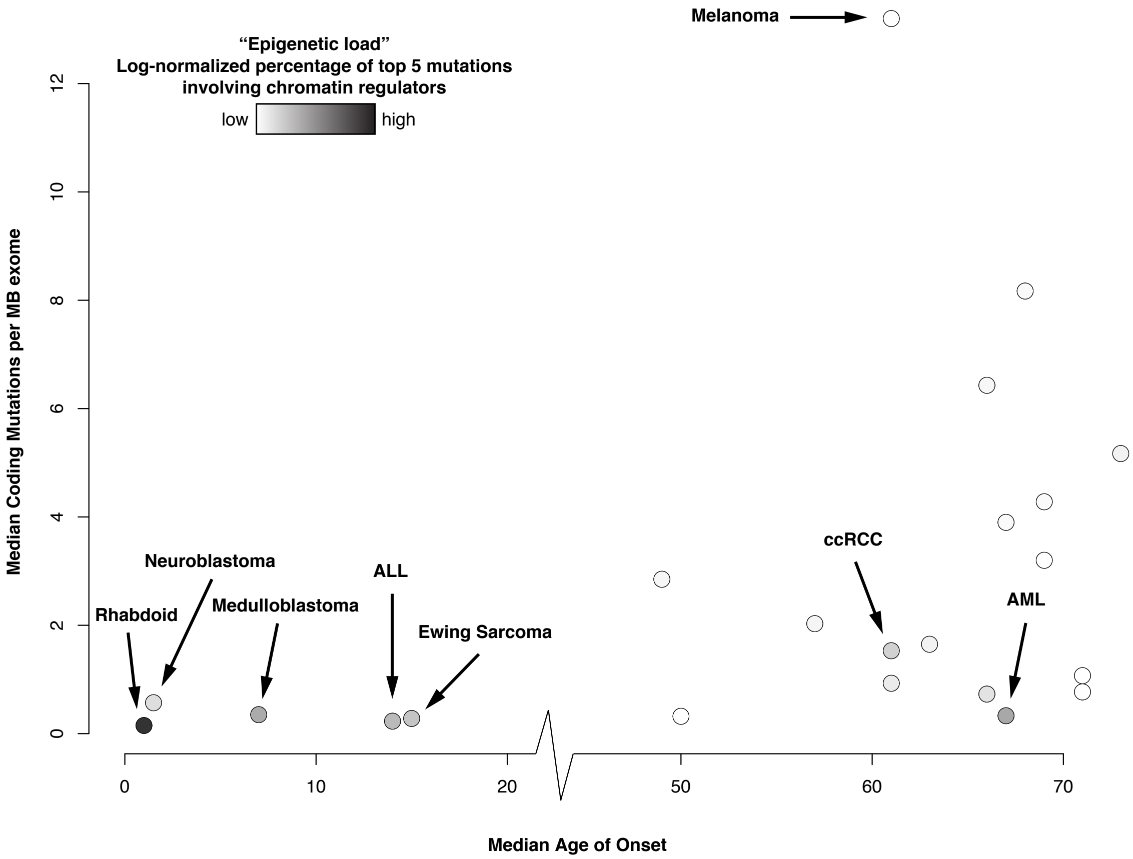


Figure 1.1: Disproportionate frequency of chromatin regulator mutations in pediatric and hematological cancers. The median age of onset (x-axis) is plotted against the median number of coding mutations per megabase of exome ("mutational load", y-axis). Points are colored based on their "epigenetic load", computed as the frequency at which a chromatin regulator mutation appeared in the five most abundant mutation types for a particular cancer, normalized by the total mutational load for that cancer.

destabilization of nucleosomes from chromatin through the action of transcriptional regulators. The isolation of nucleosome-depleted regions ("open chromatin") thus identifies functional gene regulatory elements across the genome. Open chromatin has traditionally been assayed via preferential digestion by nucleases such as DNase I [64] [65] [66] [67] [68] [69] [70] [71]. An alternative methodology for the isolation of regulatory elements is termed FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) [10] [15] [72] [73] [74] [75]. The technique has now been used in a wide range of eukaryotes, from *Plasmodium* [76] to maize [77], and we recently demonstrated its efficacy in the Kaposi Sarcoma-associated Herpesvirus

[78]. We have now extended the FAIRE technique to permit studies of chromatin accessibility in both primary human tissues and tumors as well as Formalin-Fixed Paraffin-Embedded (FFPE) tissue specimens.

1.5 Epigenetic therapies in cancer

Cancers can be treated by compounds that act through chromatin or DNA-modifying enzymes such as inhibitors of DNA methylation and histone deacetylases. Trapoxin, and variants thereof, as well as other compounds have been known since 1996 [79] to inhibit histone deacetylase activity [80], and inhibitors of DNA methylation (such as 5-azacytidine) were first synthesized in 1974 [81]. More recently, novel classes of histone deacetylase inhibitors, histone methyltransferase inhibitors (e.g. EPZ-004777 against DOT1L [82] and EPZ-6438 against EZH2 [83]), DNA methyltransferase inhibitors, and bromodomain inhibitors (e.g. JQ1 against BRD4 [84], have proven effective in certain cancer contexts. This exciting new wave of small molecule inhibitors will provide novel therapeutic options, especially when used in conjunction with other vetted compounds, for many forms of cancer. It will be of great importance to study the effects of these compounds on the epigenome itself in cancer cells in the development of specific biological therapies that target chromatin.

1.6 Thesis contributions

The experiments described here show that transcription factor chimerism leads to chromatin dysregulation in Ewing Sarcoma (Chapter 2), that mutations in chromatin regulatory proteins can lead to changes in chromatin accessibility and widespread RNA processing defects in clear cell Renal Cell Carcinoma (Chapter 3), and that modifications to the FAIRE procedure make this technique compatible with archived clinical specimens and tissue biopsies (Chapter 4). The ability to assay chromatin accessibility in archival specimens will allow us to follow the effects of cancer therapies longitudinally in single patients, perform large-scale studies of rare diseases, and perhaps lead employment of FAIRE as a high-throughput clinical diagnostic. Moreover, these data have enabled high-throughput screens for compounds that affect chromatin accessibility, and contribute generally to a greater understanding of chromatin biology and dysregulation in human cancers (Chapter 5).

This work has been a highly collaborative effort. In Chapter 2, all ChIP experiments, Western blots, gene expression microarrays, viral-mediated knockdown/re-expression experiments, and quantitative PCR was performed by Mukund Patel and Andrew McFadden. I performed all computational analyses and interpretation of high-throughput sequencing data as well as all FAIRE experiments.

In Chapter 3, Kate Hacker prepared all RNA for sequencing, analyzed and interpreted data from tissue microarrays, performed all altered splicing validation, functionally annotated all SETD2 mutations, prepared many libraries for high-throughput sequencing, and contributed significantly to biological interpretation of the data. Darshan Singh analyzed TCGA RNA-seq data for altered splicing and Joel Parker analyzed genotyping data from our tumor cohort for mutations. I performed all FAIRE and genotyping experiments, prepared many libraries for high-throughput sequencing, analyzed all FAIRE and RNA data, developed the Intron Retention Score and hierarchical clustering methods, analyzed H3K36me3 ChIP-seq data and TCGA DNA methylation data, and led the integration and biological interpretation of the data.

In Chapter 4, I performed all experiments and analyzed all of the data.

CHAPTER 2

TUMOR-SPECIFIC RETARGETING OF AN ONCOGENIC TRANSCRIPTION FACTOR CHIMERA RESULTS IN DYSREGULATION OF CHROMATIN AND TRANSCRIPTION

2.1 Introduction

Recurrent chromosomal translocations have been associated with an increasingly wide range of human cancers. Commonly involving genes encoding transcriptional regulators, translocations can deregulate gene expression and generate structurally novel oncogenic fusion proteins [85]. The transforming activity of these chimeric genes typically reveals cell type specificity, suggesting that certain lineages are permissive for transformation. Studies of oncogenic transcription factors have typically focused only on the fusion products or their target genes and often in heterologous cells, limiting insights into the relative influence of chimerism and cell lineage.

Ewing Sarcoma, a bone tumor of children and young adults, is characterized by translocations that fuse a member of the TET family to a member of the ETS transcription factor family [86] [87] [88]. Identified in 80–85% of Ewing Sarcoma, t(11;22)(q24;q12) results in an in-frame fusion of *EWSR1* to *FLI1* [86]. EWS-FLI has been shown to be a potent transcriptional modulator critical for transformation [89] [90]. Structure-function experiments have demonstrated that the EWSR1 domain contributes transactivation activity whereas the FLI1 domain directs DNA binding, and both are required for transformation [91] [92]. EWS-FLI mediates oncogenesis by directly or indirectly regulating genes necessary for transformation. Despite evidence that EWS-FLI is necessary for transformation, ectopic expression of EWS-FLI fails to activate similar genetic programs or transform most human cell lines, indicating that cell specificity is a major determinant of EWS-FLI activity [93] [94] [95].

FLI1, a member of the ETS family, is an important developmental transcription factor [96]. FLI1 deletion in mice results in embryonic death from hemorrhage associated with aberrant hematopoiesis and vasculogenesis, supporting a role in endothelial development [97] [98] [99]. Translocations involving ETS members have been implicated in other cancers, including prostate adenocarcinoma [100]. The function of EWSR1 is less well-understood, however, reports suggest participation in transcription or RNA splicing [101]. EWSR1-deficient mice die prior to weaning and show defects in B-cell development and meiosis [102]. Other translocations involving EWSR1 have been identified, resulting in chimeras with ATF1 and WT1 in Clear Cell Sarcoma and Desmoplastic Small Round Cell Tumors, respectively [103] [104] [100].

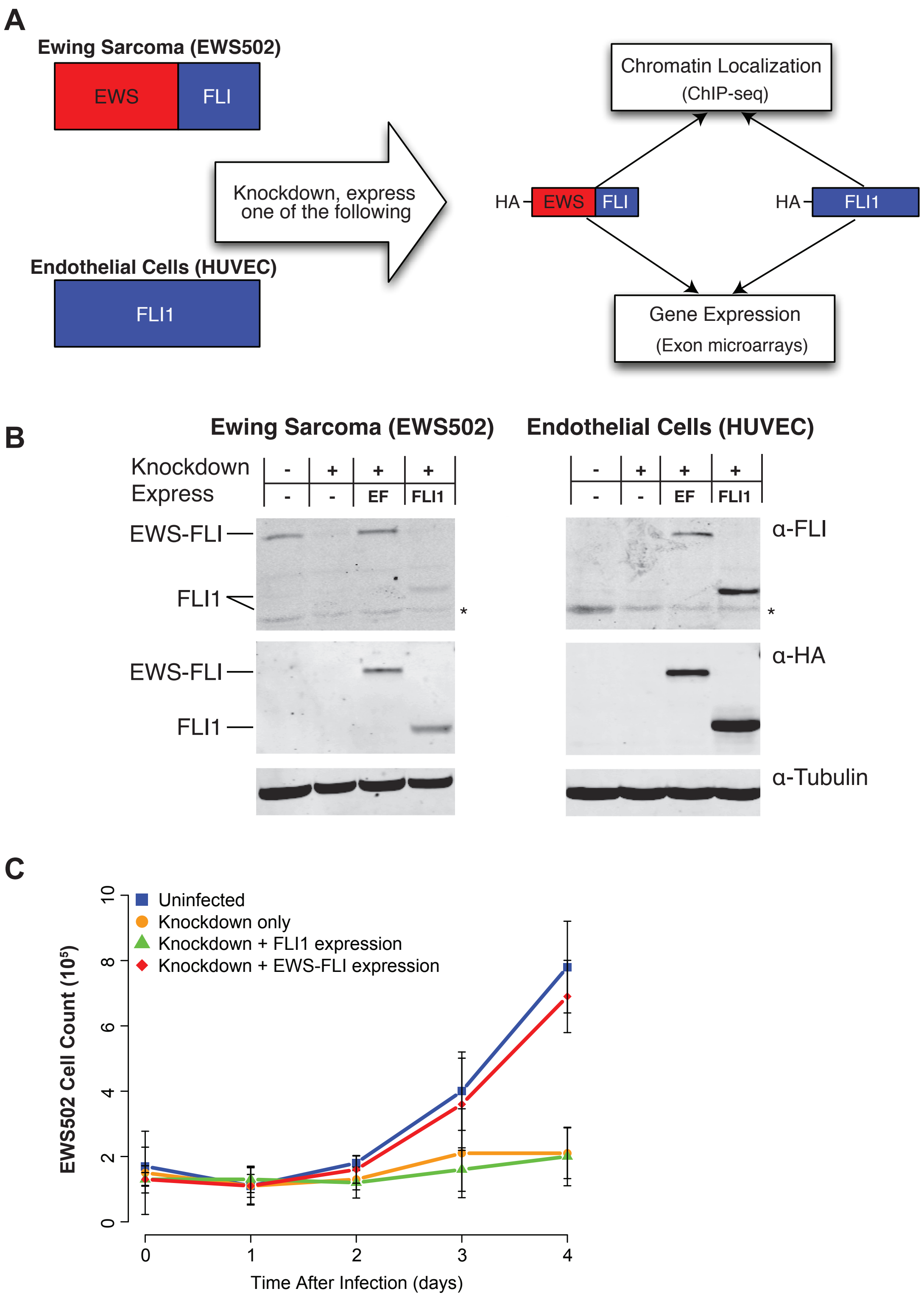
To characterize the changes in genomic localization and transcriptional output due to chimerism, we compared EWS-FLI with FLI1 in Ewing Sarcoma and human primary endothelial cells. We integrated genomic targeting with gene expression profiling and found that in tumor cells, EWS-FLI associated with distinct genomic regions lacking canonical ETS binding sites and activating a set of genes associated with a transformed phenotype. However, in endothelial cells, genomic targeting and gene regulation were similar to that of FLI1. We then examined the influence of epigenetics on this differential targeting by analyzing nucleosome occupancy and histone modifications. We found that in Ewing cells, EWS-FLI-targeted sites exhibited features characteristic of enhancer elements and were bound by RNA Polymerase II. Moreover, EWS-FLI silencing resulted in increased nucleosome occupancy of these regions. In endothelial cells, this same set of regions are normally associated with repressive chromatin, but become nucleosome depleted upon EWS-FLI expression. These data establish EWS-FLI as a pioneer factor capable of inducing and maintaining epigenetic reprogramming.

2.2 Results

2.2.1 Chimerism and cell lineage influence genomic targeting

To compare EWS-FLI with its parental protein FLI1, we developed a lentiviral delivery approach that permitted concurrent silencing of endogenous EWS-FLI or FLI1 and expression of an epitope-tagged version of EWS-FLI or FLI1 (Figure 2.1A). Lentiviral knockdown-replacement was performed in a Ewing Sarcoma cell line (EWS502) and primary human endothelial cells (HUVEC). HUVEC were selected as they abundantly express FLI1, and FLI1 has been implicated in endothelial development [98] [105] [99]. Genomic localization of each protein was examined by chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq). Gene expression was also examined using exon microarrays. The lentiviral knockdown-replacement strategy offered a number of experimental benefits to facilitate comparative genomic analyses. First, viral transduction enabled the titration of protein expression, avoiding overexpression while achieving efficient knockdown (Figure 2.1B). Second, expression of a shRNA directed to the $3'$ UTR of FLI1 (able to target both endogenous EWS-FLI and FLI1 but not the transduced genes which do not contain the $3'$ UTR) in all experimental conditions minimized the possibility for the detection of off-target effects. Finally, the use of a common and robust antibody for chromatin immunoprecipitation circumvented issues of antibody sensitivity, specificity, and antigenic variability, factors that can complicate downstream comparisons, as recently demonstrated [106].

EWS-FLI and FLI1 were expressed to approximate endogenous protein levels in both cell types (Figure 2.1B). We examined cell proliferation after EWS-FLI knockdown in the presence or absence of ectopically expressed EWS-FLI or FLI1. In the tumor cells, transduced EWS-FLI, but not FLI1, rescued the growth arrest resulting from endogenous EWS-FLI silencing (Figure 2.1C and Figure 2.21). These data support previous reports indicating the


A
Ewing Sarcoma (EWS502)
EWS
FLI
Endothelial Cells (HUVEC)
FLI1
Knockdown, express one of the following
Chromatin Localization
(ChIP-seq)
HA
EWS
FLI
HA
FLI1
Gene Expression
(Exon microarrays)
B
Ewing Sarcoma (EWS502)
Endothelial Cells (HUVEC)
Knockdown
Express
- + + +
- - EF FLI1
EWS-FLI
FLI1
*
α-FLI
EWS-FLI
FLI1
α-HA
α-Tubulin
C
Uninfected
Knockdown only
Knockdown + FLI1 expression
Knockdown + EWS-FLI expression
EWS502 Cell Count (10⁵)
0
2
4
6
8
10
0
1
2
3
4
Time After Infection (days)

requirement of EWS-FLI for cell proliferation [107] [108] [109] [90]. Inhibition of FLI1 expression or ectopic expression of EWS-FLI did not affect the proliferation of endothelial cells under the conditions tested (data not shown).

Differential activities of EWS-FLI and FLI1 could result from either of two mechanisms. The transcription factors could target similar genomic sites due to their common DNA binding domain but vary in their ability to modulate gene expression. Alternatively, chimerism could result in genomic retargeting such that differences in transcriptional output would result from variation in the sites of chromatin association. To test these two hypotheses, we performed ChIP-seq for EWS-FLI and FLI1 in both EWS502 and HUVEC. Analyzing only high quality, uniquely aligned reads, sites of genomic enrichment for each factor were determined using the Zero Inflated Negative Binomial Algorithm (ZINBA), a flexible statistical model that adjusts for the effects of GC content, mapability, and copy number variation [110]. We identified 7,172 and 13,878 potential EWS-FLI binding regions in EWS502 and HUVEC, respectively. FLI1 bound 18,958 regions in EWS502 and 39,439 regions in HUVEC (Figure 2.3A). The greater number of EWS-FLI binding sites identified in this study compared to previous ChIP-chip and ChIP-seq approaches [111] [112] [113] likely reflects greater sequencing depth and enhanced antibody sensitivity. Despite the use of different tumor cells, nearly 75% of the sites previously identified by ChIP-seq [112] overlap the regions bound by EWS-FLI in this study.

Examination of specific genomic loci demonstrated the contribution of chimerism and cell lineage to targeting (Figure 2.3A–B). For example, a site near *NR0B1* previously shown to

Figure 2.1: Experimental schema for lineage-specific transcription factor silencing and expression. A. Ewing Sarcoma (EWS502) cells and primary human endothelial cells (HUVEC) were transduced with lentivirus expressing FLI1 3′UTR-directed shRNA and HA epitope-tagged versions of EWS-FLI, FLI1, or EWSR1. B. Anti-FLI1 or anti-HA immunoblots of Ewing Sarcoma cells (EWS502) or endothelial cells (HUVEC) demonstrating concurrent silencing and replacement with HA-EWS-FLI (EF), HA-FLI1, or HA-EWSR1 (EWS). Tubulin serves as a loading control. Asterisks indicate where a background band runs at a similar molecular weight as endogenous FLI1. C. After EWS-FLI1 silencing alone (Knockdown) or together with ectopic EWS-FLI1 or FLI1 expression, EWS502 cells were counted. EWS-FLI expression, but not FLI1, rescues the effect of knockdown on proliferation.

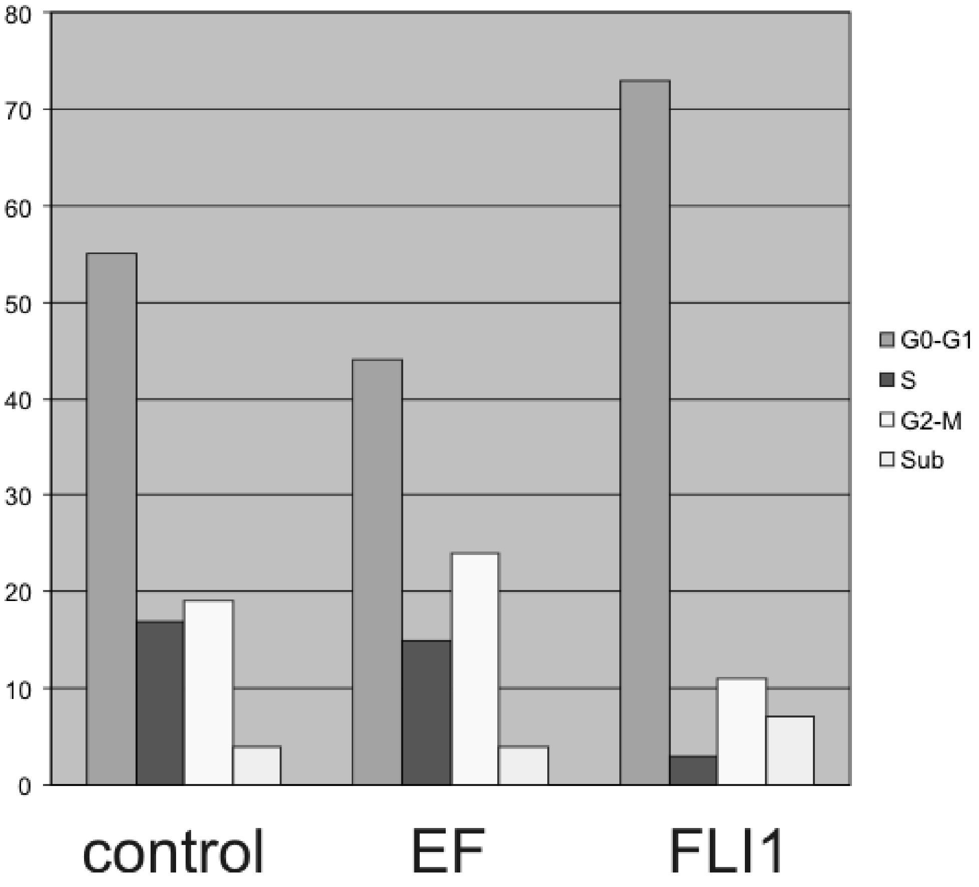


Figure 2.2: Cell cycle profile of Ewing Sarcoma cells. The percentage of cells in G0/G1, S, G2-M, or sub-G0 (Sub) based on propidium iodine staining and flow cytometry were calculated for uninfected A673 cells ("control") and cells in which endogenous EWS-FLI was silenced with concurrent HA-EWS-FLI ("EF") or HA-FLI1 ("FLI1") expression.

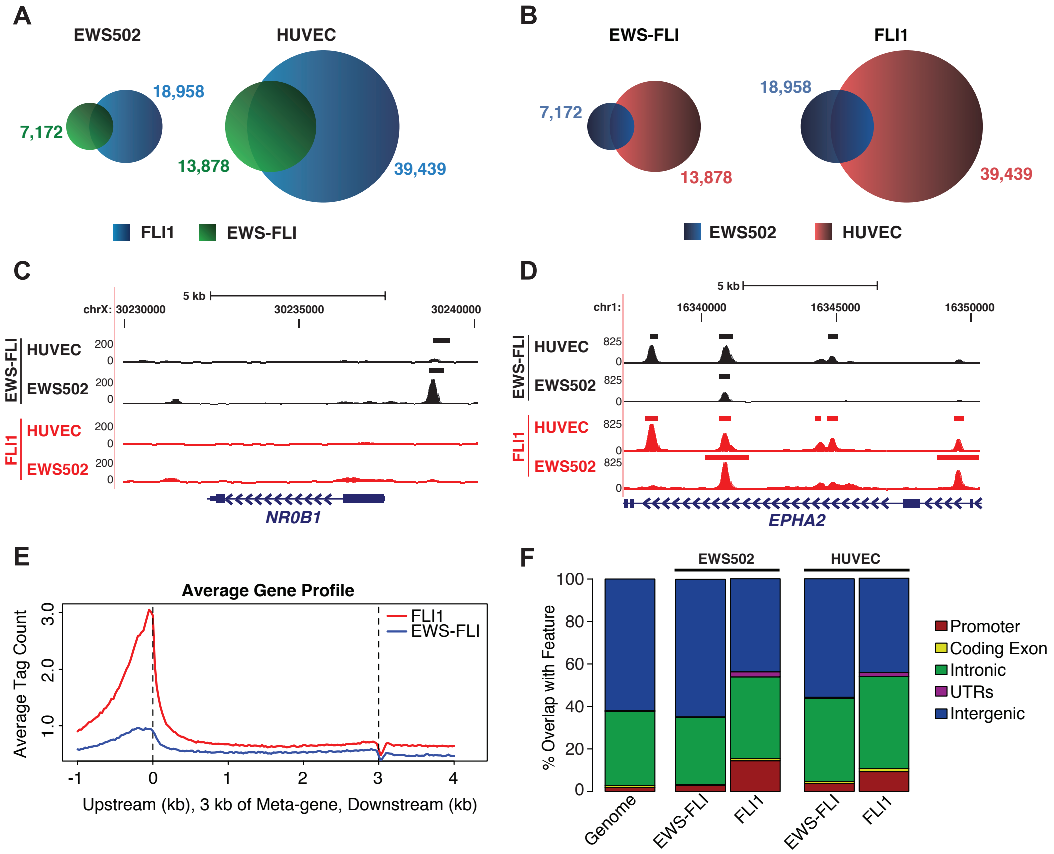


Figure 2.3: Chimerism alters ETS-mediated targeting. A–B. Venn diagrams showing the number of unique and overlapping EWS-FLI and FLI1 binding regions within the same cell type (A) or across cell types (B). C–D. UCSC Genome Browser screenshots of EWS-FLI and FLI1 ChIP-seq signal at two genes *NR0B1* (B) and *EPHA2* (C). Horizontal bars indicate targeted sites identified by ZINBA. Tag counts are shown in the Y-axis. E. Meta-gene profile of EWS-FLI and FLI1 ChIP-seq reads. 1 kb upstream of the TSS through 1 kb downstream of transcriptional termination is represented. F. Percent overlap of ZINBA-identified EWS-FLI and FLI1 binding sites with major functional genomic features. Genomic distribution of features (Genome) is shown for comparison.

be occupied by EWS-FLI [89] [114] [111] [115] was bound by EWS-FLI but not FLI1 in both cell types (Figure 2.3C and Figure 2.4). In contrast, sites around the ephrin receptor, *EPHA2*, revealed a more complex pattern (Figure 2.3D and Figure 2.5). Sites exclusive to one transcription factor or cell type were identified, as were sites common to both transcription factors and cell types. Overall, in the tumor cells, 46% of EWS-FLI sites overlap FLI1 sites, whereas in HUVEC 75% of EWS-FLI sites overlap with FLI1 (Figure 2.3A). Comparing targeting across cell types, 45% of EWS-FLI and 55% of FLI1 sites were shared between EWS502 and HUVEC (Figure 2.3B).

Genomic localization was examined by comparing the raw ChIP-seq signal over all genes (Figure 2.3E). In tumor cells, FLI1 signal was greater at transcriptional start sites (TSS), in the proximal upstream region, and through the gene body compared to EWS-FLI. Given the relative absence of EWS-FLI signal at these genic regions, we compared the overall genomic distribution of binding sites (Figure 2.3F). Again in tumor cells, FLI1 showed greater association with promoters and $5'$ and $3'$ UTRs than EWS-FLI. Compared to FLI1, EWS-FLI bound more frequently at distal intergenic regions (>60%). Although EWS-FLI and FLI1 shared occupancy at a high fraction of sites in endothelial cells, EWS-FLI demonstrated slightly less association with introns and more with intergenic regions than FLI1 (Figure 2.3F). These data suggest that in both cancer and normal cells, FLI1 targets genic sites, and chimerism leads to retargeting to intergenic regions. However, chimerism-induced retargeting is significantly mitigated by cell lineage.

2.2.2 EWS-FLI and FLI regulate divergent gene programs

To explore the transcriptional implications of genomic retargeting, EWS-FLI and FLI1-associated gene expression changes in both cell types were examined using exon microarrays. Differentially regulated genes were identified by comparing RNA from cells in which the endogenous transcription factor had been silenced to those expressing either EWS-FLI or FLI1. Observed transcriptional changes may reflect both direct and indirect effects of transcription

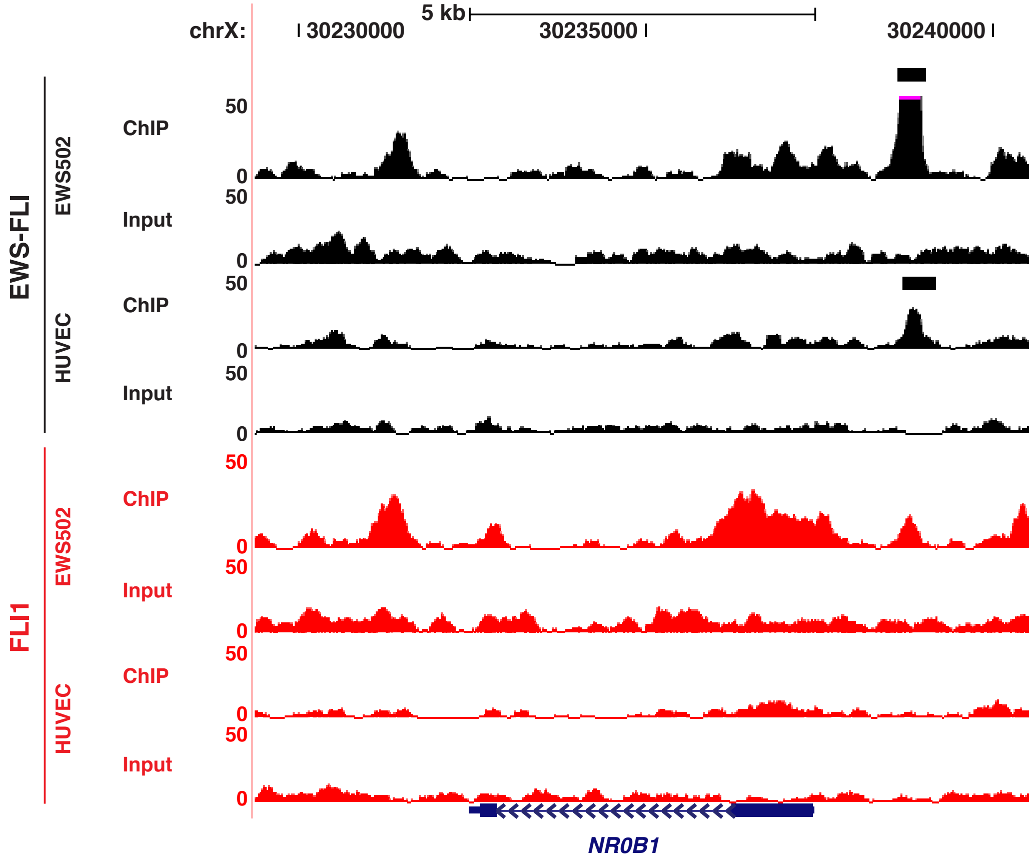


Figure 2.4: The range of factor- and cell-type-specific binding; *NR0B1*. EWS-FLI (black) and FLI1 (red) ChIP and input control signal in both Ewing Sarcoma cells and HUVECs at *NR0B1*. Viewing range is cut at 50 reads. Horizontal bars represent bound sites as identified by ZINBA. Scale bar and schema of each gene are depicted at the top and bottom of the represented tracks.

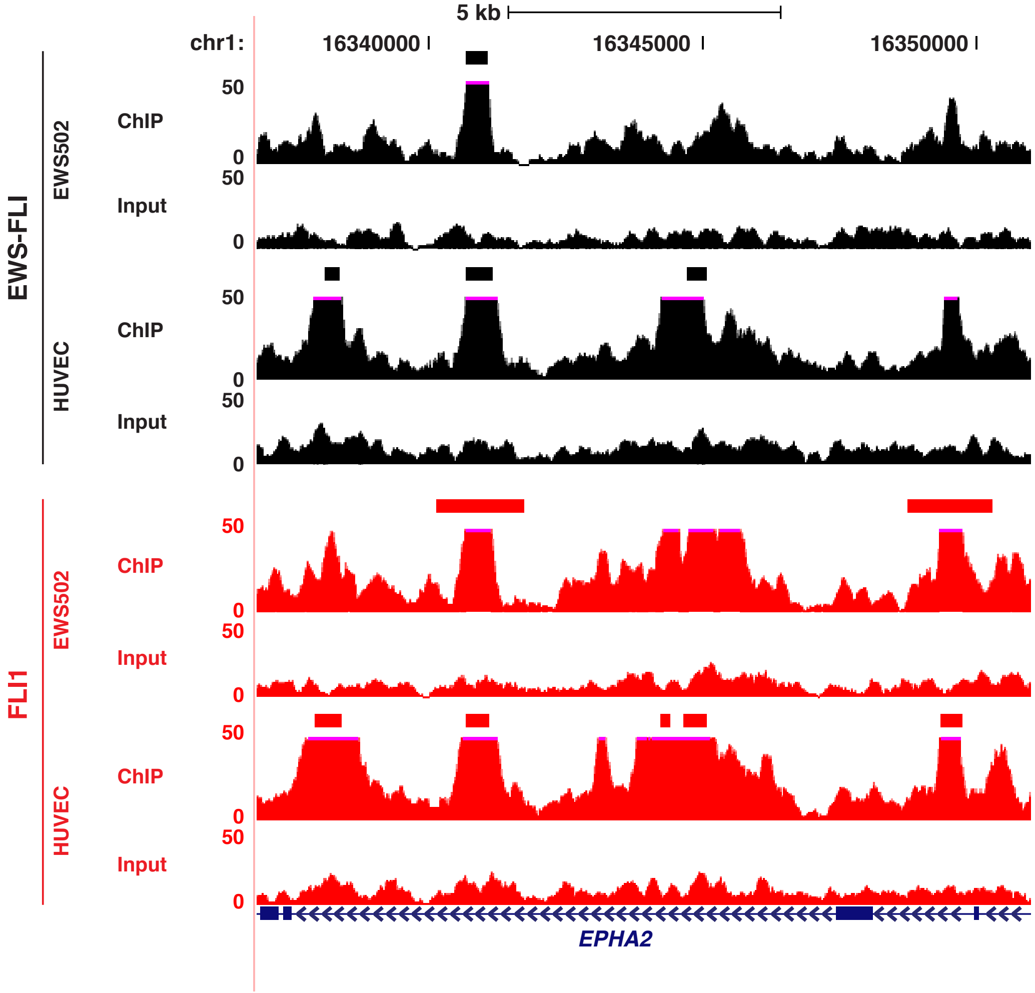


Figure 2.5: The range of factor- and cell-type-specific binding; *EPHA2*. EWS-FLI (black) and FLI1 (red) ChIP and input control signal in both Ewing Sarcoma cells and HUVECs at *EPHA2*. Viewing range is cut at 50 reads. Horizontal bars represent bound sites as identified by ZINBA. Scale bar and schema of each gene are depicted at the top and bottom of the represented tracks.

factor expression. Although EWS-FLI occupied fewer genomic sites than FLI1, it modulated the expression of more genes in both cell types. This difference was greatest in tumor cells in which EWS-FLI altered the expression of three times as many genes as FLI1 (Figure 2.6A). Genes regulated by FLI1 were mostly distinct from those regulated by EWS-FLI, with 40–45% shared in either cell type. However, of the genes commonly modulated by either factor in HUVEC, 97% were regulated concordantly, whereas in tumor cells, opposing effects on gene expression were frequently observed (41% of coregulated genes) (Figure 2.6B). Cell-type-specific regulation was also evident. Only 34% of genes differentially expressed by EWS-FLI were shared across the two cell types, whereas only 12% of FLI1 differentially expressed genes were shared.

The classes of genes regulated by EWS-FLI and FLI1 also differed significantly (Figure 2.7). Approximately one-third of the genes modulated by EWS-FLI in tumor cells were implicated in cancer or cell cycle regulation; the identification of these categories supports of previous studies of gene regulation by EWS-FLI [116] [90] [95] [117]. In contrast, FLI1 expression in tumor cells induced genes associated with hematopoiesis, hematological system development and function, and cellular development, including genes of the ephrin, thrombin, and relaxin signaling pathways. In endothelial cells, similar gene ontologies were modulated by both transcription factors. These data suggest that cell type influences the impact of chimerism on transcriptional output.

2.2.3 Differentially targeted regions are marked by DNA sequence and regulatory variation

Since genomic sites of EWS-FLI and FLI1 occupancy were mostly distinct, we hypothesized that additional factors might specify EWS-FLI or FLI1 targeting in a transcription factor- or cell-type-specific manner. We employed a computational strategy that selected binding sites that most discriminated transcription factor or cell type and then performed hierarchical clustering using the normalized ChIP-seq signals for each region (Figure 2.8A). Six major clusters of binding sites emerged. These clusters exhibited both transcription factor- and cell-type-

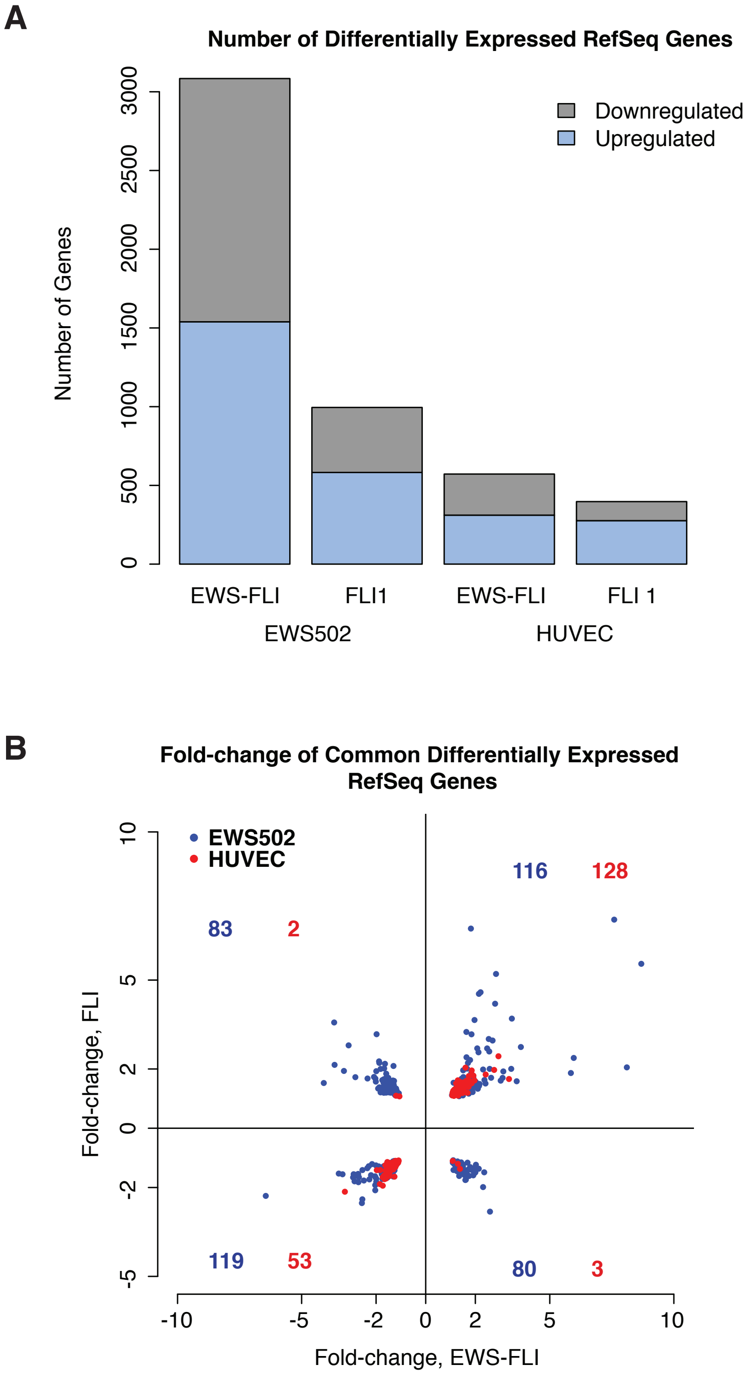


Figure 2.6: Differentially expressed genes in endothelial and Ewing cells. A. Number of up- and down-regulated RefSeq genes identified in each cell type for each transcription factor. B. Fold-change of genes commonly differentially expressed in EWS502 (blue) and HUVEC (red) cells by EWS-FLI or FLI1. Numbers shown represent gene counts per quadrant.

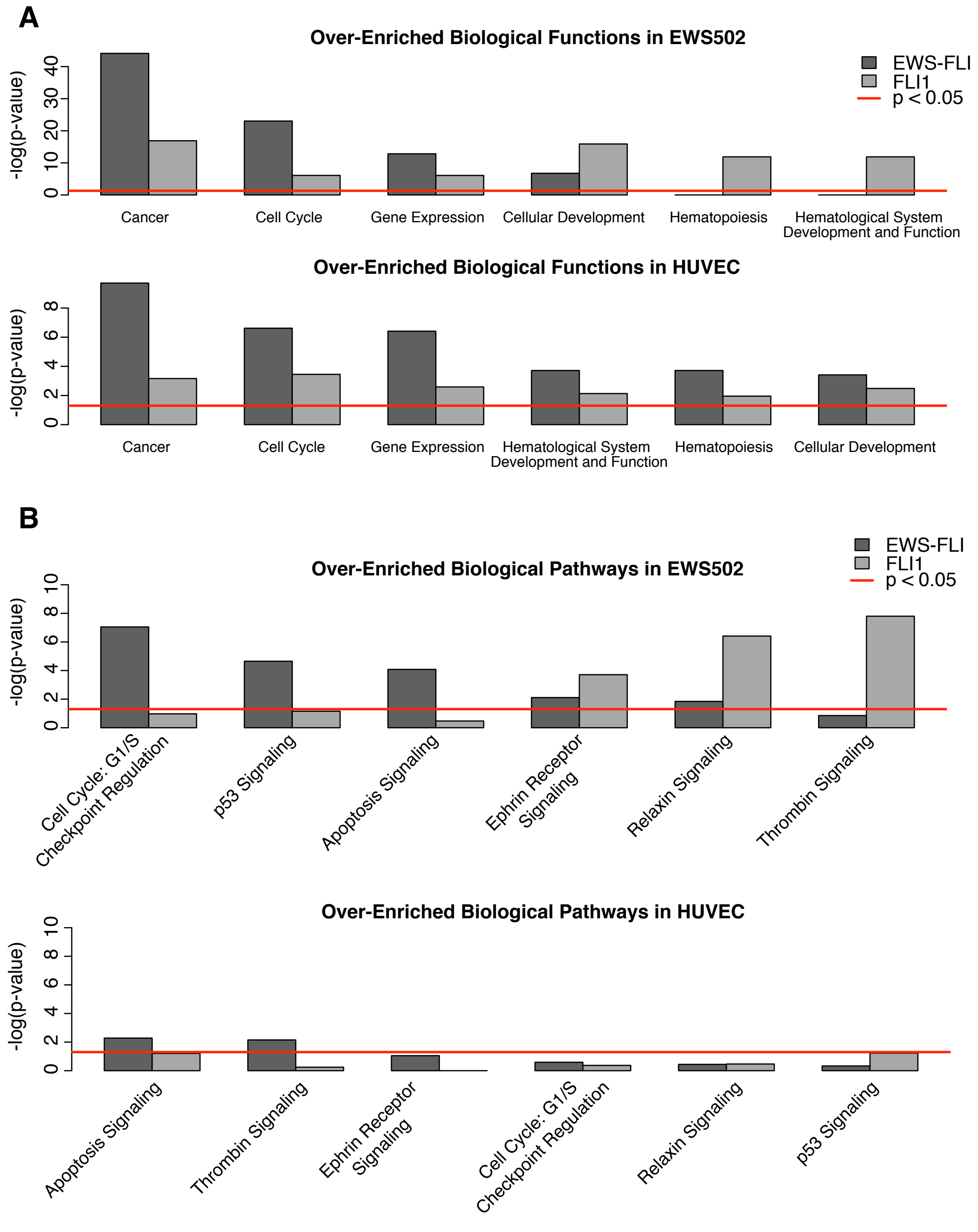
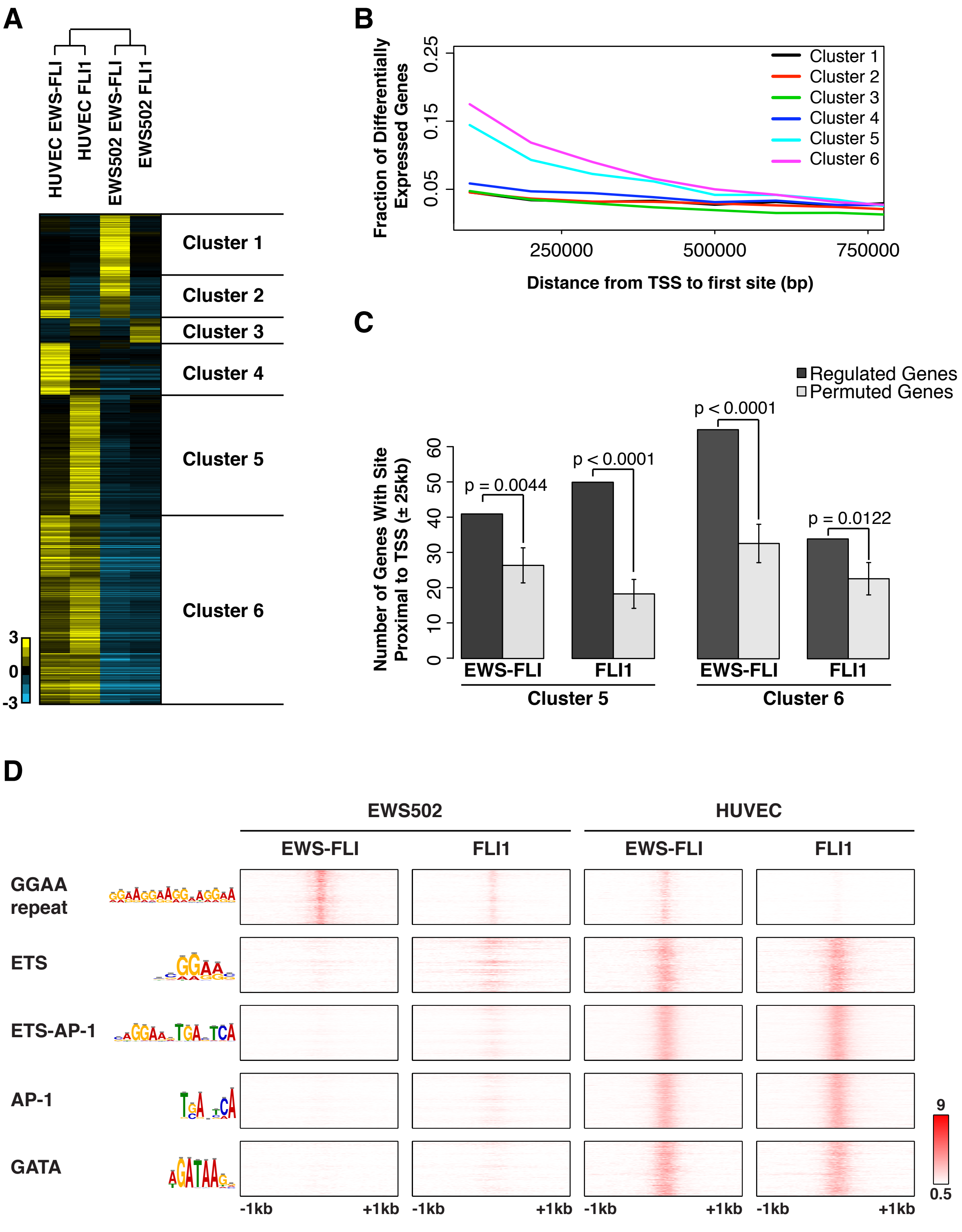


Figure 2.7: EWS-FLI and FLI1 differentially expressed genes have distinct biological functions. A. Over-enriched biological functions and B. Over-enriched biological pathways identified by Ingenuity Pathway Analysis indicate that EWS-FLI activates cancer-related genes and pathways, while FLI1 activates genes involved in normal endothelial growth in both cell types. Significance line ($p < 0.05$) is drawn in red.

dependence, with cell type being the primary determinant for the majority of differentially bound sites. Sites in clusters 1–3 revealed higher signals in tumor cells whereas those in clusters 4–6 (representing 74% of the sites) were enriched in endothelial cells. The finding of HUVEC-specific clusters bound by both EWS-FLI and FLI1 further supports that in a normal cellular environment these transcription factors share similar targeting.

Testing for associations between each cluster and gene expression demonstrated that sites in clusters 5 and 6 tended to be located near the union set of differentially expressed genes from both cell types (Figure 2.8B). Approximately 15% of differentially regulated genes contained at least one of these sites within 25 kb. Furthermore, genes that contained a TSS flanked by a site in clusters 5 or 6 (within 25 kb) were significantly more likely to be regulated by the expression of either EWS-FLI or FLI1 in HUVEC (Figure 2.8C). Genes harboring cluster 6 sites were frequently upregulated (82% and 88% for EWS-FLI and FLI1, respectively), however, genes proximal to cluster 5 sites lacked this skew toward upregulation. Interestingly, these data suggest that although FLI1 targets both cluster 5 and 6 sites equally, the potential occupancy of EWS-FLI at these sites characterizes functionally distinct elements. Since the sequence composition of clusters 5 and 6 were indistinguishable (see below), it is possible that chromatin differences that permit EWS-FLI binding also favor enhancer activity.

Using the Genomic Regions Enrichment of Annotations Tool (GREAT; [118]), we observed that regions defined by these clusters were strongly associated with specific biologically relevant ontologies independent of our gene expression data (Figure 2.10). In support of a direct regulatory role, EWS-FLI-specific binding sites (clusters 1 and 2) were significantly associated with genes regulated in cells engineered to express EWS-FLI. Interestingly, cluster 2 which demonstrated binding in both HUVEC and EWS502 was associated with genes involved in mesodermal and craniofacial development whereas sites in clusters 5 and 6 (specific for HUVEC), were largely associated with genes known to play roles in vascular development and those responsive to hypoxia and HIF1A overexpression. These data suggest that cellular


A
HUVEC EWS-FLI
HUVEC FLI1
EWS502 EWS-FLI
EWS502 FLI1
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
3
0
-3
B
Fraction of Differentially Expressed Genes
0.05
0.15
0.25
250000
500000
750000
Distance from TSS to first site (bp)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
C
Number of Genes With Site Proximal to TSS (± 25kb)
0
10
20
30
40
50
60
■ Regulated Genes
□ Permuted Genes
p = 0.0044
p < 0.0001
p < 0.0001
p = 0.0122
EWS-FLI
FLI1
EWS-FLI
FLI1
Cluster 5
Cluster 6
D
EWS502
HUVEC
EWS-FLI
FLI1
EWS-FLI
FLI1
GGAA repeat
ETS
ETS-AP-1
AP-1
GATA
-1kb
+1kb
-1kb
+1kb
-1kb
+1kb
-1kb
+1kb
9
0.5

ontology influences genomic targeting and corroborate our expression-based gene ontology observations (Figure 2.7).

To identify features of the differentially bound regions, comparative *de novo* motif detection for each cluster was performed using the 200 bp region surrounding the maximal signal for each region [119]. Identified motifs were matched to known motifs in TRANSFAC using TOMTOM [120]. 94% of sites in clusters 1 and 2, which exhibited an EWS-FLI-specific pattern in the tumor cells, contained a tetranucleotide repeat harboring the core of the ETS binding site (Figure 2.8D and see below). Sequence elements identified in clusters 3–6 were more highly varied commonly containing the canonical ETS motif.

2.2.4 Chimerism retargets EWS-FLI to tandem tetranucleotide repeats

De novo motif detection on the sequences uniquely bound by EWS-FLI in sarcoma cells represented in clusters 1 and 2 identified a GGAA-containing tetranucleotide microsatellite repeat. EWS-FLI binding to these sequences had been observed in recent studies [111] [112] [113]. The number of tandem repeats bound by EWS-FLI was higher than expected by chance in both cell types, although tumor cells demonstrated greater enrichment (Figure 2.9A). Examination of perfect sequential repeats revealed maximum enrichment at approximately 14

Figure 2.8: Hierarchical clustering identifies cell- and transcription factor-specific variation in genomic targeting. A. Hierarchical clustering of 6,525 binding sites that exhibited the widest variation in signal across transcription factors or cell types. Each row was median-centered and colored based on the average read count across the region. B. Distance (bp) from the transcriptional start site of the union set of differentially expressed genes to the closet site from clusters 1–6. C. Number of EWS-FLI or FLI1 differentially expressed genes in HUVEC containing a cluster 5 (left) or cluster 6 (right) site within 25kb of its TSS, compared to 10,000 permutations of all RefSeq genes. D. Normalized $\log_2$ ChIP-seq signal around the midpoint of identified *de novo* transcription factor motifs derived from the sequences underlying sites in each cluster. Clusters 1 and 2 were merged for the composite GGAA microsatellite motif (1,362 rows). Clusters 3–6 were merged for ETS (682 rows), ETS-AP1 (2,780 rows), AP1 (1,903 rows), and GATA (812 rows). Color was assigned on a $\log_2$ scale from 0.5 to 9.

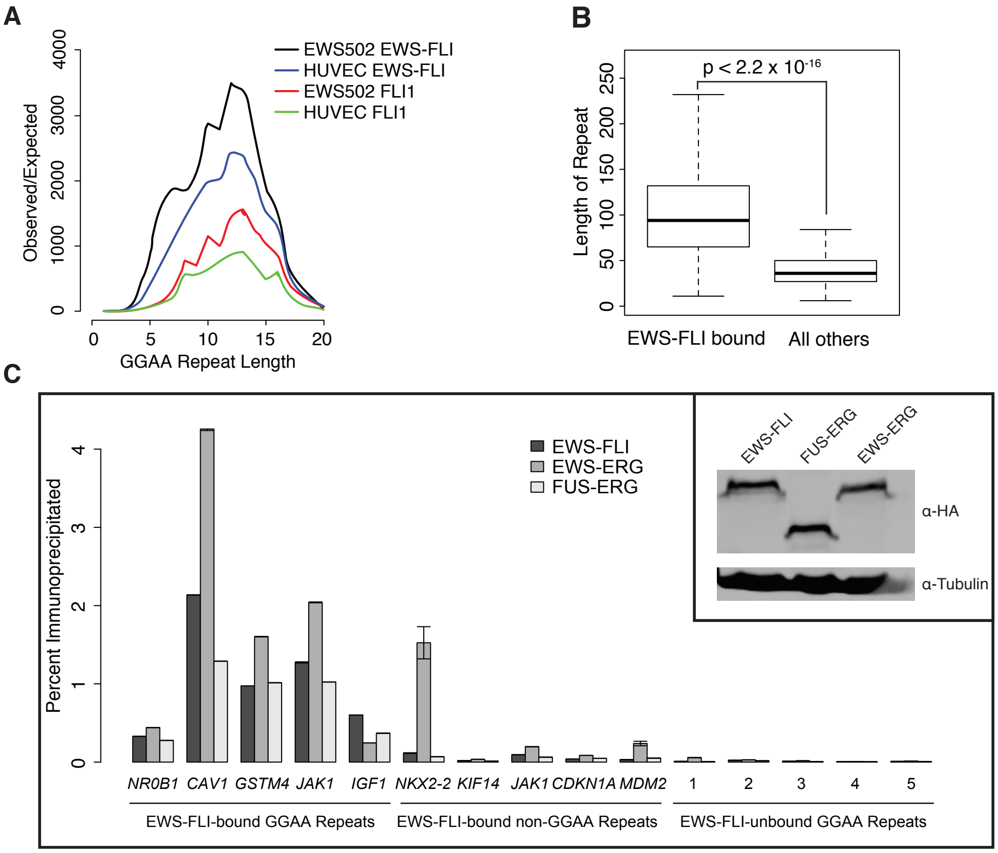
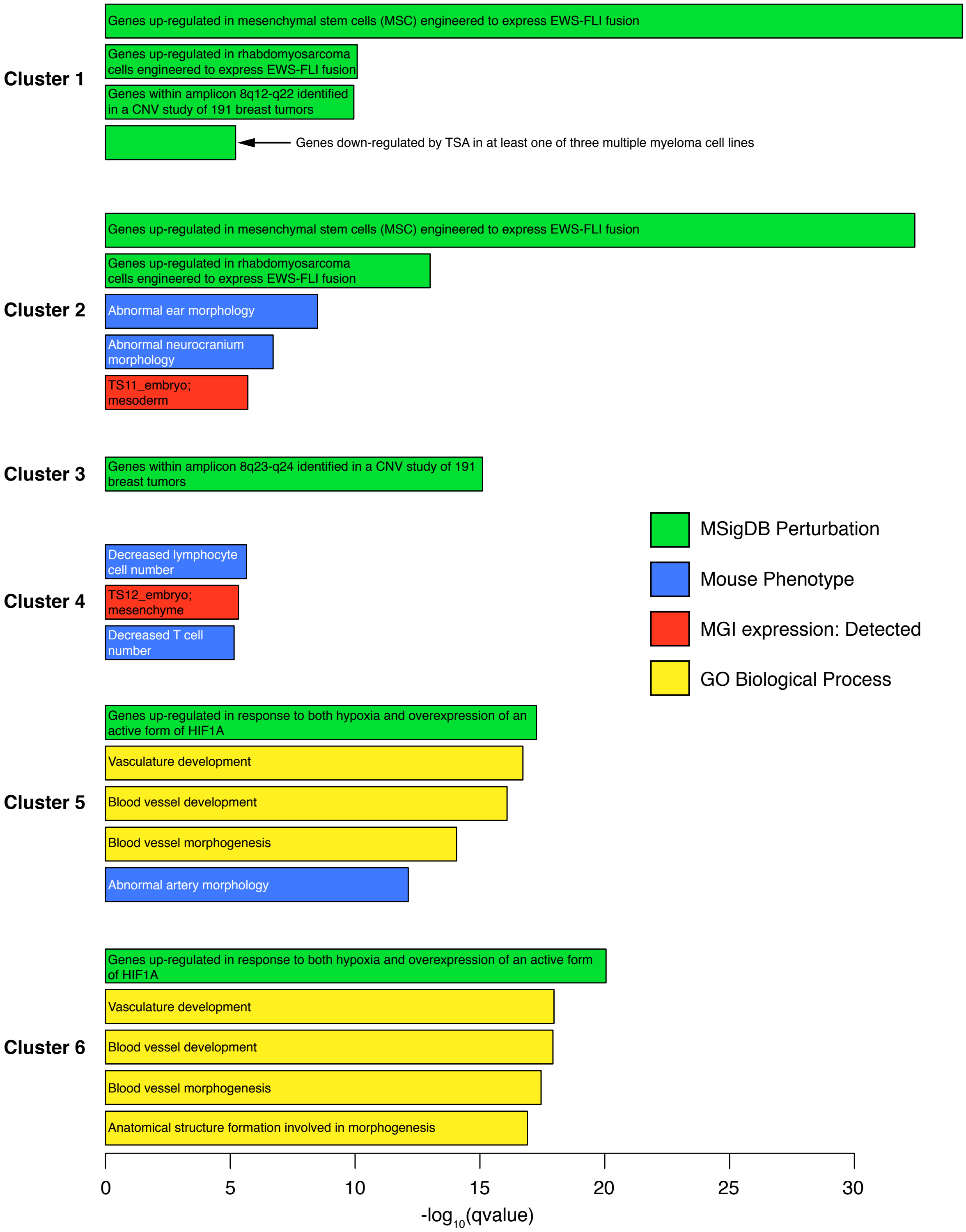


Figure 2.9: EWS-ETS fusions target GGAA-containing microsatellite repeats. A. Tandem GGAA repeats identified in EWS-FLI and FLI1 binding sites in EWS502 and HUVEC were compared to those detected by 1000 permutations of the identical number of regions over the mappable genome, maintaining chromosomal distribution. All lengths exceeding one repeat were significant to $p < 0.0001$. To permit plotting lengths for which the permuted value was zero, 0.1 was added to each observed and expected value. B. The lengths of repeat regions annotated by RepeatMasker bound by EWS-FLI in EWS502 were compared to those unbound in mappable regions of the genome. Regions bound by EWS-FLI contained significantly longer repeats as measured by t-test. C. ChIP-qPCR on chromatin isolated from EWS502 cells expressing the various Ewing Sarcoma fusions. Results are shown as a percent of input control. Overall, greater binding is identified to EWS-FLI bound regions near differentially expressed genes that contained GGAA repeats (*NR0B1*, *CAV1*, *GSTM4*, *JAK1*, *IGF1*) compared to those that bound EWS-FLI but did not harbor a repeat (*NKX2-2*, *KIF14*, *JAK1*, *CDKN1A*, *MDM2*). Five control repeat-containing regions are included, and error bars represent standard error of three replicates. (Inset) Western blot showing exogenous expression of HA-EWS-FLI, HA-FUS-ERG, and HA-EWS-ERG in EWS502 cells. Tubulin serves as a loading control.


Cluster 1
Genes up-regulated in mesenchymal stem cells (MSC) engineered to express EWS-FLI fusion
Genes up-regulated in rhabdomyosarcoma cells engineered to express EWS-FLI fusion
Genes within amplicon 8q12-q22 identified in a CNV study of 191 breast tumors
Genes down-regulated by TSA in at least one of three multiple myeloma cell lines
Cluster 2
Genes up-regulated in mesenchymal stem cells (MSC) engineered to express EWS-FLI fusion
Genes up-regulated in rhabdomyosarcoma cells engineered to express EWS-FLI fusion
Abnormal ear morphology
Abnormal neurocranium morphology
TS11_embryo; mesoderm
Cluster 3
Genes within amplicon 8q23-q24 identified in a CNV study of 191 breast tumors
Cluster 4
Decreased lymphocyte cell number
TS12_embryo; mesenchyme
Decreased T cell number
Cluster 5
Genes up-regulated in response to both hypoxia and overexpression of an active form of HIF1A
Vasculature development
Blood vessel development
Blood vessel morphogenesis
Abnormal artery morphology
Cluster 6
Genes up-regulated in response to both hypoxia and overexpression of an active form of HIF1A
Vasculature development
Blood vessel development
Blood vessel morphogenesis
Anatomical structure formation involved in morphogenesis
🟩 MSigDB Perturbation
🟦 Mouse Phenotype
🟥 MGI expression: Detected
🟨 GO Biological Process
0
5
10
15
20
25
30
-log10(qvalue)

tandem elements. Periodicity in the length of enriched repeats was observed with a preference for 8, 12 and 14 repeat units. GGAA-containing repeats as annotated by RepeatMasker that were bound by EWS-FLI in EWS502 were significantly longer than those not bound, with a median length of 100 bp (Figure 2.9B). The difference in lengths reflects the analysis of either perfect or imperfect repeats. Unexpectedly, FLI1 also bound these repeats, although with much lower frequency in both cell types, suggesting that the ability to target these sites is not exclusive to EWS-FLI but rather reflects the enhancement of a native characteristic.

We directly compared the binding of EWS-FLI and other fusions characteristic of Ewing Sarcoma at the tetranucleotide repeat-containing sites with sites containing the canonical high affinity ETS motif. The TET-ETS fusions EWS-ERG and FUS-ERG [87] [121] were epitope-tagged and expressed at similar levels as endogenous EWS-FLI in conjunction with EWS-FLI silencing (Figure 2.9C, inset). All fusion proteins tested demonstrated a greater enrichment at sites containing tandem repeats than canonical high affinity sites (Figure 2.9C). These data corroborate EWS-ERG ChIP [113] and support the general property of TET-ETS fusions to occupy these elements in a chromatinized genomic context. Moreover, the data suggest that repeat-containing sites are more likely to be bound than the canonical sites.

In light of recent studies suggesting a length requirement for microsatellite enhancer function [114] [111], we examined EWS-FLI and FLI1 sites containing five or more repeats. Approximately 30% of regions uniquely bound by EWS-FLI in either cell type contained these long tandem repeats, whereas they were present in only 0.2% and 0.04% of regions unique to FLI1 in EWS502 and HUVEC, respectively. In agreement with the previous studies, we found that these GGAA repeats were more proximally located to genes upregulated

Figure 2.10: Annotation of clusters of binding sites using GREAT shows biologically relevant associations with ontologies. Only those terms with FDR-corrected q-values more significant than 10^{-5} and in the top 5 significant terms are shown. Bars are color-coded by the ontology from which they were derived (MSigDB Perturbation, green; Mouse Phenotype, blue; MGI expression: Detected, red; GO Biological Process, yellow) and statistical significance is expressed as $-\log_{10}$(q-value).

by EWS-FLI (Figure 2.11) with FLI1 exhibiting a similar trend (data not shown). FLI1 bound more proximally to FLI1-modulated genes compared to EWS-FLI around EWS-FLI-regulated genes (Figure 2.12), suggesting that in the context of chromatin, tetranucleotide repeats may function primarily as transcriptional enhancers and can be located distally from genes.

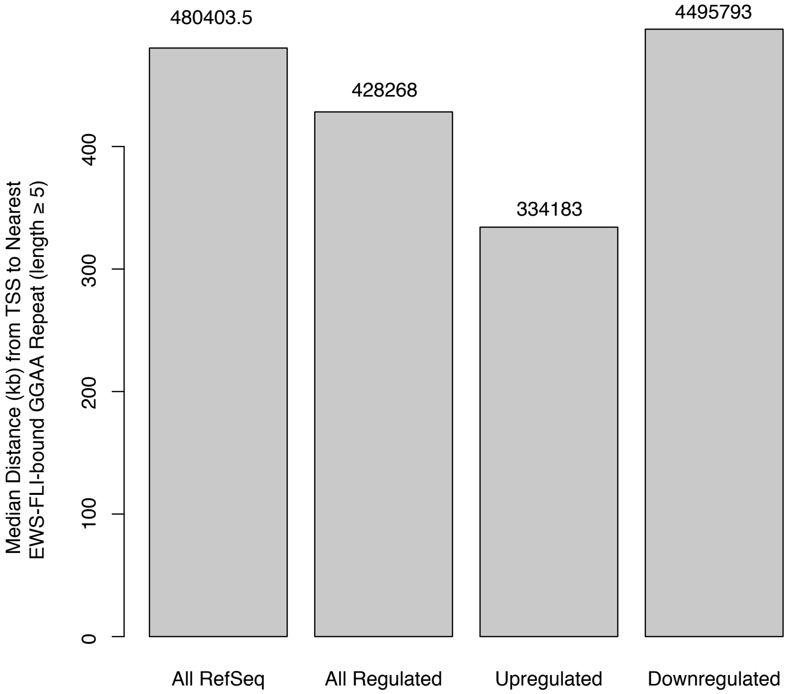


Figure 2.11: Upregulated genes are closer to EWS-FLI binding sites. Median distance from TSS of all genes, all differentially expressed, all upregulated, and all downregulated RefSeq genes to the nearest EWS-FLI-bound GGAA repeat with length greater than or equal to 5.

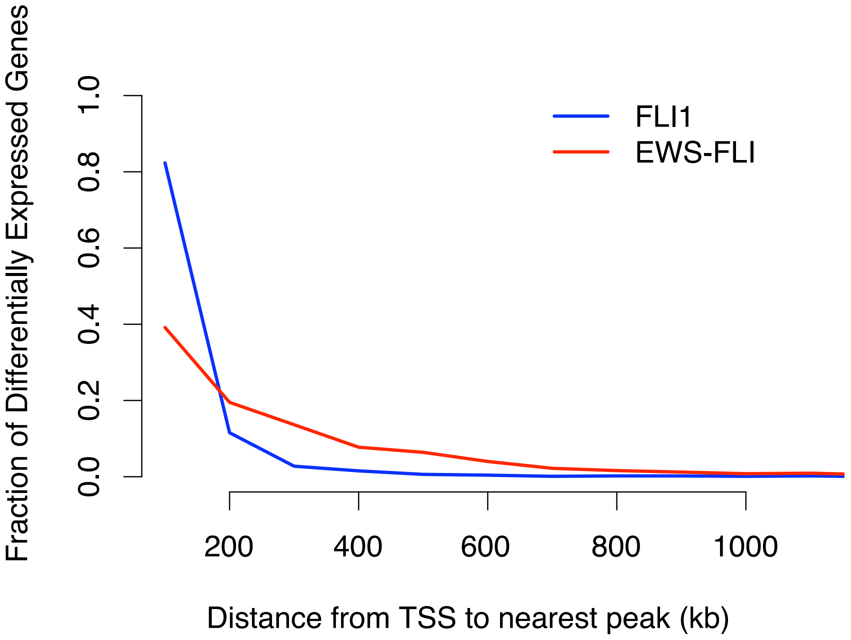


Figure 2.12: FLI1 binding sites are closer to FLI1 differentially expressed genes. Distance from the TSS of a gene differentially expressed by FLI1 (blue) or EWS-FLI to the nearest FLI1 or EWS-FLI binding site, respectively. The fraction of genes containing at least one site within the denoted distance is presented.

Preference of EWS-FLI for sites containing tetranucleotide repeats may lead to selection of, extended polymorphic repeats during tumor development such that their actual length would differ from the reference genome or other cell types. Previous studies that examined a small number of randomly selected tetranucleotide repeats failed to demonstrate a difference between tumor cells and the reference genome [111] [112]. We compared the lengths of several tetranucleotide repeats occupied by EWS-FLI in tumor cells with the same regions in HUVEC and the reference genome (hg18). Regions amenable to evaluation were limited due to the challenges inherent in primer design for repetitive regions. However, one region, located in an intron of *IGF1*, exhibited mono- or biallelic presence of a sequence longer than that predicted by the reference genome in 4 of 7 Ewing cell lines (Figure 2.13A). Sequencing of this region from EWS502 cells confirmed that the difference resulted from nine additional repeats (Figure 2.13B). Interestingly, we observed an extremely faint band running at approximately the same molecular size as the expanded region in the pooled endothelial cells. It is possible that expansion (relative to the reference genome) represents an allelic variant present in the population.

2.2.5 Combinatorial DNA binding motifs distinguish endothelial cell targeted sites

Canonical ETS motifs were identified in 72% of cluster 3 (which was largely specific to FLI1 in tumor and endothelial cells) and clusters 4–6 (which specific to endothelial cells but bound by both proteins) sites. ETS motifs in cluster 3 demonstrate that FLI1 retains the ability for context-dependent targeting even in sarcoma cells. Strikingly, the motif for the AP1 complex was detected at nearly the same frequency as ETS at cluster 3–6 sites. Remarkably, of the sites containing a computationally derived AP1 motif, 76% overlapped ChIP-seq-derived binding sites for c-Jun, a member of the AP1 complex, in HUVEC [122]. In addition to isolated AP1 motifs, composite ETS and AP1 motifs were observed at approximately 46% of the sites in clusters 3–6. We explored variation in the spacing between the ETS and AP1 motifs; approximately 25% of the sites revealed separation of 1 bp with spacing increments of 2 to 10 bp each accounting for an additional 9 to 12% of sites. The composite nature of this ETS-AP1 motif is suggestive of cooperative binding at these sites. The GATA motif was also observed in approximately 15% of sites from clusters 3–6. The ETS motif found within these

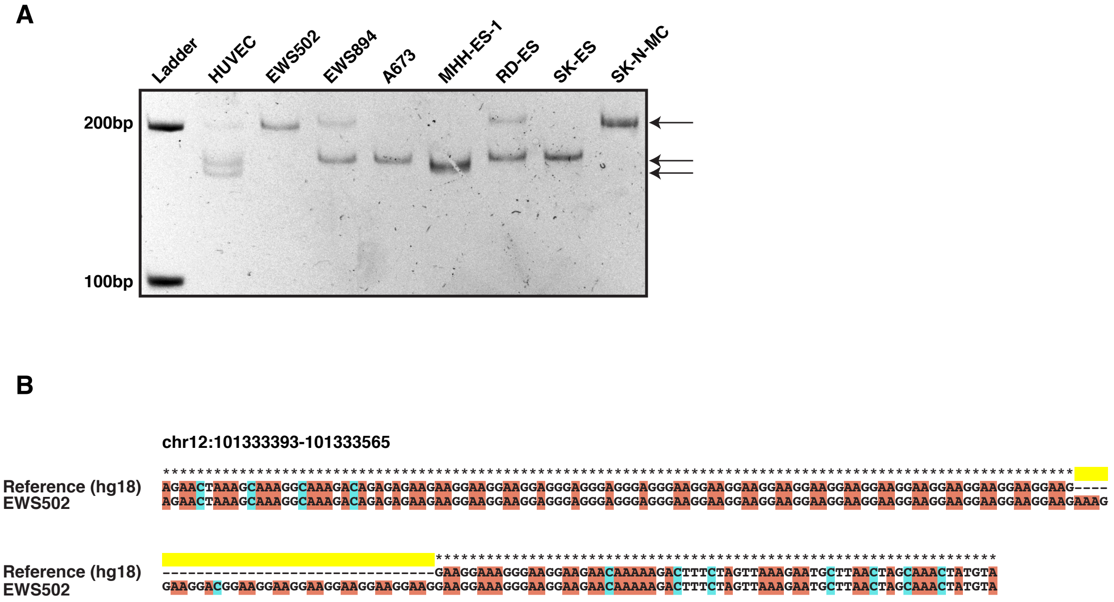

A
Ladder
HUVEC
EWS502
EWS894
A673
MHH-ES-1
RD-ES
SK-ES
SK-N-MC
200bp
100bp
B
chr12:101333393-101333565
Reference (hg18)
EWS502
AGAACTAAAGCAAAGGCAAAGACAGAGAGAAGAAGGAAGGAAGGAGGGAGGGAGGGAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAG----
AGAACTAAAGCAAAGGCAAAGACAGAGAGAAGAAGGAAGGAAGGAGGGAGGGAGGGAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGAAAG
Reference (hg18)
EWS502
--------------------------------GAAGGAAAGGGAAGGAAGAACAAAAAGACTTTCTAGTTAAAGAATGCTTAACTAGCAAACTATGTA
GAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGGGAAGGAAGAACAAAAAGACTTTCTAGTTAAAGAATGCTTAACTAGCAAACTATGTA

Figure 2.13: EWS-FLI-bound tetranucleotide repeats demonstrate repeat length polymorphism. Length of repeat found within IGF1 across 7 Ewing Sarcoma cell lines (EWS502, EWS894, A673, MHH-ES-1, RD-ES, SK-ES, SK-N-MC) and compared to endothelial cells (HUVEC). Lengths determined by PCR using primers flanking repetitive region and resolved on an 8% acrylamide gel. B. Sequence of repeat region from EWS502 cells compared to reference genomic sequence (hg18). Multiple sequence alignment was performed using ClustalX with default parameters. Exact sequence matches denoted by "*", regions of difference highlighted in yellow.

regions, demonstrated context specific sequence variation. ETS sites in isolation typically contained a C preceding the GGAA core, matching the canonical ETS motif [113]. However, ETS motifs in composite sites with AP1 were preceded by an A. Similar motif variation had been observed in the tandem binding sites of ETS-1 with RUNX1 [123].

We then compared the intensity, location, and specificity of regions containing consensus binding sites by plotting normalized ChIP-seq signals around a union set of sites that share a specific motif (Figure 2.8D). Tetranucleotide repeats were preferentially and centrally bound by EWS-FLI. In the tumor cells, FLI1 demonstrated a modest ability to interact with some of these sites as previously noted (Figure 2.9A). In endothelial cells, the signal intensities and positions of FLI1 and EWS-FLI were similar around ETS, ETS-AP1, AP1, and GATA motif-containing sites. In the tumor cells, FLI1 bound these sites although with far less signal intensity, again demonstrating the tendency of FLI1 to function normally in tumor cells. Since *de novo* motif identification may preclude detection of less common motifs, we examined signals from HUVEC around sites containing computationally-predicted motifs for Myc:Max, NFB, STAT, PPAR, HNF4A, and CREB [124]. Although these sites represented less than 1% of those analyzed, similar patterns of EWS-FLI and FLI1 signal were detected (Figure 2.14). All motif associations were lost when the sites were permuted (Figure 2.15). These data suggest a large network of cooperative interactions for FLI1 binding, most frequently AP1 and GATA. Sites selective for EWS-FLI occupancy in sarcoma cells were distinguished in function, location, and composition from those sites that characterize endothelial targeting.

2.2.6 Epigenetic factors distinguish microsatellite repeats in Ewing Sarcoma

Since cell lineage dominated the variation in targeting of both chimeric and parental transcription factors, we explored features of epigenetics and chromatin configuration that could underlie these differences. We performed ChIP-seq on Ewing Sarcoma cells for histone marks associated with active (H3K4me1, H3K4me2, and H3K4me3) or repressed (H3K27me3) chromatin. We also performed Formaldehyde Assisted Isolation of Regulatory Elements coupled with next-generation sequencing (FAIRE-seq) to identify regions of nucleosome depletion that characterize active regulatory elements. Consistent with other cell types, we found that both methylation on H3K4 and FAIRE demonstrated a strong association with gene expression, whereas H3K27me3 was inversely correlated with transcription (Figure 2.16A).

Deregulation of homeobox genes is a common attribute of cancer [125]. We compared the epigenetic status of the four *HOX* clusters comparing it to the patterns from embryonic

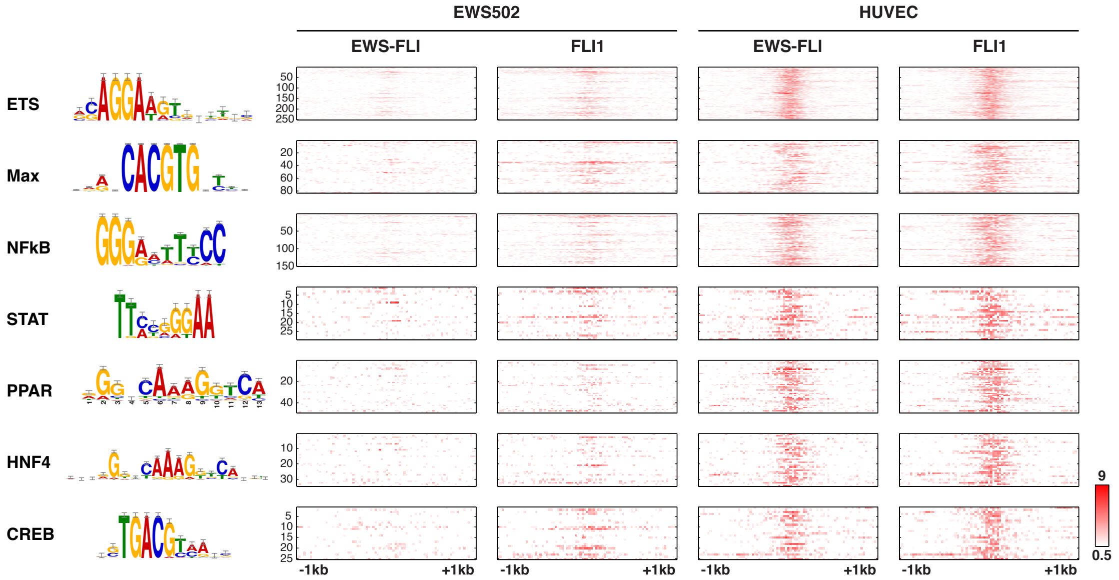


Figure 2.14: EWS-FLI and FLI1 occupied similar sites in a normal cellular context. Heatmap showing normalized ChIP-seq signals of EWS-FLI or FLI1 in both Ewing Sarcoma cells and HUVECs around computationally predicted transcription factor binding sites of ETS, Max, NFkB, STAT, PPAR, HNF4, and CREB. Sequence logos corresponding to the computationally predicted motif are shown on the left. Color was assigned on a $\log_2$ scale from 0.5 to 9.

stem cells and HUVEC [122] (Figure 2.17). Interestingly, at the *HOXA* cluster, we detected a bivalent signal similar to that observed in embryonic stem cells [126]. At the other *HOX* clusters however, there was an overall lack of H3K27me3 and enrichment for H3K4me2 and H3K4me3. This activation was not specific for a set of homeobox genes, as seen in differentiated cells (HUVEC). Widespread activation across nearly all genes in each cluster may contribute to the dedifferentiated or more stem-like state of Ewing sarcomas. The loss of H3K27me3 signal was not observed genome-wide (e.g. *PAX2* and *WNT3A* loci, Figure 2.18).

We focused our analysis on regions containing the tetranucleotide repeats since they represented the most prominent feature distinguishing EWS-FLI targeting. Comparing the epigenetic and chromatin status of the repeats in Ewing Sarcoma cells with HUVEC and other

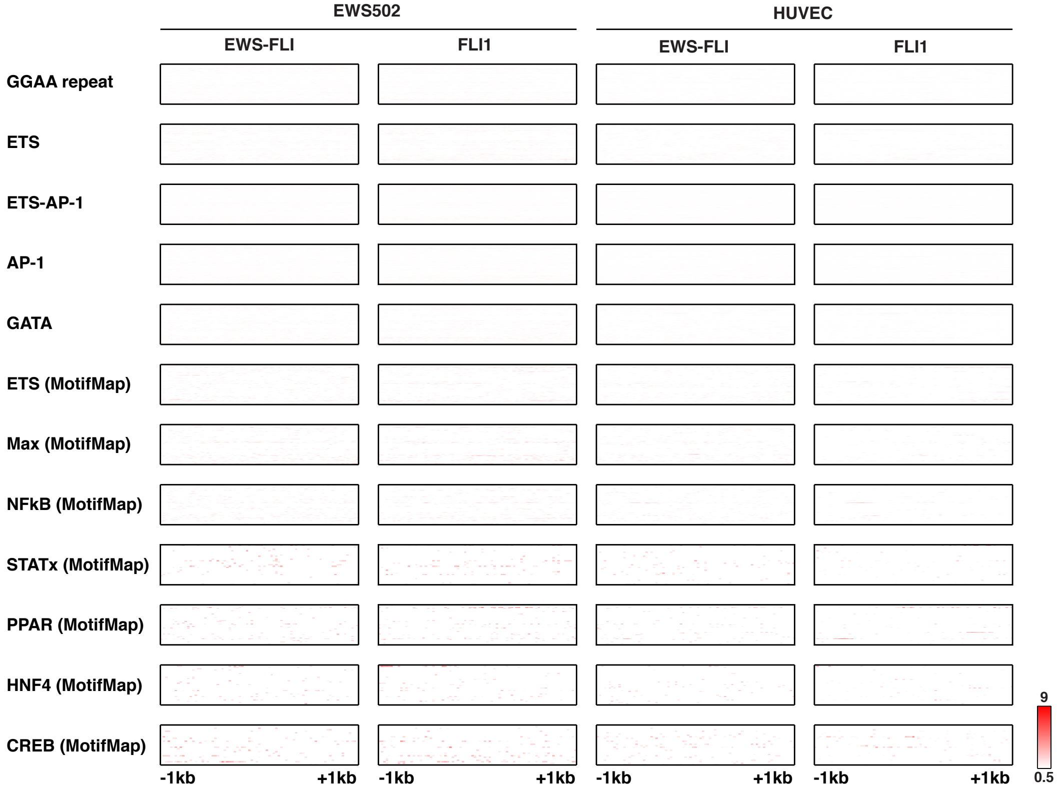
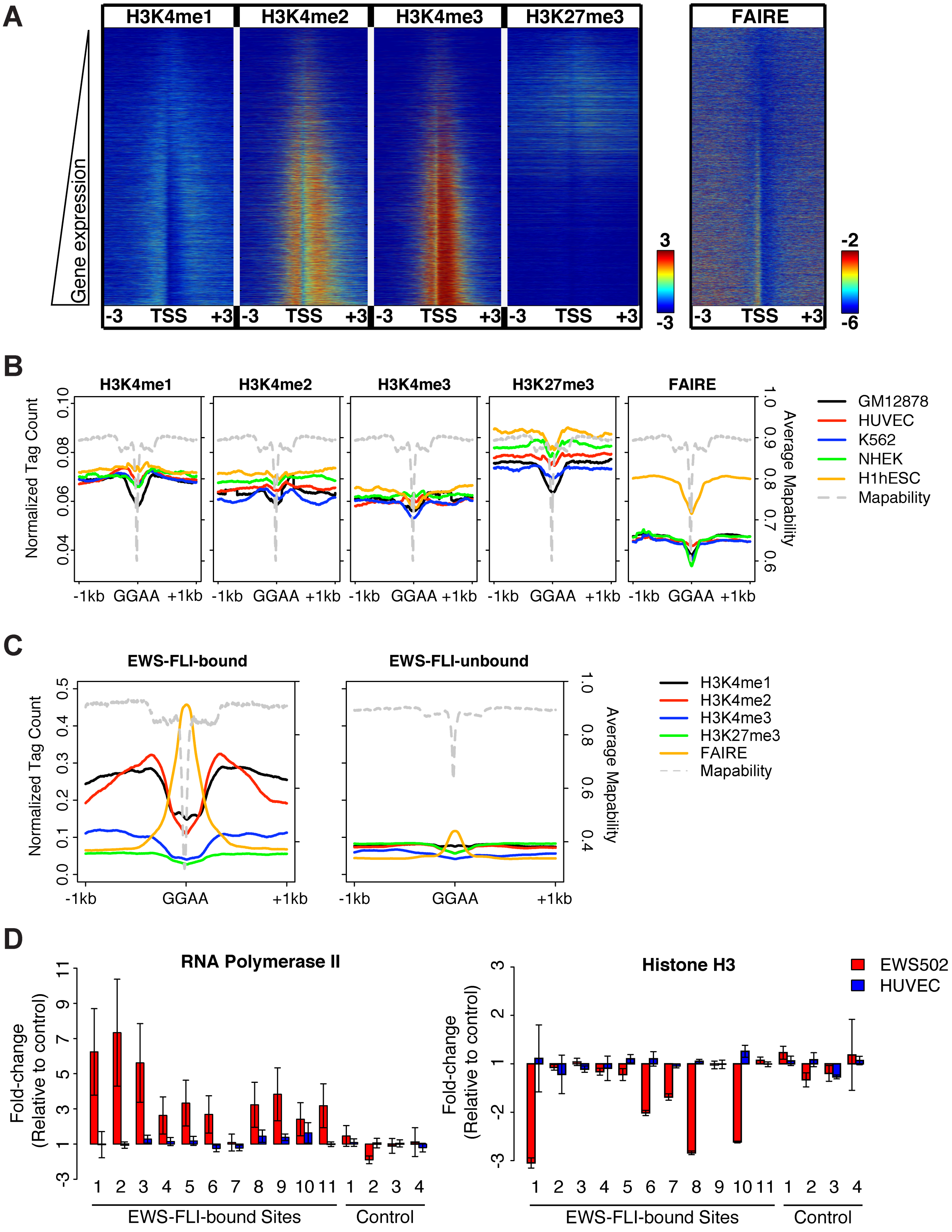


Figure 2.15: Permutation abolished EWS-FLI and FLI1 signal enrichment at all identified sites. One permutation of GGAA repeats, ETS, ETS-AP-1, AP-1, GATA, as well as the 7 computationally predicted motifs presented in Figure 2.14 shows the observed pattern is non-random. Color was assigned on a $\log_2$ scale from 0.5 to 9.

cell lines assayed as part of the ENCODE consortium, we observed that in virtually all cell types repressive marks were common at microsatellite and other repetitive elements (Figure 2.16B, Figure 2.19) [4] [127] [53]. In Ewing cells, however, strong H3K4me1 and H3K4me2 signals flanked those repeats that were bound by EWS-FLI relative to those that were not bound although the proportion of active histone marks that directly overlapped repeats was similar to other cell types (Figure 2.16C and Figure 2.19). Moreover, EWS-FLI-bound sites were largely devoid of H3K27me3 and were nucleosome depleted. Relative to HUVEC and control regions, these sites were also bound by RNA polymerase II (Figure 2.16D). To confirm nucleosome depletion, we also performed pan-histone H3 ChIP, which demonstrated overall histone H3 depletion at several sites (Figure 2.16D). Intriguingly, other classes of repetitive


A
H3K4me1
H3K4me2
H3K4me3
H3K27me3
FAIRE
Gene expression
-3 TSS +3
3
-3
-2
-6
B
H3K4me1
H3K4me2
H3K4me3
H3K27me3
FAIRE
Normalized Tag Count
0.04
0.06
0.08
0.10
Average Mapability
0.6
0.7
0.8
0.9
1.0
-1kb GGAA +1kb
GM12878
HUVEC
K562
NHEK
H1hESC
Mapability
C
EWS-FLI-bound
EWS-FLI-unbound
Normalized Tag Count
0.0
0.1
0.2
0.3
0.4
0.5
Average Mapability
0.4
0.6
0.8
1.0
-1kb
GGAA
+1kb
H3K4me1
H3K4me2
H3K4me3
H3K27me3
FAIRE
Mapability
D
RNA Polymerase II
Fold-change
(Relative to control)
-3
1
3
5
7
9
11
1 2 3 4 5 6 7 8 9 10 11 1 2 3 4
EWS-FLI-bound Sites
Control
Histone H3
-3
-2
1
2
3
EWS502
HUVEC

elements were also enriched by FAIRE, including SINEs, LINEs, and other types of simple and microsatellite repeats (data not shown). Together, these data support the utilization of the microsatellites as transcriptional enhancers. We also observed a striking difference in DNA sequence encoded nucleosome occupancy between EWS-FLI and FLI1-bound sites (Figure 2.20). Whereas FLI-bound sequences demonstrated high nucleosome occupancy, a feature of regulatory elements in higher eukaryotes [128], this pattern was not seen for EWS-FLI-bound regions, further supporting the unique attributes of EWS-FLI targeted domains.

2.2.7 EWS-FLI targets enhancer like elements altering and maintaining the local chromatin environment

In addition to analyzing microsatellite regions we also assessed the chromatin structure and epigenetic status of each class of regions identified by differential chromatin targeting (Figure 2.8A). These data further support that the Ewing-specific clusters (clusters 1–2) exhibit an epigenetic pattern resembling that of an enhancer element only in tumor cells (Figure 2.21A). Conversely, the endothelial-specific clusters (clusters 4–6) show an enhancer-like pattern unique to HUVEC. Distinct from the other clusters, sites in cluster 3, which exhibited binding by FLI1 in both cell types, were marked by H3K4me3. This epigenetic signature

Figure 2.16: Deregulation of repetitive elements in Ewing Sarcoma. A. Heatmap of normalized ChIP and FAIRE signal 3 kb around TSS ranked by gene expression in Ewing cells. Color was assigned on a $\log_2$ scale of -3 to 3 for ChIP and -6 to 2 for FAIRE. B. Normalized ChIP and FAIRE signals around the centers of GGAA repeats in five ENCODE cell lines (GM12878, black; HUVEC, red; K562, blue; NHEK, green; H1hESC, orange). Mapability of the underlying DNA sequence is represented on a scale of 0 (ambiguous) to 1 (unique) and is plotted in grey. C. Normalized ChIP and FAIRE signals around the centers of EWS-FLI-bound (left) or -unbound (right) GGAA repeats in Ewing Sarcoma cells. Mapability of the underlying DNA sequence is represented on a scale of 0 (ambiguous) to 1 (unique) and is plotted in grey. D. Enrichment of EWS-FLI-bound GGAA repeats for RNA Polymerase II (left) and histone H3 (right) in Ewing cells (red) and HUVEC (blue), as assayed by ChIP-qPCR. All values are represented as the fold-change relative to the average of the negative controls; fold-change values are centered on 1. Error bars represent the standard error from three technical replicates.

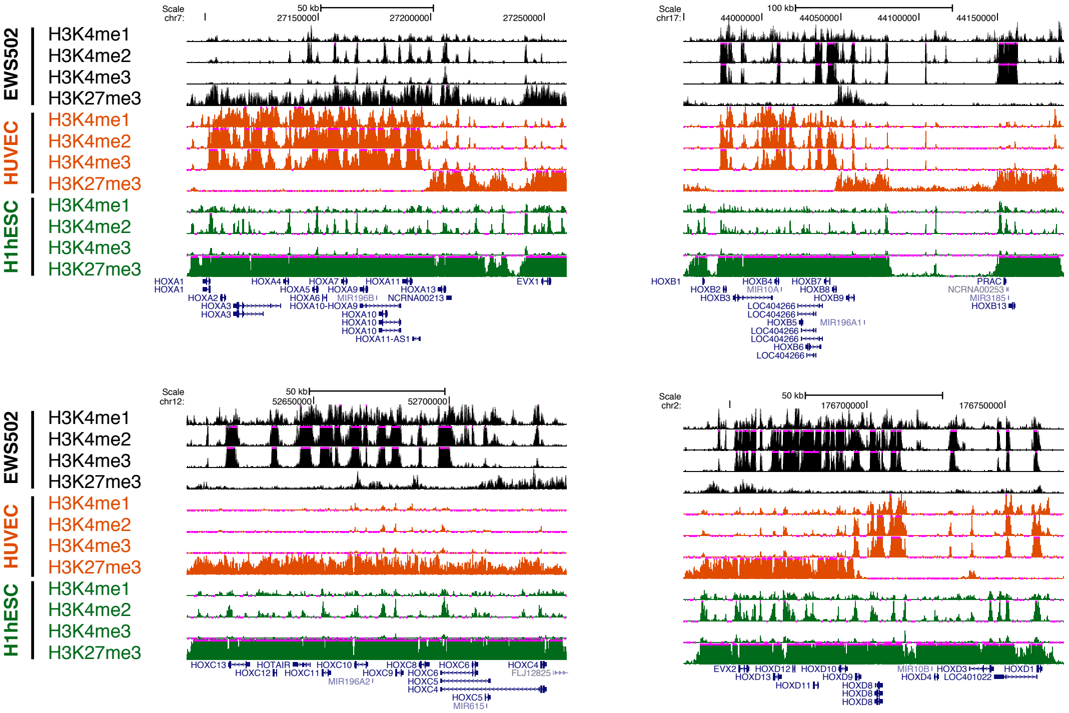


Figure 2.17: UCSC Genome Browser snapshots of epigenetic patterns in EWS502 (black), HUVEC (orange), and H1hESC (green). Three of the four HOX clusters show activation and aberrant H3K27me3 patterns relative to normal cell types. Values are presented on a scale of 0 to 50 reads.

suggests that FLI1-specific sites tend to be located proximal to transcriptional start sites.

Since EWS-FLI-specific chromatin domains were normally nucleosome occupied in HUVEC, we asked if EWS-FLI could directly alter chromatin conformation. EWS-FLI expression in HUVEC was associated with increased FAIRE enrichment (indicative of decreased nucleosome occupancy) at some of the closed chromatin domains, relative to control domains (Figure 2.21A). Moreover, silencing of EWS-FLI in tumor cells resulted in increased nucleosome occupancy at all sites tested. These findings suggest that chimerism confers nucleosome displacement activity, and continued EWS-FLI expression is required for the maintanence of an open chromatin configuration at these sites. This activity may be mediated through the recruitment of histone modifying enzymes, including histone methyltransferases and/or

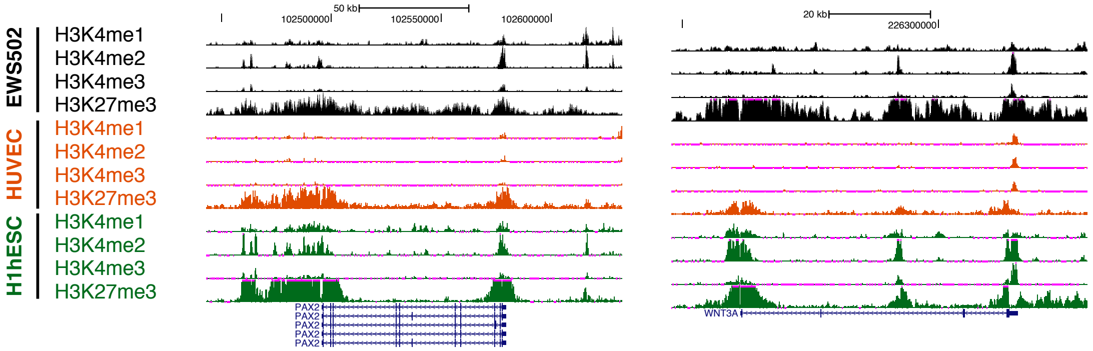


Figure 2.18: UCSC Genome Browser snapshots of epigenetic patterns in EWS502 (black), HUVEC (orange), and H1hESC (green). Normal H3K27me3 at PAX2 (left) and WNT3A (right), consistent with normal cell types. Values are presented on a scale of 0 to 50 reads.

demethylases.

2.3 Discussion

Lineage-specific outcomes are observed when chimeric transcription factors are expressed in various cell types, suggesting a major cell-specific influence on activity. One cell type may be permissive for transformation whereas other cells may not tolerate expression resulting in growth arrest or apoptosis. Cellular factors that influence activity may not be evident from studies of transcription factor-chromatin targeting limited to a single transcriptional regulator in a single cell type.

Employing an integrated genomic strategy to compare the oncoprotein EWS-FLI with its

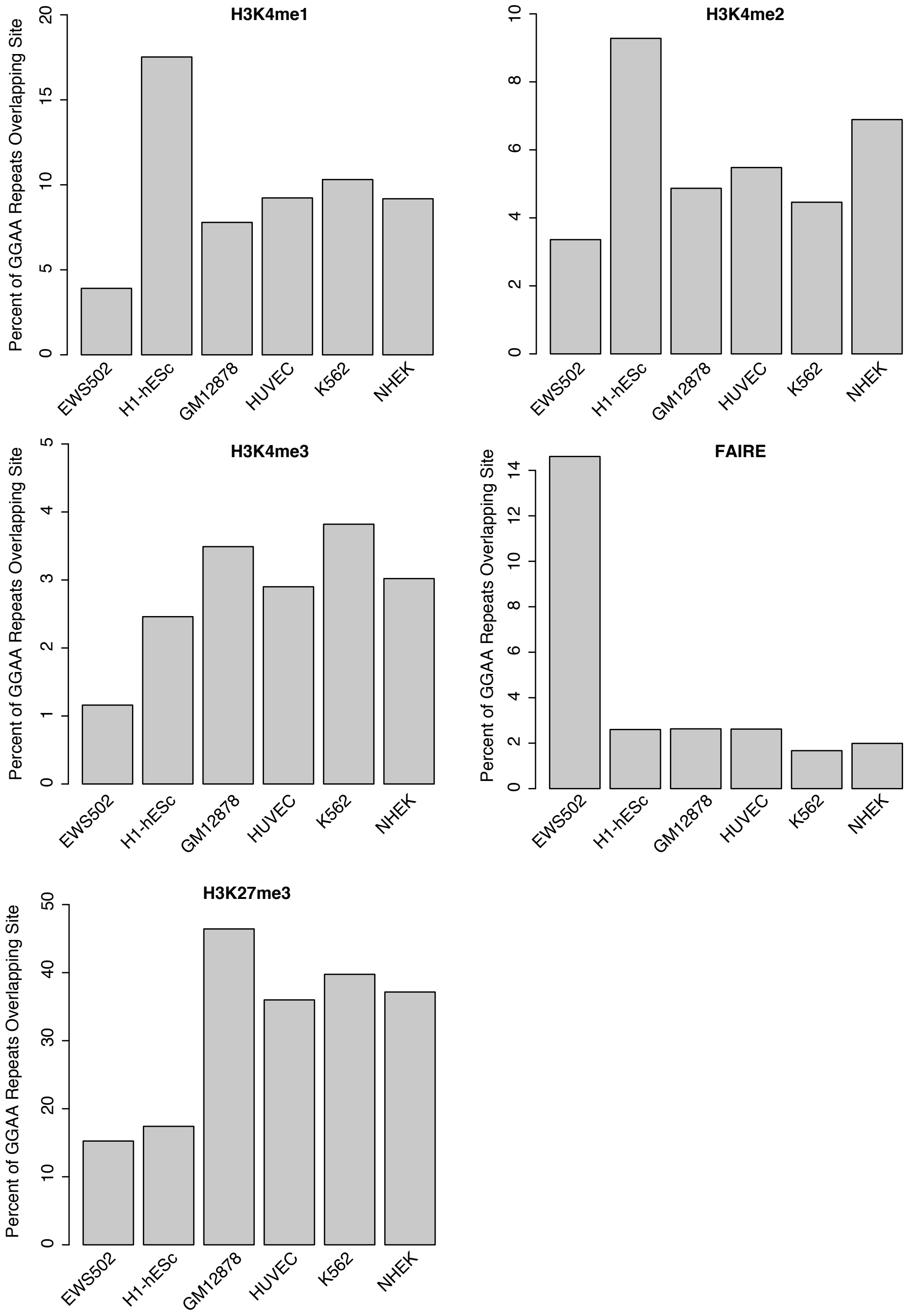


Figure 2.19: Epigenetics and chromatin accessibility of GGAA microsatellites. Percentage of GGAA repeats overlapping areas of significant enrichment for H3K4me1, H3K4me2, H3K4me3, H3K27me3, and FAIRE are presented for EWS502, HUVEC, H1hESC, K562, NHEK, and GM12878 cells.

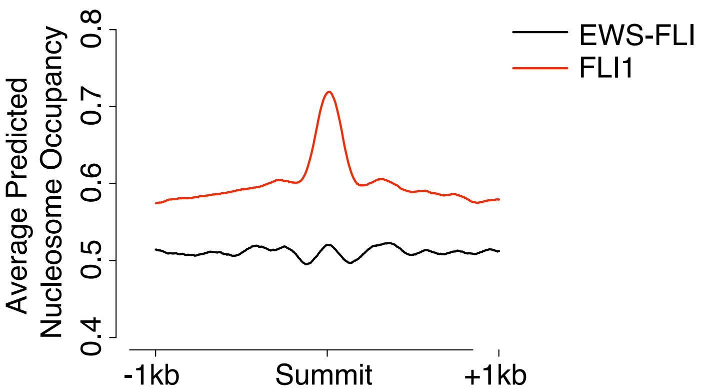


Figure 2.20: Predicted nucleosome occupancy of EWS-FLI and FLI1 binding sites. Average nucleosome occupancy predicted on DNA sequence [128] surrounding the summits of EWS-FLI and FLI1 binding sites in EWS502.

parental protein FLI1 in two relevant human cell types, we were able to separate the influence of chimerism and cell type on genomic targeting and function. In tumor cells, chimerism resulted in genomic retargeting, with approximately 40% of EWS-FLI binding sites containing a tetranucleotide repeat composed of the core ETS motif. Although FLI1 can bind to these repeats, the majority of FLI1 sites contained the canonical ETS motif. By contrast, in endothelial cells, targeting of both proteins demonstrated remarkable similarity. EWS-FLI and FLI1 localized to sites containing the canonical ETS motif as well as sites marked by AP1 and GATA motifs. Binding to a number of other less common DNA motifs suggests an extended network of interacting cooperative transcription factors. Given the abundance of ETS motif-containing sites in the genome, these interactions likely regulate cell lineage-specific patterns

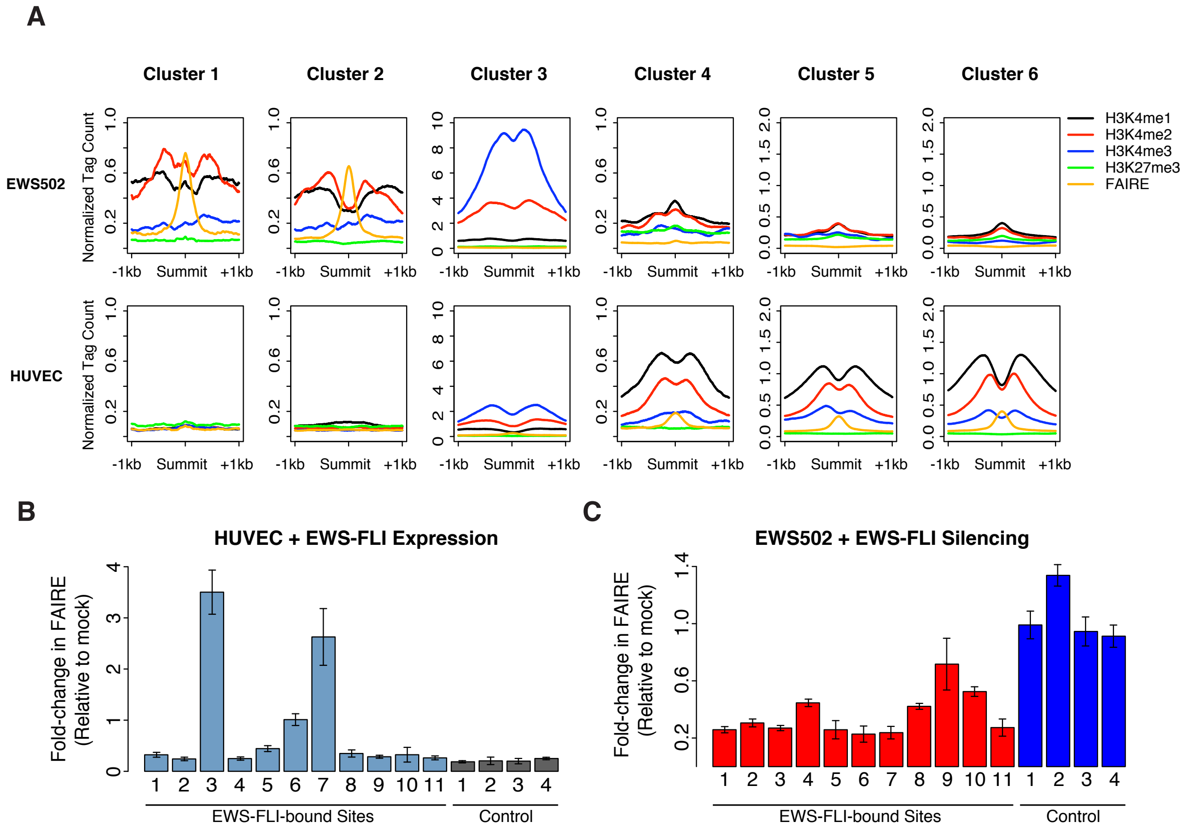


Figure 2.21: EWS-FLI is capable of epigenetic reprogramming. A. Normalized signals for H3K4me1 (black), H3K4me2 (red), H3K4me3 (blue), H3K27me3 (green), and FAIRE (orange) from both EWS502 and HUVEC are plotted for the 2 kb region surrounding the summits of sites identified by hierarchical clustering. B. Change in FAIRE enrichment at EWS-FLI-bound GGAA repeats following EWS-FLI expression in HUVEC. All values are represented as fold-change relative to scrambled shRNA control. Error bars represent the standard error from three technical replicates. C. Change in FAIRE enrichment at EWS-FLI-bound GGAA repeats following EWS-FLI silencing in EWS502. All values are represented as fold-change relative to scrambled shRNA control. Error bars represent the standard error from three technical replicates.

of genomic targeting. Although FLI1 can bind the tetranucleotide repeats both in vivo and in vitro, in reporter-based assays it fails to show activity these repeats [111] [129] suggesting that FLI1 requires the cooperation of other sequence specific transcription factors to activate transcription. The association of ETS proteins with AP1 had been observed [130] [131] [132], and the functional association of EWS-FLI or FLI1 with AP1 and GATA1 has been demonstrated in cellular transformation and hematopoietic development [132] [133] [134]. However, our data also suggest selectivity in cooperating transcription factors. Although studies of ETS-1 identified cooperative binding with RUNX1 [123] and PAX5 [135], neither relationship was evident in this study.

Differential targeting of EWS-FLI was influenced by epigenetic factors. EWS-FLI bound microsatellite regions in tumor cells that were atypically marked with an enhancer like signature, bound RNA polymerase II and resided in nucleosome depleted regions. Our data suggest that EWSR1 chimerism conferred nucleosome modification activity to EWS-FLI and is required for altering the local chromatin landscape resulting in nucleosome depletion or destabilization. However, the observation of widespread FAIRE enrichment of repetitive regions suggests that other factors may initially create a favorable chromatin arrangement permitting EWS-FLI targeting, a question currently being explored. The presence of RNA polymerase II suggests that these regions may be transcribed, a feature recently shown to be common among human epithelial cancers [136].

This ability of EWS-FLI to alter chromatin structure is similar to that of FOXA1 or GATA-4, which bind their cognate sites and affect chromatin configuration [137] [138] [139] [140] [141] [142]. Since EWS-FLI does not contain the conserved motif thought to be required for core histone interactions, its activity may be mediated through the recruitment of histone modifying enzymes. This mechanism is reminiscent of that of EVI1 or ZNF274, which can recruit histone methyltransferases involved in gene silencing [143] [144] and YY1, which can recruit a histone H4-specific methyltransferase leading to gene activation [145]. Interestingly ERG, another member of the ETS family has been shown to interact with the protein ESET which contains a SET domain and in turn recruits other chromatin remodeling factors such as HDAC1, HDAC2, and mSin3B [146] [147].

This study demonstrates the prospect of translational cancer genomics. The persistent "addiction" of the tumor to aberrant transcription offers a unique therapeutic opportunity. Consequently, genomic dysreguation through EWS-FLI-specific enhancers mediated by novel chromatin modifying activity offers the potential for targeted small molecule design. Also, comparative chromatin immunoprecipitation and the comprehensive identification of regulatory elements by FAIRE offer strategies to narrow the search space for regions of the genome that might play a role in tumor development. One such example is the repeat near IGF1 that is bound by EWS-FLI in cancer cells and differed in length from the reference genome. This finding suggests that length polymorphisms may influence EWS-FLI targeting and gene regulation as has been found for the GGAA repeat length polymorphism observed near *NR0B1* for which expression correlated with the number of repeats [115]. The identification of an extended tandem repeat proximal to *IGF1* may be of significance for disease development and treatment, since EWS-FLI-mediated IGF1 expression and signaling has been implicated in Ewing Sarcoma development [148] [149], and inhibition of IGF1 signaling is being studied as a potential therapeutic strategy. Such polymorphisms could arise *de novo* during tumor development or represent an allelic selection in individuals, and the observed selection for longer repeats could represent a mechanism to augment target gene expression. Further work will be necessary to determine the functional significance of polymorphisms or other mutations on disease susceptibility, onset, progression, and treatment.

2.4 Methods

2.4.1 Cell culture

EWS502 were cultured in RPMI 1640 supplemented with 10% FBS, HUVEC cells were cultured in Vasculife Basal Media (Lifeline Technologies) supplemented with 10% FBS. Both cell lines were maintained at standard growth conditions of 37°C and 5% CO_2.

2.4.2 Lentiviral knockdown-expression

A short hairpin region complementary to the 3′ untranslated region of FLI1 (5′-TGCCCATCCTGCACACTTACTTCAAGAGAGTAAGTGTGCAGGATGGGCTTTTTC-3′ sense strand) together with PCR-generated HA-tagged EWS-FLI, HA-tagged EWR1, and HA-tagged FLI1 were cloned into pLL5.5 [150]. Lentivirus was produced in HEK293T cells as described [150]. EWS502 or HUVEC cells were infected with lentivirus in the presence of polybrene (6 µg/mL) for 3 h after which media was changed. Chromatin or RNA was isolated at 72 h (see below).

2.4.3 Chromatin Immunoprecipitation (ChIP) and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)

Chromatin was isolated and immunoprecipitation was performed as described in [151] using 2 µg of anti-HA antibody (Abcam ab9110), anti-H3K4me1 (Abcam ab8895), anti-H3K4me2 (Abcam ab32356), anti-H3K4me3 (Abcam ab12209), anti-H3K27me3 (Millipore 07-449), anti-RNA Pol II (Abcam ab103968, or H3 (kindly provided by the Strahl lab) . Immunoprecipitated DNA was prepared for high-throughput sequencing per manufacturer's recommendations (Illumina) including DNA purification using AMPure XP beads (Agencourt) before PCR amplification. Quantitative PCR was performed as described (Absolute SYBR Green ROX Mix, Thermo Scientific). PCR primers are available upon request. FAIRE was performed on three independent cell harvests as previously described [73], and isolated DNA was prepared for sequencing as above.

2.4.4 Quality Control and Reference Genome Alignment

Reads from chromatin immunoprecipitations were aligned to the reference human genome (hg18) with Bowtie [152] using default parameters, and unambiguously placed reads were retained. Biological replicates were then merged, cross-replicate correlation was assessed, and reads were extended *in silico* to a final length of 200 bp. Any extended reads that overlapped large-scale repetitive elements were then removed. Reads from FAIRE were allowed to potentially map to up to four genomic locations, but the best scoring alignment was chosen. Biological replicates were then merged, cross-replicate correlation was assessed, and reads were extended *in silico* to a final length of 134 bp. Any extended reads that overlapped large-scale repetitive elements were then removed.

2.4.5 Peak calling and permutation

Areas of significant enrichment were identified using the Zero Inflated Negative Binomial Algorithm (ZINBA, [110]). A window size and offset of 250 bp and 50 bp, respectively, were utilized for EWSR1 and FLI1 and 500 bp and 125 bp for EWS-FLI. In all cases, a mixture regression model was created using a combination of the input control, GC content, and a background derived copy number variation model. Windows with q-values exceeding 0.95–0.99 were considered statistically significant, and peak boundaries were further refined when necessary. Additional parameters were specified to account for the broad nature of H3K27me3 domains. The percentage of peaks and average signal over a meta-gene were calculated using CEAS [153] and plotted in R. For analyses of GGAA repeat length, peak coordinates were permuted 1000 times across the uniquely-mappable genome while maintaining chromosomal distribution using BEDTools [154]. The frequency of tandem GGAA/TTCC repeats was computed for lengths 1–25 and compared to that of the test peak coordinates to compute a two-sided p-value.

2.4.6 Hierarchical Clustering and Motif Identification

A union set of all EWS-FLI and FLI1 peaks for each cell type were merged using Galaxy [155]. For each of the 51,085 regions, we retrieved the average number of sequencing reads from each experiment and filtered for regions where the standard deviation and interquartile range exceeded 0.75 and 0.5, respectively. The resulting 6,525 regions were median-centered and hierarchically clustered using average linkage and Pearson correlation. The resulting dendrogram and heatmap were created in Java Treeview [156]. Regions identified by clustering analysis were narrowed to a refined window immediately around the location of binding by intersecting the union set of all 200 bp windows around the site of greatest signal (summit). *De novo* motif detection was performed using CisFinder [119] using the 200 bp flanking sequence as background. Motif heatmaps were created by calculating the input-normalized number of sequencing reads for each sample in the 2 kb region surrounding each identified motif location.

2.4.7 Flow Cytometry

48 h after lentiviral infection, Ewing sarcoma cells (A673) were trypsinized and washed once with PBS then permeabilized and fixed in 70% ethanol. Cells were washed with PBS and resuspended in PBS with propidium iodide (concentration) and RNase. Cells were analyzed (CyAn) and the cell cycle profile was quantified (ModFit LT, Verity House).

CHAPTER 3

VARIATION IN CHROMATIN ACCESSIBILITY IN HUMAN KIDNEY CANCER LINKS H3K36 METHYLTRANSFERASE LOSS WITH WIDESPREAD RNA PROCESSING DEFECTS

3.1 Introduction

Large-scale cancer sequencing studies continue to identify mutations in genes encoding chromatin regulatory proteins in a wide variety of human cancers. The downstream molecular consequences of these mutations, however, remain unknown. Clear cell Renal Cell Carcinoma (ccRCC) is a particularly relevant model for the study of chromatin regulation in cancer for several reasons. First, relative to mutations in other classes of genes, ccRCC are marked by frequent mutation of chromatin regulators [42] [32] [157] [158]. Three of the more commonly mutated genes in ccRCC include chromatin modifiers *SETD2*, *PBRM1*, and *BAP1* [42] [32] [157] [159], suggesting that alterations at the level of chromatin may play a prominent role in the development of ccRCC [42] [32]. Mutation-associated changes in chromatin organization may promote oncogenesis in novel ways, and it has been suggested that specific chromatin regulator mutations may confer differences in patient survival or associate with more advanced disease [56]. However, the downstream effect of these mutations on tumor chromatin biology remains unknown. Second, this cancer is tightly associated with a distinct transcriptional program resulting from the inactivation of the von Hippel-Lindau (*VHL*) tumor suppressor gene [160] [161] [162] [163]. The loss of *VHL* results in the stabilization of hypoxia inducible factors (HIFs), transcription factors that activate a complex program of downstream targets, including vascular endothelial growth factor (VEGF) and other genes [164] [165] [163]. Third, besides *VHL* and chromatin regulators, mutations in other cancer-associated pathways are generally absent from ccRCC tumors.

Elucidating the functional consequences of mutations in genes encoding chromatin regulatory proteins on chromatin organization and transcription in human tumor specimens requires the application of techniques developed for cultured cells to primary human tissues. Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) interrogates chromatin accessibility by isolating nucleosome-depleted regions of DNA [166] [72] [73] [10] [74]. These regions harbor regulatory elements such as active transcriptional start sites, transcriptional enhancers, insulators, silencers, and locus control regions [72] [73] [10] [15] [54] [74]. As a component of the ENCODE project, FAIRE has been used to identify regulatory elements across a wide range of cell lines [54] [167]. However, the application of FAIRE to primary human tissue or to explore the association between chromatin and genetic alterations in cancer has yet to be evaluated.

We modified FAIRE for use on primary human clinical samples to define the chromatin landscape in a large cohort of ccRCC tumors and matched normal tissues. We identified tumor- and normal-kidney-specific classes of chromatin accessibility changes, as well as those associated with chromatin modifier mutations. We focused our study on SETD2, which trimethylates lysine-36 on histone H3 (H3K36me3) [168] [169] [170] [171] [172] [31]. Associated with the RNA polymerase II complex, SETD2-dependent methylation tends to occur towards the 3' ends of genes and over nucleosomes located at exons [171] [173] [174]. SETD2 and H3K36me3 seem to play a role in co-transcriptional RNA processing. In cell-culture-based studies, silencing of SETD2 or readers of H3K36me3 has been associated with differential exon inclusion for individual genes [175] [176] and alternative transcription start site utilization [177]. However, the consequence of SETD2 deficiency on chromatin organization and RNA processing remains to be explored on a genome-wide scale and in a disease-relevant model. *SETD2* is mutated in approximately 12% of primary human ccRCC tumors and results in H3K36me3 deficiency [178]. A similar rate of *SETD2* mutation has also been observed in high-grade gliomas [179]. A recent study of intratumor heterogeneity in ccRCC identified distinct *SETD2* mutations in all subsections of the same tumor suggesting the importance of disrupting SETD2 function for a subset of tumors [178].

We found that *SETD2* mutation was associated with chromatin accessibility differences preferentially in gene bodies, and these genes frequently exhibited RNA processing defects. Nearly 25% of all expressed genes demonstrated aberrancies in splicing, including exon skipping, intron retention, and alternative transcription start and termination sites. We observed that misspliced exons were marked by a striking increase in chromatin accessibility immediately upstream of the aberrant splice and a loss of nucleosome occupancy directly over the exon. This study represents the first investigation of chromatin organization in human tumors to identify the impact of chromatin modifier mutations on the genomic landscape. Understanding chromatin dysregulation in cancer may ultimately inform the application of emerging classes of chromatin-targeted small molecules in renal cancer.

3.2 Results

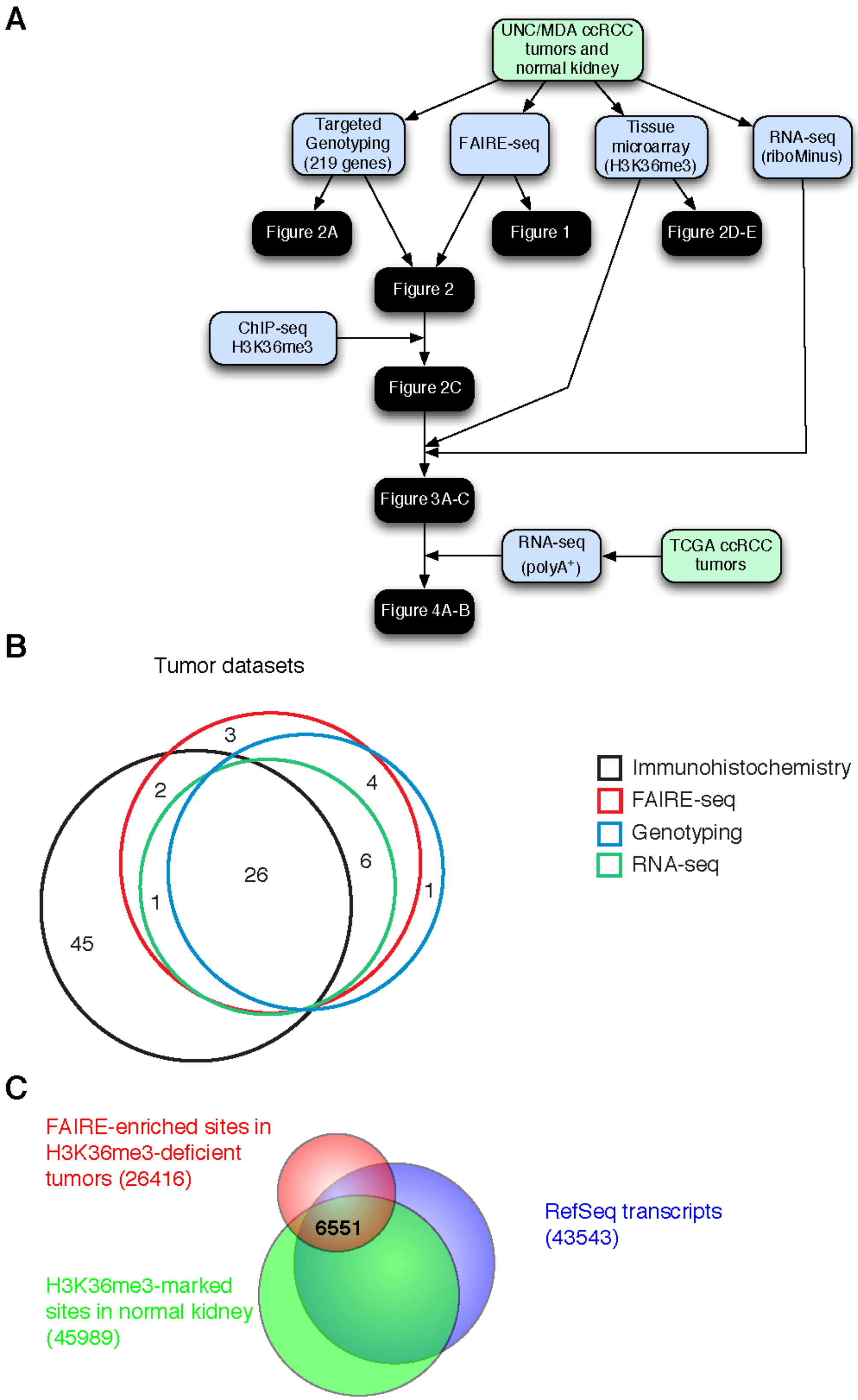
3.2.1 Differences in chromatin accessibility between tumors and normal kidney tissue corroborate the underlying role of HIF in ccRCC

We performed FAIRE-seq on 42 primary ccRCC tumor samples as well as uninvolved matched normal kidney from seven of these patients (Figure 3.1A–B). We identified approximately 11,000 500-bp genomic intervals with differences in chromatin accessibility that discriminated tumors from normal kidney (2-sided t-test, $p < 0.01$) (Figure 3.2A–B). For approximately 70% of these regions, FAIRE signal was increased in the tumor samples, indicative of nucleosome depletion. Using hierarchical clustering, three clusters of genomic loci emerged: two were marked by tumor-specific nucleosome depletion (Clusters 1 and 2), and another was characterized by nucleosome depletion in normal kidney tissue but not in tumors (Cluster 3). Virtually all tumors exhibited nucleosome depletion at the sites in Cluster 1, whereas approximately 50% of tumors demonstrated FAIRE enrichment at regions in Cluster 2.

We then examined each cluster for shared biological associations among the loci and adjacent genes. Regions in each cluster were associated with genes (GREAT, [118]). For sites in Cluster 1, 2,274 genes were identified, many of which members of several cancer-associated gene sets. Particularly interesting in the setting of ccRCC, where HIF transcription factor family stabilization and activation of hypoxia response genes is a central feature of this tumor type, we found the most significantly associated genes in this cluster were involved in HIF activation and hypoxia regulation (Figure 3.2C, full list of associations for each cluster in Figure 3.3). This association was not observed for regions in Cluster 2 or 3 (Figure 3.3). Analysis of the sequences in Cluster 1 identified several highly enriched Transcription Factor (TF) motifs [181], including the hypoxia response element consensus binding sequence (Figure 3.2D). We additionally found that previously identified HIF1A and HIF2A (EPAS1) binding sites [180] only significantly overlapped loci in Cluster 1 ($p < 0.001$, Figure 3.2B, E). The detection of features associated with the hypoxia response through variation in chromatin accessibility is consistent with the unique link between HIF activity and ccRCC, and these results demonstrate the ability of FAIRE to detect central biological pathways through the identification variations in chromatin organization in an unbiased fashion.

3.2.2 *SETD2* mutations link H3K36me3 loss with changes in chromatin accessibility

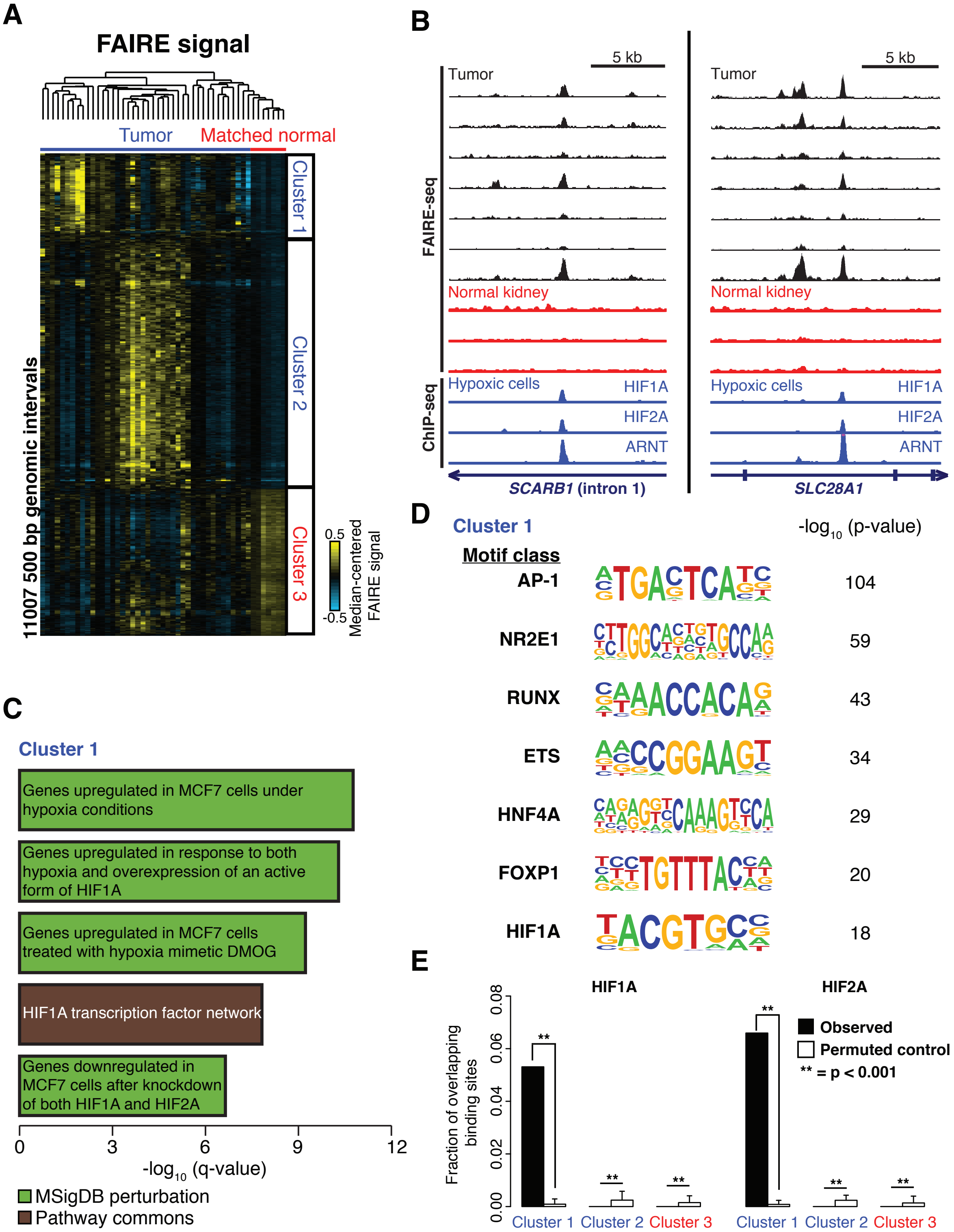
To identify mutations in chromatin modifiers within tumor samples, we genotyped 33 unique ccRCC tumors (from our cohort of 42 above) and the same 7 matched normal kidney specimens (Figure 3.1A–B). We classified sequence variants based on predicted ability to confer severe protein structural changes, including frameshift, nonsense, and mutations altering an annotated splice site (“high severity”), as well as missense mutations (“moderate severity”)


A
UNC/MDA ccRCC tumors and normal kidney
Targeted Genotyping (219 genes)
FAIRE-seq
Tissue microarray (H3K36me3)
RNA-seq (riboMinus)
Figure 2A
Figure 1
Figure 2D-E
Figure 2
ChIP-seq H3K36me3
Figure 2C
Figure 3A-C
RNA-seq (polyA+)
TCGA ccRCC tumors
Figure 4A-B
B
Tumor datasets
3
2
4
26
6
1
1
45
□ Immunohistochemistry
□ FAIRE-seq
□ Genotyping
□ RNA-seq
C
FAIRE-enriched sites in H3K36me3-deficient tumors (26416)
6551
RefSeq transcripts (43543)
H3K36me3-marked sites in normal kidney (45989)

(Figure 3.4A). Approximately half of the *SETD2* mutations in these classes were predicted to disrupt the catalytic SET domain. High- and moderate-severity mutations were also observed in other domains in *SETD2* including the SRI domain, which mediates the interaction with RNA Polymerase II. A prediction of copy number using the genotyping data also revealed that with the exception of one tumor (which displayed one high and one moderate severity mutation) loss of heterozygozity coincided with mutations in SETD2 (Figure 3.6C).

We identified approximately 7,000 500-bp windows exhibiting significant variation in FAIRE enrichment between *SETD2*-mutant and *SETD2*-normal tumors (2-sided t-test, $p < 0.01$) (Figure 3.4B, Figure 3.1C). In the *SETD2*-mutant tumors, FAIRE signal at these regions was most commonly increased (80%), suggesting that *SETD2* loss is preferentially associated with greater chromatin accessibility. SETD2 trimethylates H3K36 typically at gene bodies [3] [173]. Regions with increased FAIRE signal in *SETD2*-mutated tumors (one-sided t-test, $p < 0.01$) also overlapped gene bodies (49% of sites), most of which (91%) were marked by H3K36me3 in normal kidney ($p < 0.001$ relative to permuted control) (Figure 3.5). More specifically, regions of increased chromatin accessibility associated with *SETD2* mutation were enriched directly over the same domains marked by H3K36me3 (24.5%, $p < 0.001$ relative to permuted control) (Figure 3.4C). In contrast, the regions with decreased FAIRE signal showed no association with H3K36me3, and in fact showed a significant underrepresentation relative to permuted control ($p < 0.001$). As an additional control, we tested for this association at regions with increased FAIRE signal in *PBRM1*-mutant tumors, which we

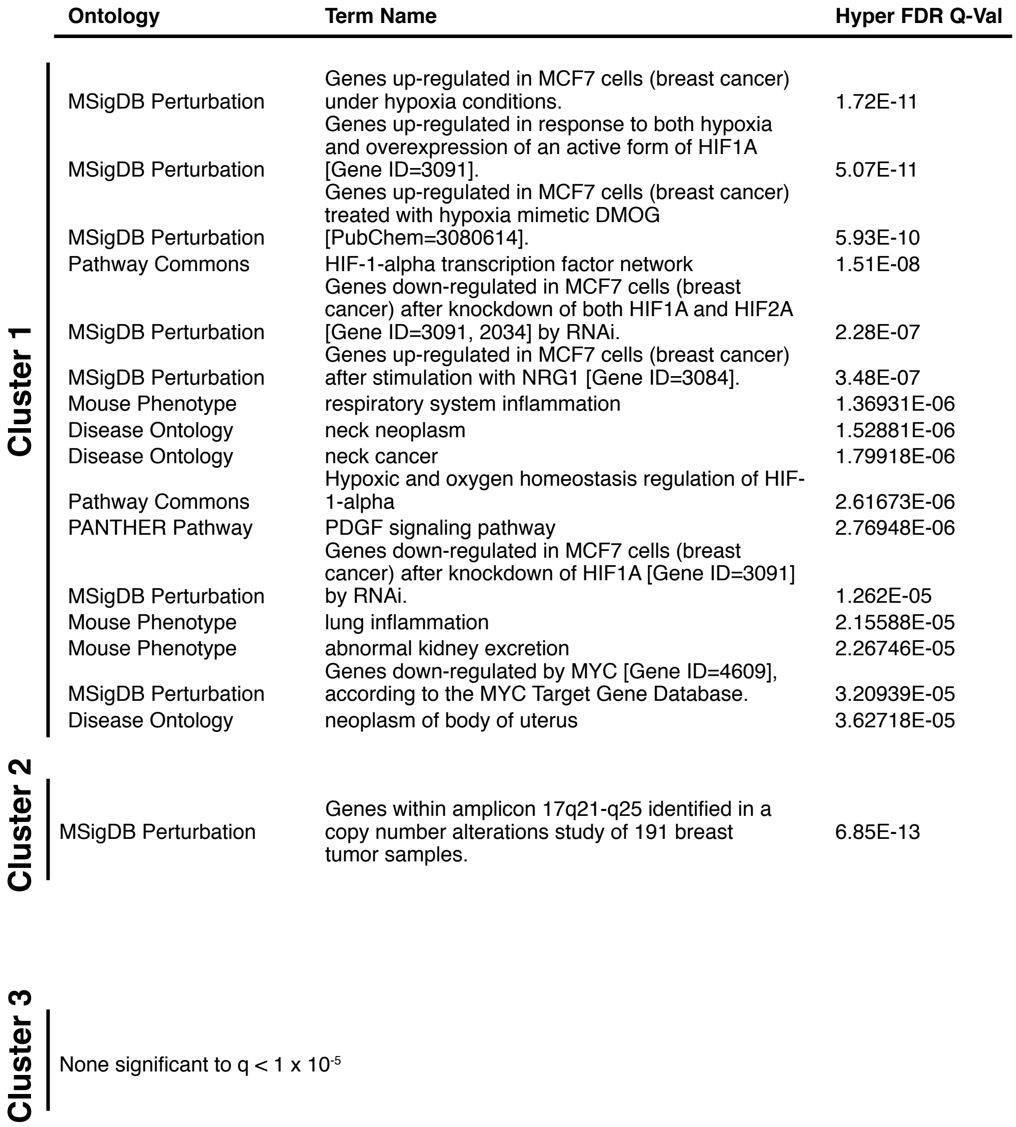
Figure 3.1: Schematic representation of dataset integration and genomic site identification. A. Flowchart depicting dataset integration utilized for each figure. Datasets are colored in green, data types in blue, and resulting figures in black. B. Venn diagram depicting how tumors were utilized for various experimental approaches. C. Venn diagram depicting the intersection of the RefSeq transcripts, H3K36me3-marked regions and genes with FAIRE enrichment in H3K36me3-deficient tumors relative to H3K36me3-normal tumors to yield the 6551 genomic sites used for determination of intron retention.


A
FAIRE signal
Tumor
Matched normal
Cluster 1
Cluster 2
Cluster 3
11007 500 bp genomic intervals
0.5
-0.5
Median-centered FAIRE signal
B
FAIRE-seq
ChIP-seq
5 kb
Tumor
Normal kidney
Hypoxic cells
HIF1A
HIF2A
ARNT
SCARB1 (intron 1)
SLC28A1
C
Cluster 1
Genes upregulated in MCF7 cells under hypoxia conditions
Genes upregulated in response to both hypoxia and overexpression of an active form of HIF1A
Genes upregulated in MCF7 cells treated with hypoxia mimetic DMOG
HIF1A transcription factor network
Genes downregulated in MCF7 cells after knockdown of both HIF1A and HIF2A
0
3
6
9
12
-log$_{10}$ (q-value)
🟩 MSigDB perturbation
🟫 Pathway commons
D
Cluster 1
Motif class
-log$_{10}$ (p-value)
AP-1
104
NR2E1
59
RUNX
43
ETS
34
HNF4A
29
FOXP1
20
HIF1A
18
E
HIF1A
HIF2A
Fraction of overlapping binding sites
0.00
0.02
0.04
0.06
0.08
**
Cluster 1
Cluster 2
Cluster 3
■ Observed
□ Permuted control
** = p < 0.001

expected to yield a divergent set of loci. Indeed, areas of increased chromatin accessibility associated with this mutation were significantly underrepresented at H3K36me3-marked regions ($p < 0.001$ relative to permuted control). Together, these data indicate that regions of nucleosome depletion associated with *SETD2* mutation preferentially occurs at genic sites normally marked by H3K36me3.

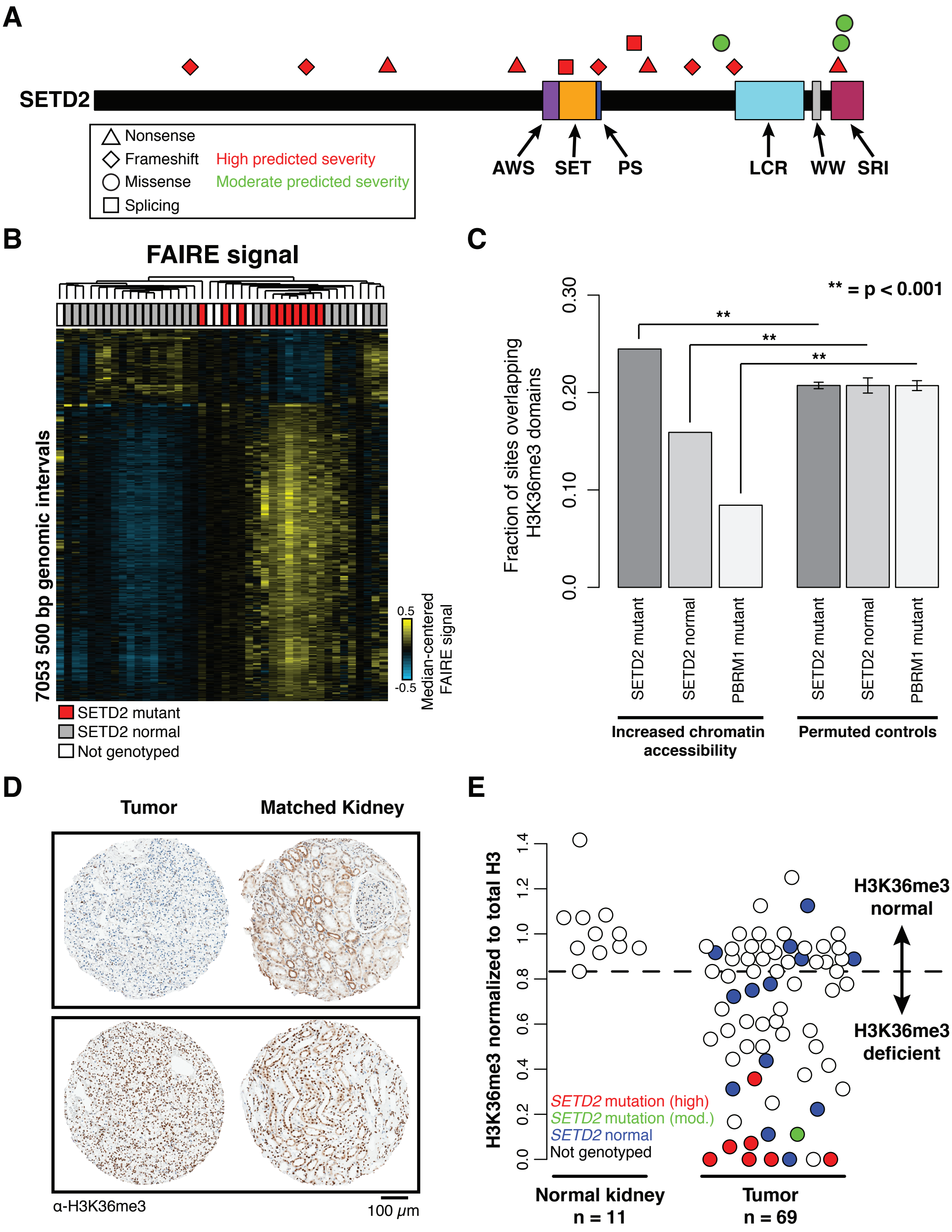
Although SETD2 is responsible for trimethylation of H3K36, other mechanisms may influence H3K36 methylation status. Moreover, the effects of specific classes of *SETD2* mutations in human tumors on H3K36 methylation in RCC are not known. We quantified H3K36me3 on a tissue microarray (69 tumors, 11 matched normal kidneys) (Figure 3.4D, Figure 3.1B). Whereas normal kidney samples demonstrated consistent nuclear H3K36me3 signal (Figure 3.4E), tumors displayed a range of staining intensity, with 53% of tumors exhibiting reduced H3K36me3 intensity. Hereafter, this group of tumors is referred to as "H3K36me3 deficient." Each of the eight tumors that contained mutations predicted to affect SETD2 activity and screened by IHC demonstrated H3K36me3 deficiency (Figure 3.4E). Tumors containing mutations before the SET domain (Q320fs, E978*, and Q1409*) displayed a complete loss of H3K36me3 signal. However, tumors with *SETD2* mutations located within the SET domain (G1681fs) or in the SRI domain (R2510L) displayed reduced H3K36me3 signal, suggesting that some mutations may cause a partial loss of function. Several tumors (8 of 13) without identified *SETD2* mutations also exhibited reduced H3K36me3 signal. SETD2

Figure 3.2: Regions of tumor-specific nucleosome eviction identify the underlying role of HIF in ccRCC. A. Hierarchical clustering of median-centered FAIRE signal in windows with significant differences between tumors and normal kidney (2-sided t-test $p < 0.01$). B. FAIRE-seq tracks for ccRCC (black) and uninvolved kidney (red) at two loci. ChIP-seq signal [180] from HIF1A, HIF2A, and ARNT are plotted in blue. C. The top five gene ontology associations ($q < 1 \times 10^{-5}$) with sites in Cluster 1 are shown. D. Transcription factor binding motifs enriched in Cluster 1 compared to local background 500 bp flanking windows ($>$ 2.5-fold over background and present in at least 10% of the Cluster 1 windows). P-values relative to local background are shown. E. Fraction of HIF1A and HIF2A binding sites [180] that overlap the loci in clusters 1, 2 and 3 compared to controls where clustering windows were permuted. Errors bars represent standard deviation (SD).



	Ontology	Term Name	Hyper FDR Q-Val
Cluster 1	MSigDB Perturbation	Genes up-regulated in MCF7 cells (breast cancer) under hypoxia conditions.	1.72E-11
	MSigDB Perturbation	Genes up-regulated in response to both hypoxia and overexpression of an active form of HIF1A [Gene ID=3091].	5.07E-11
	MSigDB Perturbation	Genes up-regulated in MCF7 cells (breast cancer) treated with hypoxia mimetic DMOG [PubChem=3080614].	5.93E-10
	Pathway Commons	HIF-1-alpha transcription factor network	1.51E-08
	MSigDB Perturbation	Genes down-regulated in MCF7 cells (breast cancer) after knockdown of both HIF1A and HIF2A [Gene ID=3091, 2034] by RNAi.	2.28E-07
	MSigDB Perturbation	Genes up-regulated in MCF7 cells (breast cancer) after stimulation with NRG1 [Gene ID=3084].	3.48E-07
	Mouse Phenotype	respiratory system inflammation	1.36931E-06
	Disease Ontology	neck neoplasm	1.52881E-06
	Disease Ontology	neck cancer	1.79918E-06
	Pathway Commons	Hypoxic and oxygen homeostasis regulation of HIF-1-alpha	2.61673E-06
	PANTHER Pathway	PDGF signaling pathway	2.76948E-06
	MSigDB Perturbation	Genes down-regulated in MCF7 cells (breast cancer) after knockdown of HIF1A [Gene ID=3091] by RNAi.	1.262E-05
	Mouse Phenotype	lung inflammation	2.15588E-05
	Mouse Phenotype	abnormal kidney excretion	2.26746E-05
	MSigDB Perturbation	Genes down-regulated by MYC [Gene ID=4609], according to the MYC Target Gene Database.	3.20939E-05
	Disease Ontology	neoplasm of body of uterus	3.62718E-05
Cluster 2	MSigDB Perturbation	Genes within amplicon 17q21-q25 identified in a copy number alterations study of 191 breast tumor samples.	6.85E-13
Cluster 3	None significant to $q < 1 \times 10^{-5}$		

Figure 3.3: Gene ontology associations with sites in Clusters 1–3. The Genomic Regions Enrichment of Annotations Tool (GREAT[118]) was used to analyze the functional significance of regions identified by FAIRE. Associations with hypergeometric FDR-adjusted q-values less than 1 x 10^{-5} are shown. Cluster 3 analysis yielded no ontologies that met this threshold.


A
SETD2
AWS
SET
PS
LCR
WW
SRI
△ Nonsense
◇ Frameshift
○ Missense
□ Splicing
High predicted severity
Moderate predicted severity
B
FAIRE signal
7053 500 bp genomic intervals
0.5
-0.5
Median-centered FAIRE signal
SETD2 mutant
SETD2 normal
Not genotyped
C
** = p < 0.001
Fraction of sites overlapping H3K36me3 domains
0.0
0.10
0.20
0.30
SETD2 mutant
SETD2 normal
PBRM1 mutant
SETD2 mutant
SETD2 normal
PBRM1 mutant
Increased chromatin accessibility
Permuted controls
D
Tumor
Matched Kidney
α-H3K36me3
100 μm
E
H3K36me3 normalized to total H3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
H3K36me3 normal
H3K36me3 deficient
SETD2 mutation (high)
SETD2 mutation (mod.)
SETD2 normal
Not genotyped
Normal kidney n = 11
Tumor n = 69

was undetectable by immunohistochemistry in two of these tumors, whereas others exhibited decreased *SETD2* mRNA, suggesting alternate mechanisms for H3K36me3 loss (Figure 3.6). We also observed evidence for *SETD2* gene hemizygosity in other H3K36me3-deficient *SETD2*-normal tumors, suggesting that loss of heterozygosity may contribute to deficiency in H3K36 methylating activity (Figure 3.6C). Interestingly, one tumor (Tumor 25 in Figure 3.6C) did not exhibit a copy number loss, carried two *SETD2* mutations (E1846*, high severity; I2499S, moderate severity), and showed a moderate H3K36me3 deficiency (an intensity value of 0.36 in Figure 3.4E). We would thus predict that at least one of these mutations is hypomorphic, thus explaining the intermediate magnitude of the H3K36me3 deficiency. Similarly, we detected two mutations in *SETD2* in another tumor (Tumor 3 in Figure 3.6C), which exhibited a global loss in H3K36me3 staining along with copy number loss. These data suggest that either the tumor cell population was heterogeneous and the remaining allele was differentially mutated in each population (as was observed in [178]) or that the one remaining allele was mutated in two locations. Together, these data illustrate that defective H3K36me3 is a common feature of ccRCC and that SETD2 genotype alone underestimates H3K36me3 deficiency.

Figure 3.4: SETD2 mutations link H3K36me3 loss with changes in chromatin accessibility. A. Schematic representation of *SETD2* mutations predicted to have high or moderate severity on protein structure. B. Hierarchical clustering of median-centered FAIRE signal in windows with significant differences between *SETD2*-mutant tumors (red) and tumors without *SETD2* mutation (gray) (2-sided t-test $p < 0.01$). Samples not genotyped are labeled in white. C. Proportions of nucleosome-depleted loci overlapping H3K36me3-marked regions compared to loci with permuted genomic coordinates. Errors bars represent SD. D. Representative immunostaining of two ccRCC tumor-normal pairs on the tissue microarray. E. Quantification of H3-normalized H3K36me3 intensity across 11 normal kidney and 69 renal tumors. Mutation severity (high, red; moderate, green; none, blue) is indicated. Samples with unknown *SETD2* mutation status are plotted in white. The threshold for H3K36me3 deficiency was set to the lowest observed intensity in normal tissue (dashed line).

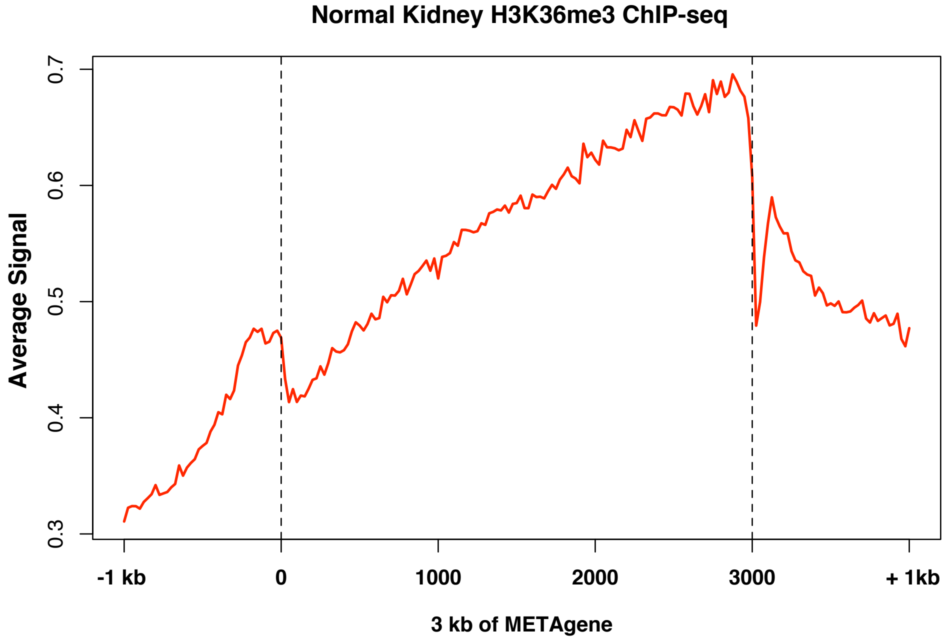
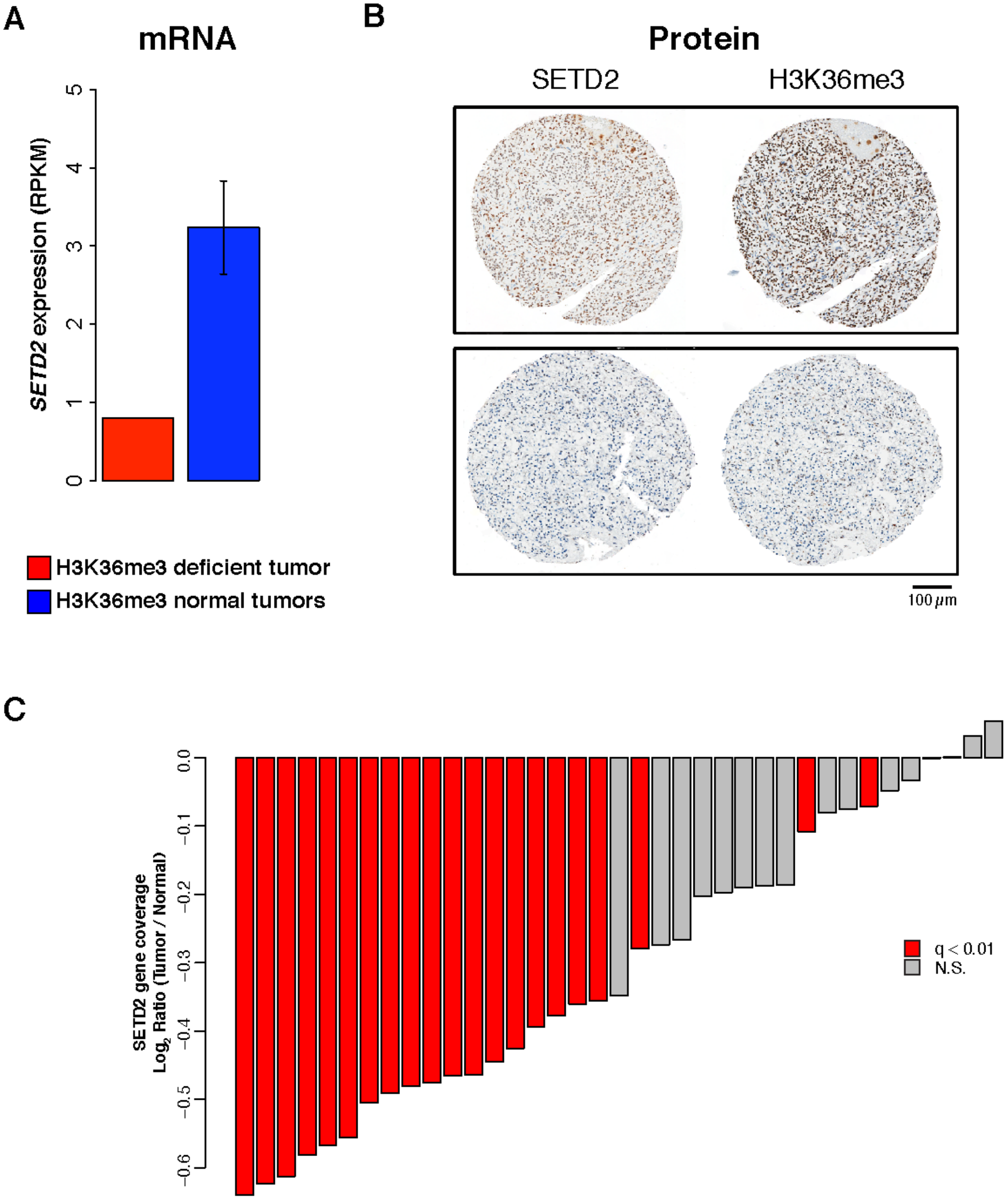


Figure 3.5: META-gene plot of H3K36me3 ChIP-seq signal from normal kidney. The average number of reads is plotted across the 3 kb of average gene length, plus 1 kb upstream and downstream, demonstrating a 3′ bias for accumulation of the H3K36me3 mark.



	Tumor 1	Tumor 2	Tumor 3	Tumor 4	Tumor 5	Tumor 6	Tumor 7	Tumor 8	Tumor 9	Tumor 10	Tumor 11	Tumor 12	Tumor 13	Tumor 14	Tumor 15	Tumor 16	Tumor 17	Tumor 18	Tumor 19	Tumor 20	Tumor 21	Tumor 22	Tumor 23	Tumor 24	Tumor 25	Tumor 26	Tumor 27	Tumor 28	Tumor 29	Tumor 30	Tumor 31	Tumor 32	Tumor 33	Tumor 34	Tumor 35	Tumor 36	Tumor 37
High severity		*	*						*	*	*		*	*				*		*					*												
Moderate severity			*			*																			*												
H3K36me3 deficient	?	*	*	*	*	*	?	?	*	?	*	*	*	?	*	*	*	*		*	*	*			*			*					*				

3.2.3 *SETD2* mutation is associated with DNA hypomethylation proximal to sites of nucleosome depletion

In many higher eukaryotes, the H3K36me3 mark is recognized by several chromatin readers, one of which is the PWWP domain of the DNA methyltransferase Dnmt3a, resulting in DNA methylation proximal to the marked histone [182]. Using ccRCC DNA methylation data from The Cancer Genome Atlas (TCGA), we observed localized changes ($p < 0.05$) in DNA methylation, primarily (>70% of probes) DNA hypomethylation, in *SETD2*-mutant tumors of the TCGA dataset at nucleosome-depleted regions identified in our *SETD2*-mutant tumors (Figure 3.7). These data link changes DNA methylation to sites of nucleosome eviction and/or loss of H3K36me3 through *SETD2* mutation. This result underscores the importance of H3K36me3 and how its loss may confer a multifaceted alteration in the epigenome.

3.2.4 Intron retention and splicing defects affect a large fraction of genes with altered chromatin accessibility in *SETD2*-mutant tumors

H3K36me3 has been previously implicated in RNA processing [175] [183] [176], an association not previously examined in primary tissues or in a disease-relevant model. We thus hypothesized that H3K36me3 deficiency would alter RNA processing and splicing in tumors specifically at genes with altered chromatin accessibility. To assess total RNA, including pre-mRNA and non-polyadenylated transcripts, we performed RNA-seq on ribosome-depleted

Figure 3.6: Decreased *SETD2* expression in *SETD2*-normal tumors results in H3K36me3 deficiency. A. *SETD2* RNA expression (RPKM) for an H3K36me3-deficient tumor without *SETD2* mutation compared to the average RPKM for H3K36me3-normal tumors. Error bars represent standard error. B. Representative immunohistochemical staining of SETD2 protein and H3K36me3 in a genotypically *SETD2*-normal tumor with H3K36me3 deficiency (bottom panel) compared to a *SETD2*-normal tumor with normal SETD2 protein and H3K36me3 levels (top panel). C. Log-ratio of gene coverage of each tumor over the average of two normal kidney samples following a log-transformation and mean-centering of the number of reads mapping to each gene. Significance of the tumor-normal difference was determined using a negative binomial test. H3K36me3 status was unknown for 5 tumors (denoted by “?”).

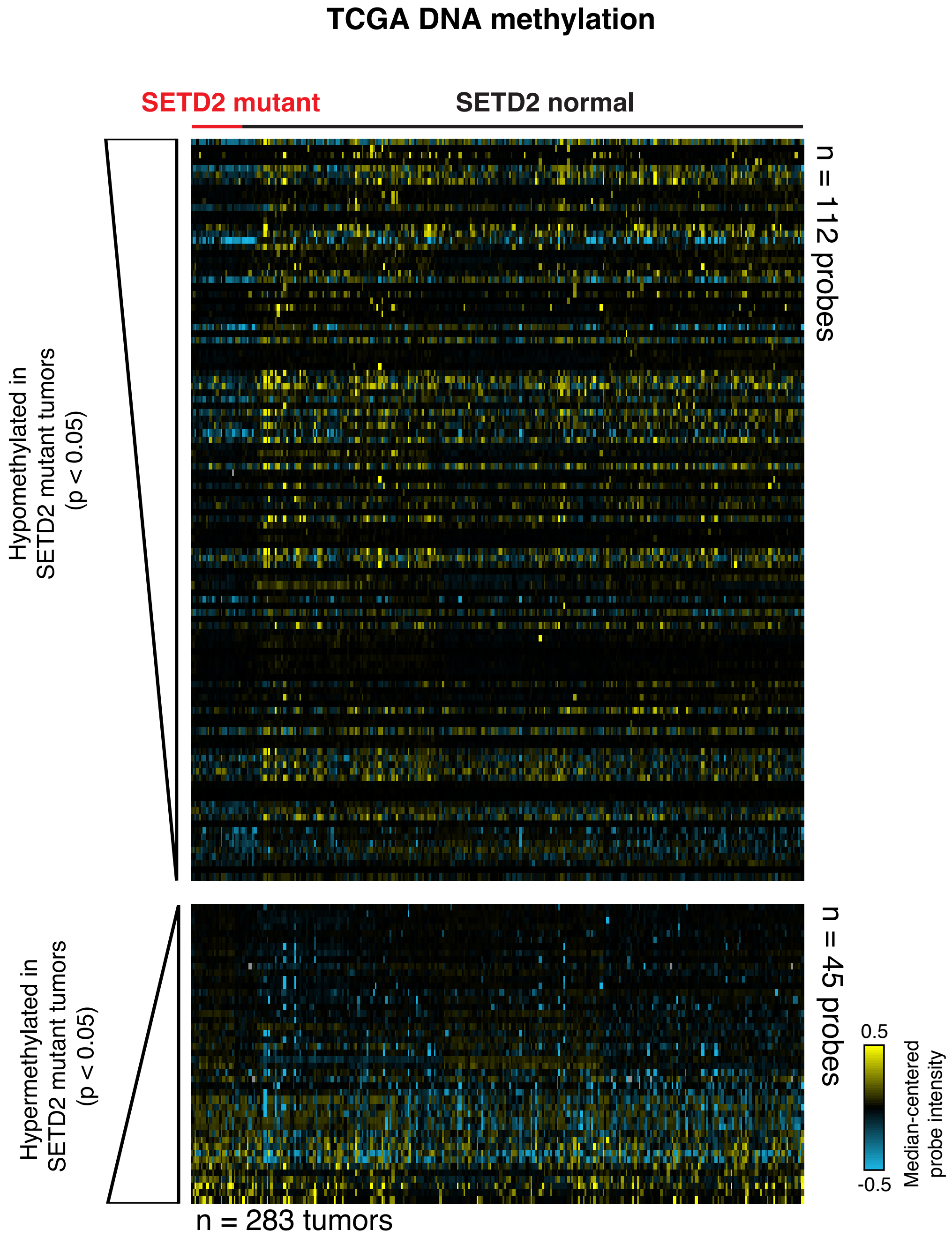


Figure 3.7: Nucleosome-depleted regions in *SETD2*-mutant tumors display localized DNA hypomethylation. Median-centered DNA methylation intensity for probes in genes both displaying FAIRE enrichment associated with *SETD2* mutation and marked by H3K36me3 in normal kidney. Data from 283 TCGA ccRCC tumors at 157 probes are presented. Specific hypomethylation of a cohort of these regions was selectively associated with *SETD2* mutation. Color was assigned on a scale of -0.5 to 0.5.

RNA from 33 tumors, all but one of which was annotated with mutational status (Figure 3.4 and Figure 3.1); six tumors without H3K36me3 status assessed by immunohistochemistry were omitted. We observed that H3K36me3-deficient tumors displayed a striking enrichment of intronic pre-mRNA signal compared to tumors with normal H3K36me3 levels. To quantify this effect, we calculated Intron Retention Scores (IRS), which reflect the ratio of intronic to exonic RNA-seq reads on a gene-by-gene basis for each tumor. IRS values range from 0 to 1, where a score of 0 represents a completely spliced message and a score of 1 represents uniform genic coverage. Intron retention was dramatically increased in the H3K36me3-deficient tumors at 95% of the transcripts (6,551 in total) marked by H3K36me3 in normal kidney and by nucleosome depletion (one-sided t-test, $p < 0.01$) in H3K36me3-deficient tumors (Figures 3.8A, 3.8C, 3.1B, 3.9A, 3.10). To confirm this result, we performed ChIP-seq from an independent normal kidney sample (Figure 3.9B). Of the 6,551 transcripts initially identified (Figure 3.1 and Figure 3.8A), 6,101 were identified using the second normal kidney sample, representing a 93% overlap. When the 6,551 transcripts by were instead stratified by *PBRM1* mutation status, widespread intron retention was not observed (Figure 3.8B), suggesting this effect is specific to tumors with H3K36me3 deficiency. Many of the affected genes are part of recognized cancer-associated pathways, including known tumor suppressors (e.g. *MET*, *PTEN*, and *TP53*), genes in the DNA repair pathway (e.g. *ATR*, *RAD50*, *POLN*, and *XRCC1*), cell cycle regulators (e.g. *CCNB1*, and *CCND3*), as well as numerous receptors and protein kinases (e.g. *BRAF*, *EGFR*, *PIK3CA*, and *TGFBR3*) (Figure 3.11).

3.2.5 Widespread RNA processing defects linked with *SETD2* mutations persist in the mature RNA pool and are marked by altered chromatin accessibility

To test whether the changes in pre-mRNA transcripts persisted into mature polyadenylated RNA, we analyzed TCGA RNA-seq data derived from polyA$^+$ mRNA from a large cohort (n=416) of ccRCC tumors. Applying a gene-model-independent algorithm for read mapping and transcript prediction [184], we observed that *SETD2*-mutant tumors exhibited significant

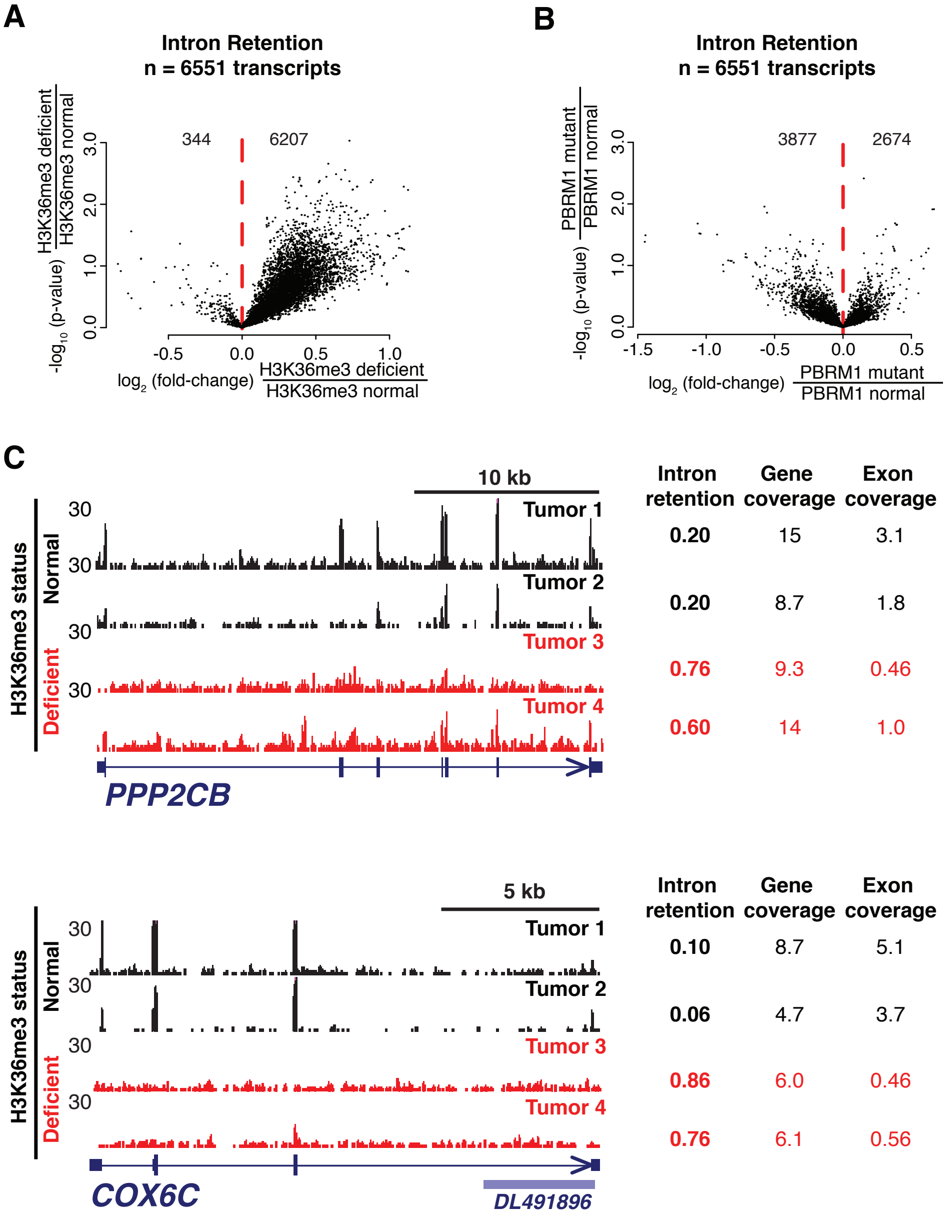


Figure 3.8: H3K36me3 deficiency is associated with intron retention. Intron Retention Scores for selected genes (Figure 3.1C) were compared between A. H3K36me3-deficient tumors and H3K36me3-normal tumors, and B. *PBRM1*-mutant and *PBRM1*-normal tumors. C. Example genes exhibiting increased intron retention in H3K36me3-deficient tumors (top, *PPP2CB*; bottom, *COX6C*). Intron retention scores, genic coverage (calculated with both intron and exon reads), and exonic coverage (calculated only with exonic reads) are provided for two H3K36me3-deficient tumors (red) and two H3K36me3-normal tumors (black).

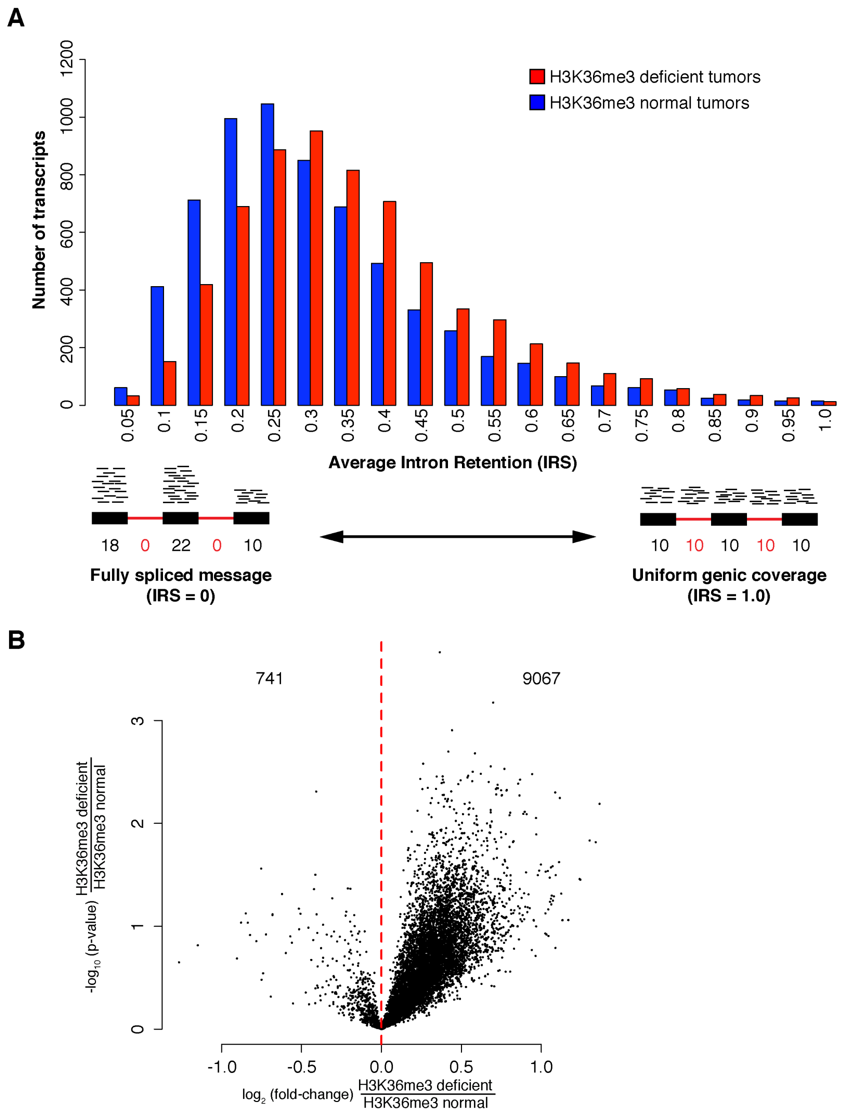


Figure 3.9: H3K36me3-deficient tumors display increased intron retention compared to H3K36me3-normal tumors. A. The number of RefSeq transcripts is plotted for each gradation of Intron Retention Score (IRS) in rRNA-depleted RNA. An average intronic retention score of 0 indicates exclusively exonic coverage (fully spliced) whereas an intronic retention score of 1.0 indicates uniform genic coverage (the absence of exonic enrichment). B. Intron Retention Scores for selected genes (Figure 3.1C) (however H3K36me3 ChIP-seq data was obtained from normal kidney of a second individual) were compared between H3K36me3-deficient tumors and H3K36me3-normal tumors.

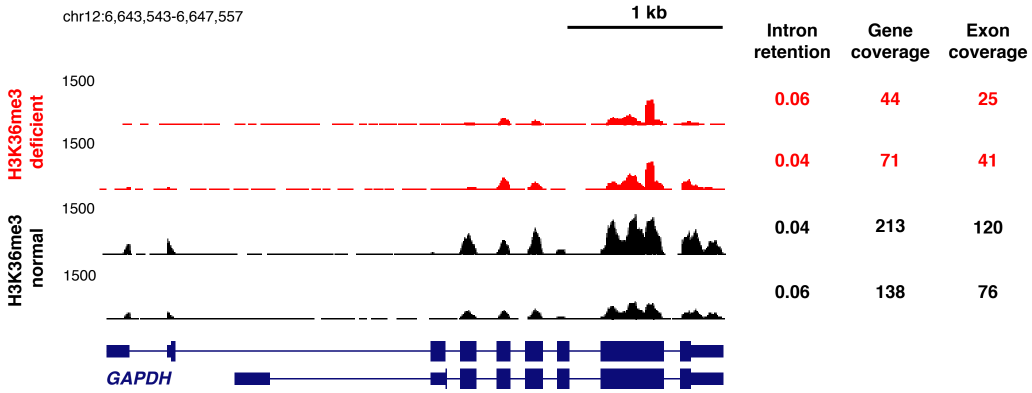


Figure 3.10: *GAPDH* exhibits low intron retention in H3K36me3-deficient tumors. For *GAPDH*, a gene not exhibiting increased intron retention in H3K36me3 deficient tumors, intron retention scores, genic coverage (calculated with both intron and exon reads), and exon coverage (calculated only with exonic reads) are provided for two H3K36me3 deficient tumors (red) and two H3K36me3 normal tumors (black).

alterations in transcript processing (3929 transcripts, Figure 3.12A–B). Alterations included intron retention (12% of altered transcripts, Figures 3.1B, 3.13), variation in exon utilization (66% of altered transcripts, Figures 3.12B–C, 3.14), and differences in transcriptional start and termination site usage (22% of altered transcripts). We also observed the generation of previously unannotated splice isoforms, which we validated by quantitative PCR in independent tumor samples (Figure 3.12D). Aberrancies in RNA processing were detected more frequently in highly expressed genes (Figure 3.15A). Low abundance messages may preclude the detection of differences in transcript processing. However, overall expression of genes

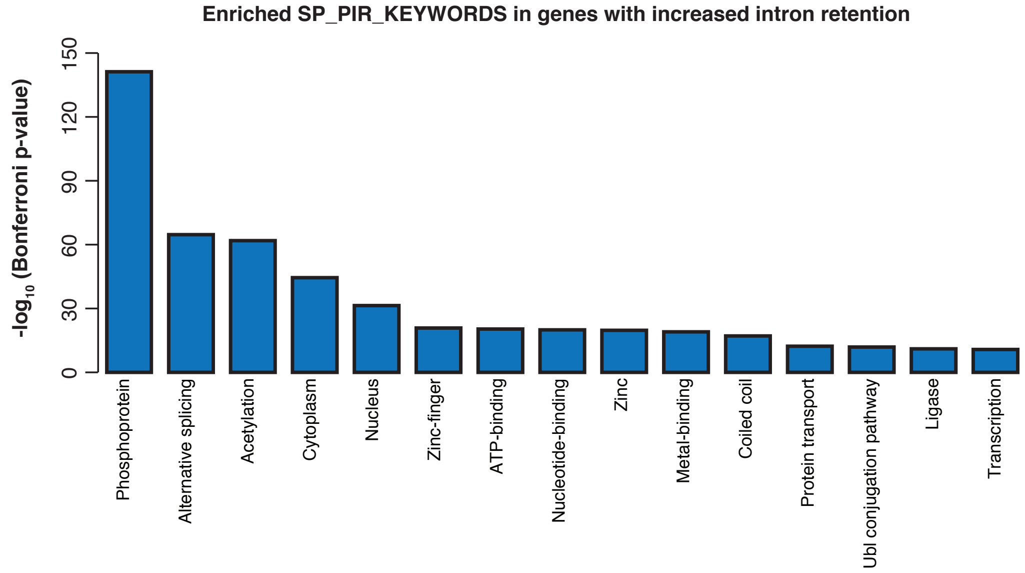


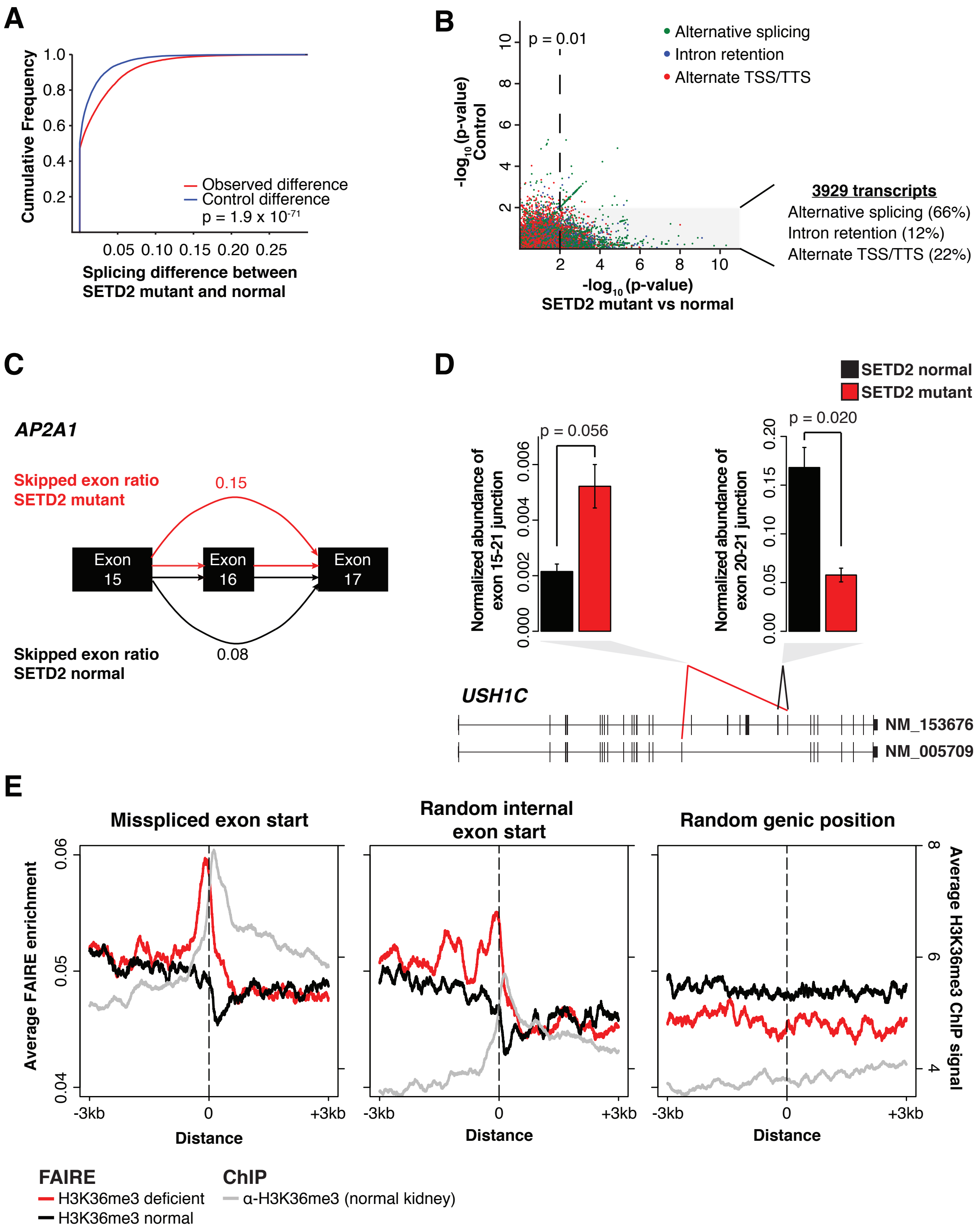
Figure 3.11: Enriched ontologies among genes with increased intron retention. Genes (n=2999) exhibiting increased intron retention between in H3K36me3-deficient tumors were assessed for associated ontologies. The most highly enriched terms among the SP_PIR_KEYWORDS ontology are presented as the -$\log_{10}$ of the Bonferroni-corrected p-value. P-values were filtered to $p < 1 \times 10^{-10}$. “Alternative splicing” refers to genes previously annotated as exhibiting alternative splicing.

exhibiting defects in RNA processing was comparable, however, between *SETD2*-mutant and *SETD2*-normal tumors (Figure 3.15B). Increased intron retention and altered splicing were additionally found more frequently in longer genes and genes with a larger number of exons (Figure 3.16A–B).

Since H3K36me3 preferentially marks well-positioned exonic nucleosomes [171] [173] [174], we analyzed chromatin accessibility around the intron-exon boundary of misspliced exons. Tumors with normal levels of H3K36me3 demonstrated an expected reduction of FAIRE signal immediately downstream of intron/exon junctions as well as a concomitant enrichment in H3K36me3 (from ChIP-seq in normal kidney), indicative of a well-positioned exonic nucleosome (Figure 3.12E, left), corroborating previous reports [173]. Strikingly, in H3K36me3-deficient tumors, evidence of the exonic nucleosome was lost and a dramatic increase in chromatin accessibility was observed immediately upstream (50 bp) of the intron/exon junction (Figure 3.12E, left, red line). This pattern was also evident, though less pronounced, at internal exon start sites of random genes (Figure 3.12E, middle) but completely absent at random genic positions (Figure 3.12E, right). Changes in chromatin accessibility even at internal exons chosen regardless of whether they exhibited a splicing defect may indicate a more widespread defect that may not always result in detectable variation in splicing. These data demonstrate the ability to detect subtle variations in chromatin organization in primary human tumors and link H3K36me3 loss with alterations in chromatin accessibility at exons.

3.3 Discussion

To identify the genomic consequences of mutated chromatin regulators, we modified and applied FAIRE-seq to a large cohort of primary kidney tumors. Using an unbiased approach, we identified variation in chromatin accessibility distinguishing tumors from normal kidney.


A
Cumulative Frequency
1.0
0.8
0.6
0.4
0.2
Observed difference
Control difference
p = 1.9 x 10^-71
0.05 0.10 0.15 0.20 0.25
Splicing difference between
SETD2 mutant and normal
B
-log10 (p-value)
Control
10
8
6
4
2
p = 0.01
Alternative splicing
Intron retention
Alternate TSS/TTS
2 4 6 8 10
-log10 (p-value)
SETD2 mutant vs normal
3929 transcripts
Alternative splicing (66%)
Intron retention (12%)
Alternate TSS/TTS (22%)
C
AP2A1
Skipped exon ratio
SETD2 mutant
0.15
Exon 15
Exon 16
Exon 17
Skipped exon ratio
SETD2 normal
0.08
D
SETD2 normal
SETD2 mutant
p = 0.056
Normalized abundance of
exon 15-21 junction
0.000 0.002 0.004 0.006
p = 0.020
Normalized abundance of
exon 20-21 junction
0.00 0.05 0.10 0.15 0.20
USH1C
NM_153676
NM_005709
E
Missplíced exon start
Random internal
exon start
Random genic position
Average FAIRE enrichment
0.04 0.05 0.06
Average H3K36me3 ChIP signal
4 6 8
-3kb 0 +3kb
Distance
FAIRE
H3K36me3 deficient
H3K36me3 normal
ChIP
α-H3K36me3 (normal kidney)

Tumor-specific open chromatin corresponded to HIF-targeted sites and was linked to genes involved in the hypoxia response. This result reflects the well-studied association of ccRCC development with *VHL* inactivation and HIF stabilization. These data also serve to validate the use of FAIRE in primary tumors to detect biologically meaningful pathways.

We then associated variation in chromatin accessibility with mutations in chromatin regulators. Focusing on *SETD2*, we observed widespread increases in chromatin accessibility especially in gene bodies typically harboring H3K36me3 in normal kidney tissue. A recent report suggested that *SETD2* silencing in cultured cells results in alternative internal transcriptional start sites [177], akin to cryptic initiation observed in yeast [185] [186]. Our data using human tumor specimens supports a more diverse model for transcriptional defects, including retention of introns, missplicing of exons, and usage of alternative transcriptional start or end sites. These defects were widespread, affecting nearly 25% of all expressed genes, and defects were more common in highly transcribed genes.

Moreover, we found a surprising increase in chromatin accessibility immediately upstream (50 bp) of misspliced exons in *SETD2*-mutated tumors. This result suggests a mechanism by

Figure 3.12: Widespread RNA processing defects linked with *SETD2* mutations persist in the mature RNA pool and are marked by altered chromatin accessibility. A. Splicing differences (see methods) between *SETD2*-mutant and *SETD2*-normal tumors (red) compared to a permuted control (blue) are plotted as a cumulative distribution function. B. Significance of the difference in ratios between *SETD2*-mutant and *SETD2*-normal tumors (x-axis) plotted against the scrambled control (y-axis). Points are colored by the class of RNA processing aberrancy. A gray box represents significance (p = 0.01) in the *SETD2* mutant-normal comparison, but not significant in the control comparison. The percentages of significant differences in transcript processing are also presented. C. Schematic of *AP2A1* splicing. Exon skipping was represented as the ratio of included exon coverage to the sum of the exon and the spliced form. The skipped exon ratio is provided for *SETD2*-mutant tumors (red) and *SETD2*-normal tumors (black). D. Quantitative PCR across two *USH1C* alternative exon utilization sites identified by RNA-seq for three *SETD2*-normal tumors (black) and two *SETD2*-mutant tumors (red). Error bars represent standard error. E. FAIRE signal plotted around the exon start (within 3kb) of misspliced exons (left), random internal exon starts (middle), and random genic positions (right) for H3K36me3-deficient tumors (red) and H3K36me3-normal tumors (black). H3K36me3 ChIP-seq signal from normal kidney tissue is plotted in gray.

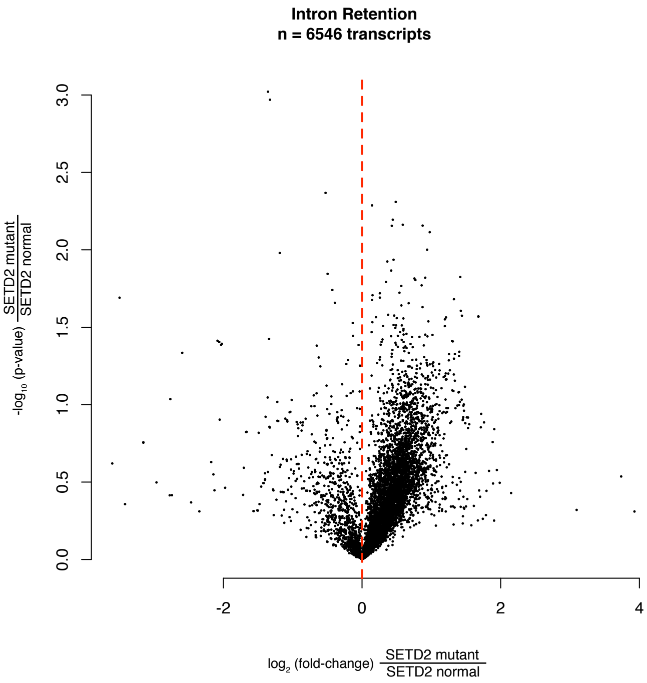


Figure 3.13: Intron retention in *SETD2*-mutant tumors persists into mature, polyadenylated RNA. Intron retention scores for 6,546 RefSeq transcripts in polyA$^+$ RNA from the TCGA dataset were averaged across two *SETD2*-mutant tumors and compared to that of three *SETD2*-normal tumors. These tumors were selected based on their inclusion and analysis in both datasets. These transcripts were marked by H3K36me3 in normal kidney and contained a site determined by FAIRE to be more nucleosome-depleted in H3K36me3-deficient tumors.

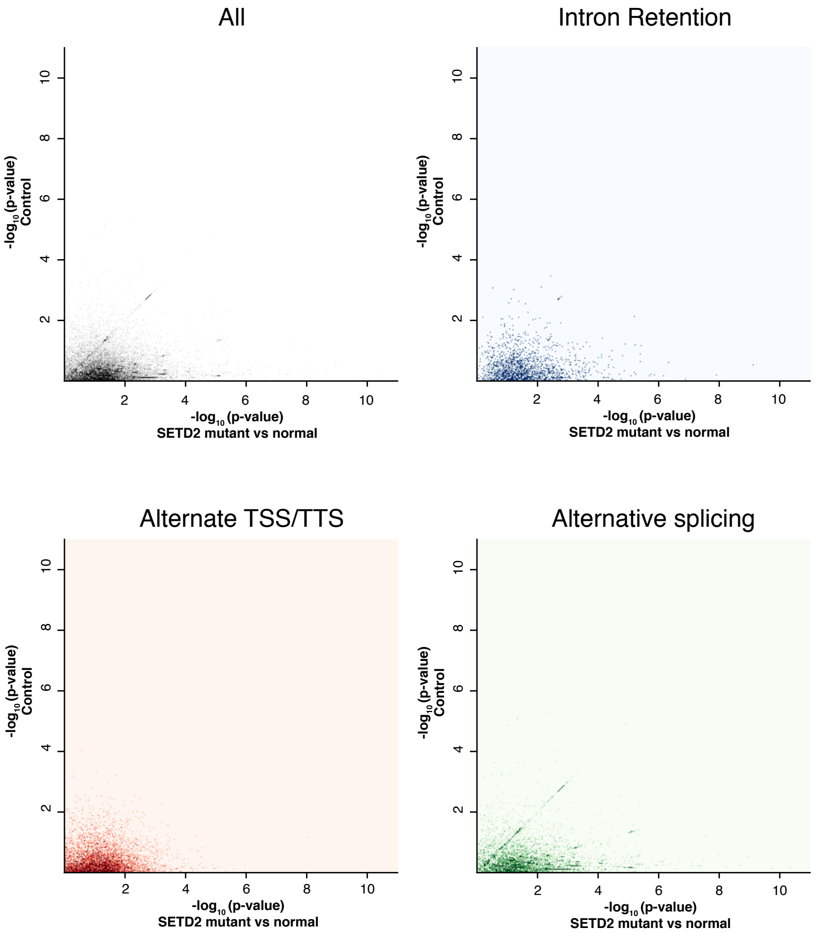


Figure 3.14: *SETD2*-mutant tumors display widespread changes in RNA processing. Significance of the difference in ratios between in *SETD2* mutant and normal tumors (x-axis) are plotted against the scrambled control (y-axis). Combined instances of altered transcript processing (black) can be subdivided as intron retention (blue), alternate transcriptional start or termination sites (red), or alternative splicing (green).

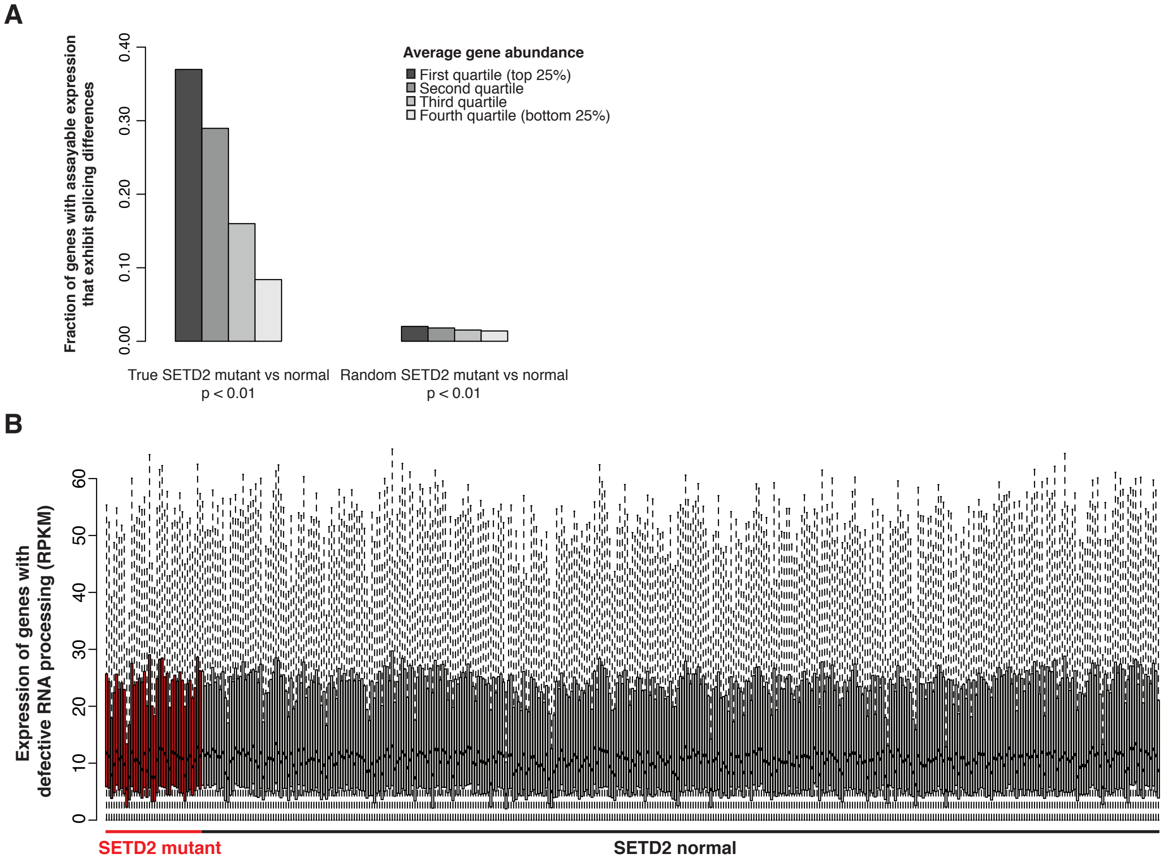


Figure 3.15: Aberrant splicing is preferentially detected in highly transcribed genes. A. RNA abundance for each gene was averaged across all tumors and normal kidney then divided into quartiles. We detected differences in splicing in approximately 38% of the first quartile of genes (top 25% of genes by expression), but only about 8% of genes in the fourth quartile (bottom 25% of genes by expression). B. Overall RNA levels (RPKM) for *SETD2*-mutant (n = 38, marked in red) and *SETD2*-normal tumors (n = 380, marked in black), showing that the expression of genes with defective RNA processing is comparable between these tumor classes.

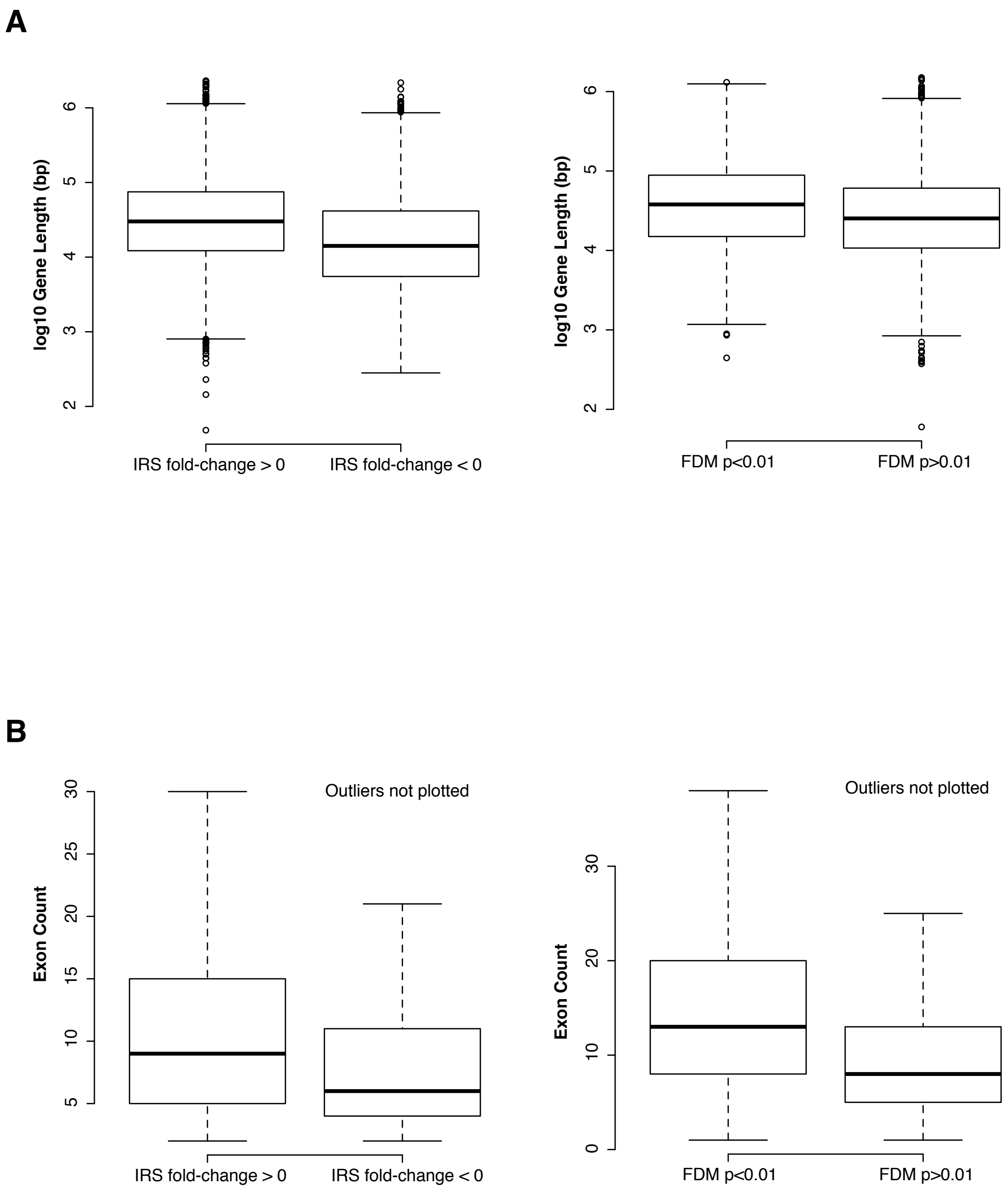


Figure 3.16: Aberrant splicing and intron retention is preferentially detected in long genes and genes with more exons. A. log$_{10}$ gene length (bp) is plotted for genes with increased IRS in H3K36me3-deficient tumors (left panel, left) and genes with decreased IRS in H3K36me3-deficient tumors (left panel, right). log$_{10}$ gene length (bp) is plotted for genes with altered splicing ($p < 0.01$) in SETD2-mutant tumors (right panel, left) and genes without altered splicing ($p > 0.01$) in SETD2-mutant tumors (right panel, right). B. Exon count is plotted for genes with increased IRS in H3K36me3-deficient tumors (left panel, left) and genes with decreased IRS in H3K36me3-deficient tumors (left panel, right). Exon count is plotted for genes with altered splicing ($p < 0.01$) in SETD2-mutant tumors (right panel, left) and genes without altered splicing ($p > 0.01$) in SETD2-mutant tumors (right panel, right).

which the altered inclusion of the downstream exon is related to nucleosome positioning over the exon itself as well as the adjacent upstream nucleosome. Nucleosome positioning and histone modifications (including H3K36me3) are known to regulate multiple processes involved with splicing, including changes in the speed or pausing of RNA polymerase [187] [188] [189] [190] [191], and the ability for splicing machinery to appropriately recognize the splice donor and acceptor. Our finding also suggests that the positioning of this upstream nucleosome may be related to trimethylation of H3K36 on the exonic nucleosome. In *S. cerevisiae*, loss of Set2 leads to destabilization of nucleosomes through hyperacetylation of gene bodies and cryptic transcriptional initiation [185] [186]. Since hyperacetylation was not observed following *SETD2* silencing [171], the increased chromatin accessibility we observed over gene bodies may therefore represent nucleosome destabilization in a hyperacetylation-independent manner. Although our results directly link *SETD2* mutation and H3K36 trimethylation to chromatin accessibility, studies that specifically examine nucleosome positioning and histone modification will be necessary to fully investigate this potential mechanism.

Though our data associate *SETD2* mutations/H3K36me3 deficiency with aberrant RNA processing, exactly how this dysregulation contributes to tumorigenesis remains unknown. A significant fraction of the deregulated transcripts include known tumor suppressors, DNA damage response proteins, and kinases. Strikingly, 58% of genes with altered splicing patterns (Figure 3.12A–B) encode annotated phosphoproteins ($p = 7.3 \times 10^{-109}$), representing an enrichment exceeding that of genes annotated as having alternate splice isoforms ($p = 2 \times 10^{-60}$), a finding also observed in genes exhibiting retained introns (Figures 3.11, 3.17). Alterations in the abundance, stability, or splicing of RNA could induce changes in the phosphoproteome and disrupt normal cellular signaling and growth checkpoints, leading to tumorigenesis. Deregulated signaling as well as transcriptional defects provide numerous putative targets for therapeutic exploitation. Additionally, the application of FAIRE, or IHC for H3K36 trimethylation, could enable the classification of clinical specimens into functional tumor subtypes.

This study advances our understanding the relationship between genetic alterations af-

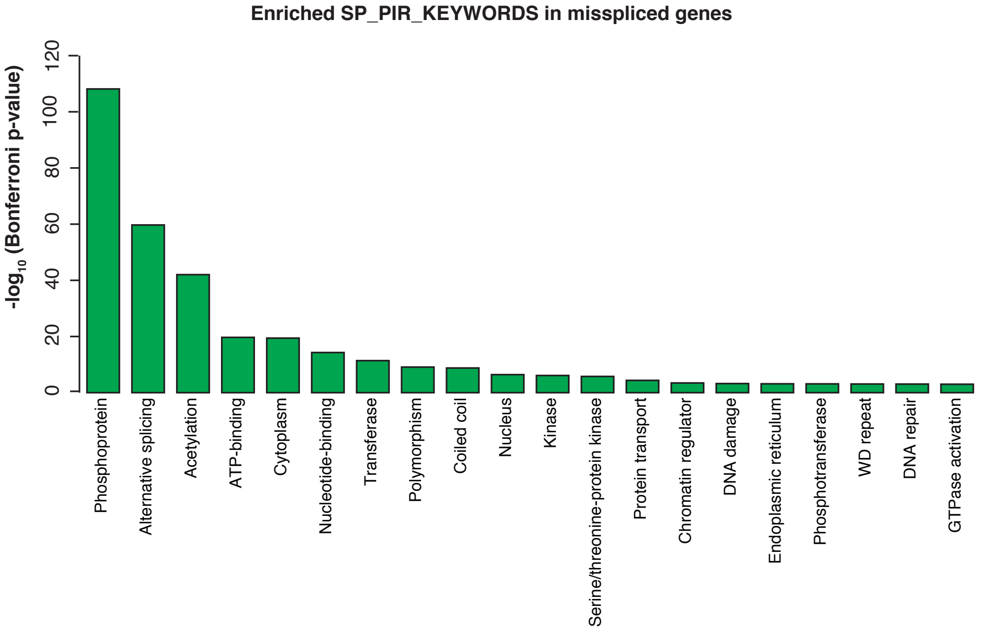


Figure 3.17: Enriched ontologies among misspliced genes. Genes exhibiting significant splicing differences between *SETD2* mutant and normal tumors ($p < 0.003$) in the TCGA cohort were assessed for associated ontologies. The most highly enriched terms in the SP_PIR_KEYWORDS ontology are presented as the -log$_{10}$ of the Bonferroni-corrected p-value. P-values were filtered to $p < 1 \times 10^{-3}$. “Alternative splicing” refers to genes previously annotated as exhibiting alternative splicing.

fecting chromatin organization and alterations in transcription. RNA processing defects in a large fraction of expressed genes, many of which are tumor-suppressors critical for cellular function, may be a common phenotype of many cancers. Comprehensive mapping of the chromatin landscape in primary tumors offers a new tool for understanding the functional consequences of chromatin modifier mutations in human disease.

3.4 Methods

3.4.1 Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) and hierarchical clustering of differentially open chromatin

FAIRE was performed as previously described [74] [75]. Sequencing was performed using 36- or 50-bp single-end reads (Illumina GAIIx or HiSeq 2000). Reads were filtered using TagDust [192] and aligned to the reference human genome (hg19) with Bowtie [193] using default parameters. Reads were counted in 500 bp sliding windows across the genome, normalized for sequencing depth, and adjusted for batch effects using Principal Components Analysis (PCA). One outlier normal kidney sample was removed at this step, and all normal kidney samples were removed for subsequent tumor-only analyses. For clustering analyses, only windows with sufficient sequencing depth (row average > 0.25) were retained; groups were compared using 1- or 2-sided t-tests ($p < 0.01$), clustered and plotted [156]. Feature intersections were computed using BEDtools [154].

3.4.2 Re-processing of HIF1A, HIF2A, and ARNT ChIP-seq data

ChIP-seq reads for HIF1A, HIF2A, and ARNT [180] were filtered using TagDust [192] and aligned to the reference human genome (hg19) using Bowtie [193] requiring unique read placement. Binding sites ($q < 0.05$) for HIF1A and HIF2A were identified using MACS [194], with a shift-size of 250 bp.

3.4.3 Ontologies associated with differentially open chromatin

Regions from Clusters 1–3 were associated with gene ontologies using GREAT [118] using all possible 500 bp windows as background. The top 5 ontologies with $q < 1 \times 10^{-5}$ were presented; full gene ontology associations are supplied in Figure 3.5.

3.4.4 Motif analysis

Significantly over-enriched known Transcription Factor (TF) motifs were identified using HOMER [181]. The 500 bp flanking region was used as local background. Only those TF motifs whose enrichment over background exceeded 2.5-fold, were present in at least 10% of the target sequence, and $q < 0.0001$ were presented in Figure 3.2D. Highly similar entries were merged.

3.4.5 SureSelect custom capture and mutation calling

Genotyping was performed using the SureSelect XT Custom Capture (Agilent). Multiplexing was achieved using TruSeq adapters (Illumina); samples were pooled prior to the capture and amplified post-capture using TruSeq PCR primers (Illumina). Blocking reagents were replaced with water to avoid cross-reactivity. Sequencing was performed using 50-bp paired-end reads (Illumina HiSeq 2000). Reads were aligned to the reference human genome (hg19) using BWA [193]. Genes were sequenced to an average coverage of 200X with 85% of the target sequenced to least at 50X. Genotypes were determined using the Genome Analysis Toolkit (GATK) [195] “Better” protocol. Only high-confidence (quality score > 100) variants predicted to have high or moderate severity and not reported in dbSNP (v129) were considered.

3.4.6 Histone methylation ChIP-seq and data processing

ChIP for H3K36me3 and input DNA from normal kidney was sequenced on the Illumina GAII. Reads were aligned to the reference genome (hg19) using Bowtie requiring unique alignment. H3K36me3 sites were called first using ZINBA [110], then merged to call broader domains by merging two or more nearby (within 5 kb) sites using Galaxy [155]. The average H3K36me3 signal across gene bodies was plotted using CEAS [153].

3.4.7 Feature overlap permutations

Significance of overlap between sites of differentially open chromatin associated with SETD2 or PBRM1 mutations and H3K36me3 sites was determined by permutation. First, the overlap between the actual set of significant windows and histone methylation was computed. Then the same number of randomly-selected windows from the full list (regardless of significance) was selected 1000 times, and an empirical p-value was determined by counting the number of times the overlap of the permuted set exceeded that of the actual set.

3.4.8 Tissue Microarray Construction and Immunohistochemistry

Tissue microarrays (TMAs) were constructed from formalin-fixed, paraffin-embedded tumor blocks from 69 ccRCC tumors and 11 matched normal kidneys collected at the time of nephrectomy. Hematoxylin and eosin stained slides were reviewed to identify a target area of ccRCC histology in each tissue block. TMAs were then constructed using 0.6 mm cores on the manual tissue microarrayer (Beecher Instruments). Tumor and normal samples were represented in triplicate. Sequential 4 µm slides were cut from each TMA.

Immunohistochemical (IHC) staining of H3K36me3, of Histone H3, and SETD2 was performed (Bond Autostainer, Leica Microsystems, Inc.) according to the manufacturer's protocol. Antigen retrieval for H3K36me3, SETD2 and Histone H3 was performed for 30 minutes in Citrate Buffer pH 6.0 (Bond AR9961) and hydrated with Bond wash buffer (AR9590). Slides were incubated with H3K36me3 antibody (Abcam, ab9050, dilution 1:2000) or Histone H3 (courtesy of the Strahl Lab, dilution 1:5000) or SETD2 (Abcam, ab31358, 1:200) for 1 hour at room temperature. Antibody detection was performed (Bond Polymer Refine Detection System, DS9800) followed by image acquisition (ScanScope CS, Aperio Technologies).

Quantification of H3K36me3, SETD2, and Histone H3 was performed independently by two reviewers who were blinded to the tissue identity. The percentage of tumor cells with positive nuclei was determined by evaluating the entire core for each sample. The degree of H3K36me3 or SETD2 staining was averaged across triplicate samples and normalized to total H3 to correct for differences in cell number. Using the minimum value of normalized H3K36me3 in normal kidney as a cutoff, tumors were stratified as either "H3K36me3-normal" or "H3K36me3-deficient" for subsequent analyses. Five additional tumors (not represented on the tissue microarray) were similarly assessed by immunohistochemistry and classified as "H3K36me3-deficient" (3 tumors) or "H3K36me3-normal" (2 tumors).

3.4.9 Intron retention estimates by RNA-seq

Total RNA was isolated from tumors (miRNeasy, Qiagen) and validated to have a median RNA Integrity Numbers (RIN) of 8.6 (minimum 6.8) using a Bioanalyzer (Agilent). Ribosomal RNA was depleted (RiboMinus, Invitrogen) and RNA was fragmented (RNA Fragmentation Reagents, Ambion). cDNA was generated (SuperScript II, Invitrogen) by random priming followed by second strand synthesis (DNA Polymerase I, Enzymatics) and purified (PCR purification kit, Qiagen). Libraries were prepared according to the manufacturer's specifications (Illumina). Sequencing was performed using 50-bp single-end reads (Illumina HiSeq 2000). Reads were aligned to the reference human genome (hg19) using TopHat [196] and gene expression was estimated by calculating RPKM, analyzing only exonic reads. Intron Retention Scores were calculated for each gene as follows:

$$2 \times \frac{\frac{\Sigma\, intronic\ coverage}{\Sigma\, intronic\ length}}{\frac{\Sigma\, intronic\ coverage}{\Sigma\, intronic\ length} + \frac{\Sigma\, exonic\ coverage}{\Sigma\, exonic\ length}}$$

3.4.10 Quantitative RT-PCR

Total RNA extracted from patient tumors (Qiagen miRNeasy) was either rRNA-depleted (Ribo-Minus, Invitrogen) or polyA-selected (Oligotex mRNA Mini Kit, Qiagen). RNA was reverse transcribed by random priming (Supercript II Reverse Transcriptase, Invitrogen) and cDNA was quantified by PCR and normalized to *ABCF* (Maxima SYBR Green/ROX qPCR Master Mix, Thermo Fisher Scientific; 7900HT Fast Real-Time PCR System, Applied Biosystems).

ABCF Forward: 5′CGCCAAGCCATGTTAGAAAATG3′

ABCF Reverse: 5′CGGCTACAATGTACAGGTCTG3′

USH1C Forward1: 5′ACCATCTCCAAACCTGTCATG3′

USH1C Forward2: 5′ATGATCAGGGAGTGGAACC3′

USH1C Reverse: 5′CCATCCTCTTCAACATCTCCTG3′

3.4.11 Differential splicing analysis

RNA-seq reads were aligned to the human reference genome using MapSplice [197]. For each gene, a splicing graph was created as previously described [184]. Each exon and splicing event was represented as an edge, and splice junctions as nodes. We computed a splicing fraction of each edge as the fraction of RNA-seq coverage in that edge divided by the total coverage of all edges sharing one node of that edge. Only edges with coverage exceeding 5 reads and genes with multiple isoforms (13,879 genes) were considered. The node exhibiting the largest difference between *SETD2* mutant and normal tumors was determined by comparing the median of each group. As a control, we created random groups of tumors of the same sizes. Splicing differences between *SETD2* mutant and normal tumors were compared to that of the control group by a Kruskal-Wallis one-way analysis of variance test. The skipped exon ratio was computed as the ratio of coverage of the included exon, and sum of coverages of the included exon and the skipping splice.

CHAPTER 4

ISOLATION OF REGULATORY ELEMENTS FROM ARCHIVAL HUMAN SPECIMENS USING FAIRE

4.1 Introduction

Archiving biopsy and other tissue samples in paraffin following extended formalin fixation (Formalin-Fixed Paraffin Embedded, FFPE) is the standard pathological procedure in hospitals and biobanks. It is estimated that over 1 billion of these samples exist worldwide [198]. The ability to store specimens long-term at room temperature and later assess cellular histology, as well as the relative ease and low cost of use leads to the predominant preference for FFPE archiving over flash-freezing tissues in liquid nitrogen and storing them at -80°C. FFPE-archived specimens, however, have undergone significant manipulations to ensure their histological integrity for long periods of time. After resection or biopsy, the tissue is placed in neutral-buffered formalin (consisting of 3–4% formaldehyde) for 4–48 hours. After fixation, the tissue is gradually dehydrated, passing through a series of graded ethanols and xylenes, then finally embedded in paraffin wax. This process, particularly the extended fixation time, can lead to nucleic acid degradation and modification or damage to DNA through formaldehyde-induced adducts [199].

It has recently been discovered that a modified preparation of chromatin for immunoprecipitation from FFPE (PAT-ChIP) results in similar genome-wide profiles of histone modifications as fresh samples [200] [201]. These results indicate that despite the extended formaldehyde fixation, chromatin not only remained sufficiently intact to probe post-translational modifications of histone proteins, but also was amenable to detection by quantitative PCR and high-throughput sequencing.

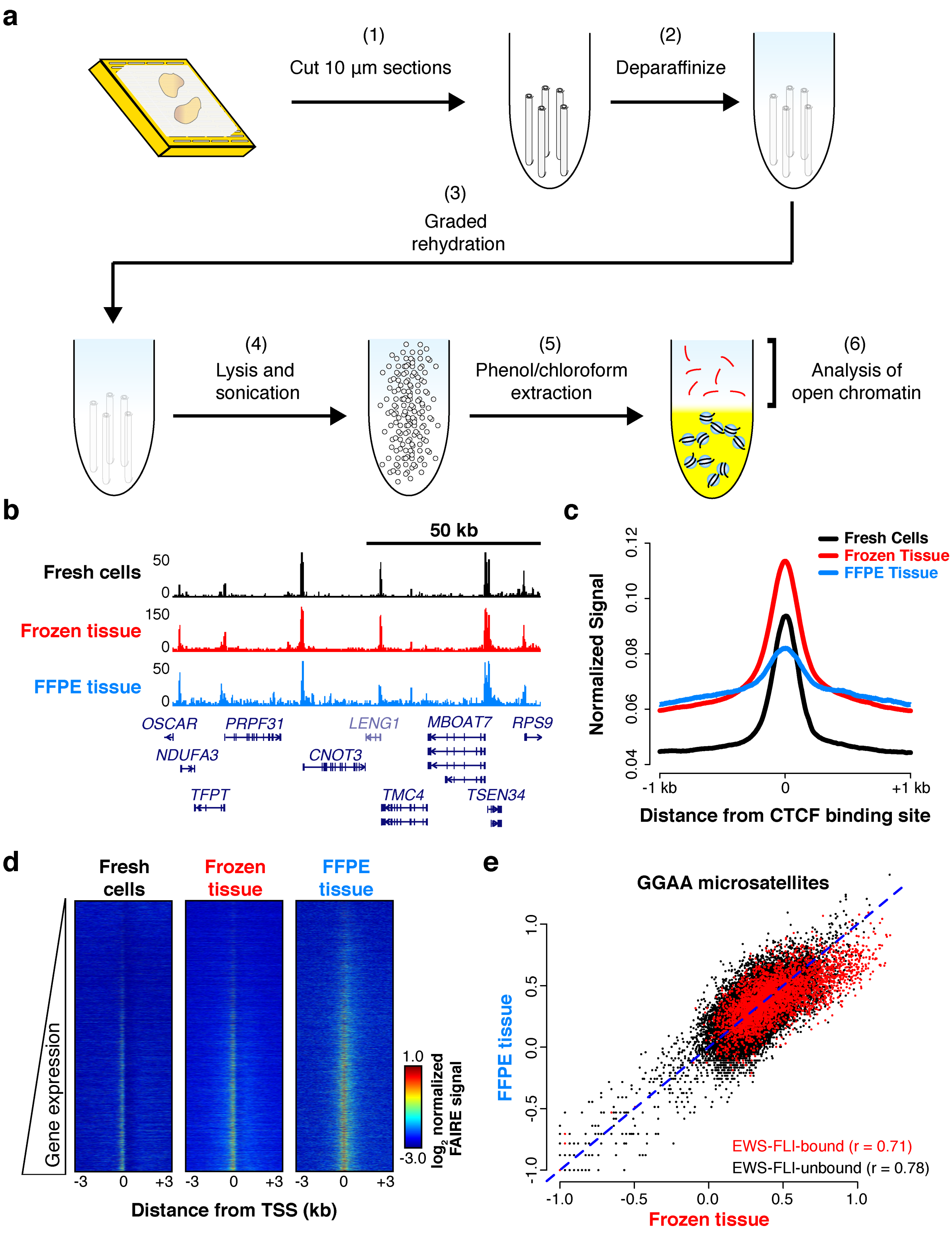
An alternative method of studying chromatin organization is Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE), which utilizes a short fixation time to ensure only histone-DNA interactions are crosslinked [166] [72] [73]. FAIRE has been used extensively to study accessible regions of chromatin in a multitude of eukaryotic cells and tissues [15] [54] [76] [78] [77] [57], and has proven effective at identifying tumor-subtype-specific differences in chromatin accessibility that were linked to RNA processing defects (Chapter 3) [202]. The extended fixation imparted by FFPE preparation, however, could damage and/or over-fragment nucleosome-free DNA or otherwise hinder our ability to identify regulatory elements from these tissues.

Here, we set out to explore whether a modified FAIRE procedure could allow us to overcome these technical challenges and detect biologically relevant regulatory elements from FFPE human tumors. In a highly controlled system that permits the direct comparison of cultured cells, frozen tissue, and FFPE tissue, we show that deparaffinization and rehydration of 10-μm FFPE sections prior to lysis, sonication, and phenol-chloroform extraction leads to the highly concordant detection of both promoter-proximal and distal locations of nucleosome depletion. Moreover, we demonstrate that FFPE-FAIRE is robust to as few as 1 x 10^6 cells, a quantity easily achievable from most specimens. Ongoing work will investigate whether the type, (e.g. carcinoma, sarcoma, blastoma), location (e.g. brain, breast, kidney, bone), or age of the FFPE tumor sample (upwards of 20 years old) plays a role in our ability to identify biologically relevant regions of chromatin accessibility. If successful, this approach will ultimately allow us to follow the effects of cancer therapies longitudinally in single patients, as well as perform large-scale studies of rare diseases. Moreover, its potential clinical relevance could allow for the employment of FAIRE as a high-throughput clinical diagnostic.

4.2 Results

4.2.1 FFPE-FAIRE shows high concordance with frozen tissue and cultured cells in controlled xenograft system

We began our study utilizing a tumor-derived cancer cell line (EWS894). Cells were subcutaneously injected into two NOD scid gamma (NSG) mice to form tumors. Upon resection, tumors from both mice were divided; half of the portions were flash-frozen in liquid nitrogen and stored at -80°C whereas the other half were crosslinked in neutral-buffered formalin for 4–6 hours. Portions of tumors from each mouse were then co-embedded in paraffin (FFPE) as per standard pathological procedures. For FFPE specimens, 10-µm sections were deparaffinized and gradually rehydrated, then lysed, sonicated, and subject to phenol/chloroform extraction as previously described (Figure 4.1a). FAIRE-seq was performed as previously described for cultured cells as well as frozen xenografts [74] [75]. We then compared the open chromatin landscape of EWS894 among cultured cells, frozen tissue, and 1 x 10^6 cell equivalents of FFPE tissue following high-throughput sequencing (Figure 4.1b–e). We found that, in general, there was consistent FAIRE enrichment across all three sample sources at promoter-proximal regions (Figure 4.1b), and that this enrichment correlated with gene expression, as has been previously shown [74] (Figure 4.1d). We also assessed FAIRE enrichment around binding sites for CTCF (Figure 4.1c) as well as a class of distal regulatory elements (GGAA microsatellite repeats) bound in this form of cancer by a translocation-derived transcription factor chimera, EWS-FLI [57] (Figure 4.1e). Although the FAIRE signal at these two classes of distal regulatory elements was somewhat reduced in FFPE tissue (Figure 4.1c, e), a high degree of correlation (Pearson r = 0.78) between frozen and FFPE tissue was nonetheless observed at EWS-FLI-bound GGAA microsatellite repeats, likely the most biologically relevant class of regulatory element in this cell type (Figure 4.1e).


a
(1)
Cut 10 µm sections
(2)
Deparaffinize
(3)
Graded
rehydration
(4)
Lysis and
sonication
(5)
Phenol/chloroform
extraction
(6)
Analysis of
open chromatin
b
50 kb
Fresh cells
50
0
Frozen tissue
150
0
FFPE tissue
50
0
OSCAR
PRPF31
LENG1
MBOAT7
RPS9
NDUFA3
CNOT3
TFPT
TMC4
TSEN34
c
Fresh Cells
Frozen Tissue
FFPE Tissue
Normalized Signal
0.04
0.06
0.08
0.10
0.12
-1 kb
0
+1 kb
Distance from CTCF binding site
d
Fresh
cells
Frozen
tissue
FFPE
tissue
Gene expression
-3
0
+3
-3
0
+3
-3
0
+3
Distance from TSS (kb)
1.0
-3.0
$\log_2$ normalized
FAIRE signal
e
GGAA microsatellites
FFPE tissue
-1.0
-0.5
0.0
0.5
1.0
-1.0
-0.5
0.0
0.5
1.0
Frozen tissue
EWS-FLI-bound (r = 0.71)
EWS-FLI-unbound (r = 0.78)

4.2.2 FFPE-FAIRE is robust to as few as 1 x 10^6 cells

We next wanted to explore the degree to which starting material quantity affects FAIRE signal. In addition to 1 x 10^6 cell equivalents (Figure 4.1), we also performed FAIRE from 5 x 10^5, 2 x 10^6, 1 x 10^7, and 2 x 10^7 cell equivalents from the same FFPE specimen. Surprisingly, we observed a decline in the signal-to-noise as the amount of starting material increased. Moreover, we observed poor library complexity in the sample corresponding to 5 x 10^5 cell equivalents (94% of reads aligned to a non-unique start coordinate; only 35% of reads aligned to non-unique start coordinates for 1 x 10^6 cell equivalents) (Figure 4.2). To quantify these differences, we developed a novel metric termed the Chromatin Integrity Number (ChIN score), akin to an *in silico* quantitative PCR experiment. Based on calculations on cell lines and tissues assayed as part of ENCODE [54], ChIN scores in excess of 0.8 have sufficient signal-to-noise at the five positive control loci tested. ChIN scores maintained stable, as did correlations of signal between frozen and FFPE tissue, from 1–2 x 10^6 cell equivalents, but both declined as starting material increased to 1–2 x 10^7 (Figure 4.2).

4.2.3 Ongoing work

We will soon be scaling up to assay chromatin accessibility across a cohort of 15 patients with diverse forms of cancer: clear cell Renal Cell Carcinoma, luminal and basal-type breast carcinoma, medulloblastoma, and Ewing Sarcoma. This will allow us to investigate whether

Figure 4.1: FFPE-FAIRE shows high concordance with frozen tissue and cultured cells in controlled xenograft system. a. Experimental schema. Deparaffinization and rehydration of 10-μm FFPE sections prior to lysis, sonication, and phenol-chloroform extraction permit the isolation of regulatory elements from FFPE tumor chromatin. b. UCSC Genome Browser screenshot demonstrating high promoter-proximal concordance among sample source types. c. Normalized FAIRE signal from fresh cells (black), frozen tissue (red), and FFPE tissue (blue) around CTCF binding sites identified by ChIP-seq in endothelial cells [14]. d. Heatmap of normalized FAIRE signal 3 kb around TSS ranked by gene expression in Ewing cells. Color was assigned on a $\log_2$ scale of -3 to 1. e. FAIRE signal over EWS-FLI-bound (red) and EWS-FLI-unbound (black) GGAA microsatellites is highly concordant (Pearson r = 0.78) between frozen tissue (x-axis) and FFPE tissue (y-axis).

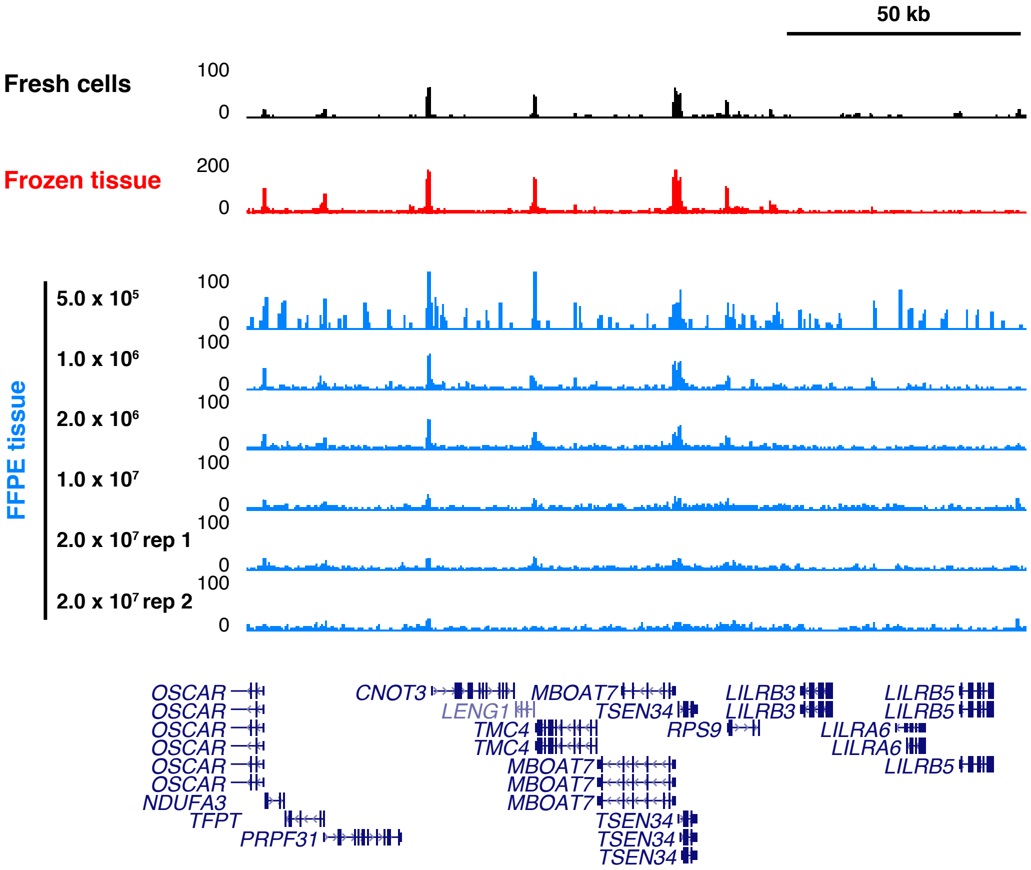


Figure 4.2: FFPE-FAIRE is robust to as few as 1 x 10^6 cells. UCSC Genome Browser screenshot of FAIRE signal from fresh cells (black), frozen tissue (red), and FFPE tissue (blue). Varying FFPE section thickness allowed for the isolation of DNA from a wide range of starting material quantity, from 5 x 10^5 to 2 x 10^7 cell equivalents. Two technical replicates were performed for 2 x 10^7 cell equivalents.

the type, (e.g. carcinoma, sarcoma, blastoma), location (e.g. brain, breast, kidney, bone), or age of the FFPE tumor sample (upwards of 20 years old) plays a role in our ability to identify biologically relevant regions of chromatin accessibility. We will compare FAIRE enrichment in a comprehensive manner between frozen and FFPE tissue for 10 of these tumors, assaying generic promoter-proximal and distal regions, as well as regions relevant to each tumor type (e.g. GGAA microsatellites, Ewing Sarcoma; HIF binding sites, Renal Cell Carcinoma; estrogen receptor (ER) binding sites, luminal subtype breast carcinoma). Additionally, genotype and gene expression data have been gathered for many of these samples; this information can be utilized to link cancer-type-specific differences in chromatin accessibility with cancer-type-specific differences in gene expression.

4.3 Discussion

We report here a modification to the Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) procedure for use on Formalin-Fixed Paraffin-Embedded (FFPE) tissue to utilize clinically annotated human specimens available in hospitals and biobanks worldwide. In a highly controlled xenograft system, we directly compared chromatin accessibility from cultured cells, frozen tissue, and FFPE tissue. We showed that both promoter-proximal and distal locations of nucleosome depletion were highly concordant among these tissue sources. Moreover, we investigated whether the quantity of starting material affected FAIRE signal and whether there was a lower limit for detection of accessible regions of chromatin from FFPE samples. We demonstrated that FFPE-FAIRE is robust to as few as 1 x 10^6 cells, a quantity easily achievable from most specimens, but FAIRE-seq library complexity degraded significantly with 5 x 10^5 cells. Ongoing work will investigate whether the type, location, or age of the FFPE tumor sample plays a role in our ability to identify biologically relevant regions of chromatin accessibility and tumor-type-specific differences in chromatin accessibility that coincide with differences in gene expression or genotype.

Our results demonstrate that despite numerous technical challenges, the modified FAIRE technique allows for the detection of many biologically relevant regulatory elements from FFPE tumor chromatin. This approach will ultimately enable novel studies of rare diseases, molecular consequences of therapeutic intervention, and pharmacological efficacy using chromatin accessibility as a readout. Additionally, we have already shown that when coupled with robotics, FAIRE can be used in an automated manner to screen small molecules designed to inhibit chromatin regulators, and that lead compounds from the screen confer inhibition of tumor cell growth in soft agar, a hallmark of oncogenic transformation [Pattenden et al, *manuscript in preparation*]. If similarly coupled with robotics, FFPE-FAIRE could be used in automated screens of small molecule inhibitors to both prospectively and retrospectively identify the best course of therapeutic action across hundreds to thousands of patients simultaneously.

4.4 Methods

4.4.1 Xenograft model and estimation of tumor nuclei density

Approximately 5 x 10^6 cancer cells (EWS894) were subcutaneously injected using matrigel bilaterally into two NOD scid gamma (NSG) mice (JAX laboratories) to form tumors. Upon resection, tumors from both mice were divided; half of the portions were flash-frozen in liquid nitrogen and stored at -80°C whereas the other half were crosslinked in neutral-buffered formalin for 4-6 hours. Portions of tumors from each mouse were then co-embedded in paraffin (FFPE) as per standard pathological procedures. Hematoxylin and Eosin staining was performed on the Leica Autostainer XL from 5-µm FFPE sections. Nuclei counts were estimated by Definiens Tissue Studio version 3.5.1. Paraffin embedding, sectioning, staining, and nuclei count estimates were performed by the UNC Translational Pathology Laboratory.

4.4.2 FFPE-FAIRE

For FFPE specimens, 10-µm sections were deparaffinized in six consecutive 10-minute washes (rocking at room temperature) in 1 mL BiOstic deparaffinization solution (MO-BIO

Laboratories, Carlsbad, CA), each followed by centrifugation at 20,000 x g for 2 minutes at room temperature. The pellets were then gradually rehydrated through graded ethanols (1 mL of 100%, 95%, 70%, 50%, 20%, 0% ethanol), each rocking for 5 minutes at room temperature followed by centrifugation at 20,000 x g for 2 minutes at room temperature. The final rehydrated pellets were then resuspended in 1 mL FAIRE lysis buffer and lysed by bead-beating, sonicated, and subjected to phenol/chloroform extraction as previously described [74] [75]. FAIRE-seq libraries were prepared for sequencing using TruSeq barcoded adapters as per manufacturers instructions (Illumina), and sequenced as single-end 50-bp reads. Reads with significant adapter contribution were removed using TagDust [192] and assessed to ensure their high quality using the FASTX-Toolkit. Reads from frozen and FFPE tissues were first aligned to the reference mouse genome (mm9) using Bowtie [203] [152] to remove any murine contamination. Non-mouse reads were then aligned to the reference human genome (hg19) using Bowtie, allowing for two mismatches and reads to align to up to four locations in the genome, though only the best-scoring alignment was used. Data visualization was achieved using the UCSC Genome Browser .

4.4.3 Reanalysis of EWS-FLI binding sites and GGAA microsatellites

EWS-FLI ChIP-seq data from a Ewing Sarcoma cell line (EWS502) was aligned and reanalyzed as was previously described [57], except to the reference human genome (hg19). GGAA microsatellites defined by RepeatMasker were divided based on their EWS-FLI binding status using BEDtools [154]. Signal at GGAA microsatellites was calculated based on the number of reads overlapping the repetitive element, normalized by total sequencing depth.

4.4.4 ChIN score calculation

FAIRE signal was computed at 500-bp windows near the TSS (positive controls) of five genes (*MBOAT7*, *CNOT3*, *BC006361*, *AURKIP1*, *EIF3F*) as well as nearby 500-bp negative control windows. The ChIN scores were then calculated as:

$$\frac{\Sigma\ signal\ at\ positive\ controls}{\Sigma\ signal\ at\ positive\ controls + \Sigma\ signal\ at\ negative\ controls}$$

Scores range from 0 to 1, where ChIN scores closest to 1.0 represent samples with optimal signal-to-noise. ChIN scores greater than 0.8 were considered to have sufficient quality.

The genomic coordinates (hg19) utilized to calculate ChIN scores are as follows:

Gene Symbol	**Control Type**	**Coordinates**
MBOAT7	positive	chr19:54693549–54694049
MBOAT7	negative	chr19:54678519–54679018
BC006361	positive	chr1:713868–714368
BC006361	negative	chr1:744569–745068
AURKIP1	positive	chr1:1310612–1311112
AURKIP1	negative	chr1:1319556–1320055
CNOT3	positive	chr19:54640870–54641370
CNOT3	negative	chr19:54648720–54649219
EIF3F	positive	chr11:8008506–8009006
EIF3F	negative	chr11:8015703–8016202

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

The advent of high-throughout DNA sequencing enabled many large-scale studies of mutation types and frequencies across thousands of cancer classes. An emerging theme is that genes that encode chromatin regulators are frequent, but also disproportionately prevalent in pediatric and hematological malignancies. The molecular consequences of chromatin regulator mutations on a genome-wide scale, and moreover, how other genetic insults drive chromatin dysregulation and potentially enhance tumorigenesis, were until now completely unknown. The preceding work leveraged two model cancer systems, Ewing Sarcoma and clear cell Renal Cell Carcinoma, through which we were able to better understand the causes and consequences of chromatin dysregulation.

In Ewing Sarcoma, a fusion oncoprotein (EWS-FLI) is formed as a result of a chromosomal translocation. This transcription factor chimera exhibited altered binding properties when compared to its DNA-binding parental protein FLI1 despite identical ETS-family DNA binding domains. Instead of localizing to canonical ETS motifs, EWS-FLI bound a subset of microsatellite repeats, ones that contain a multimerization of the GGAA core of the ETS sequence motif. We found that this differential targeting was influenced by epigenetic factors. These microsatellite regions were atypically marked with an enhancer-like signature, were bound by RNA polymerase II, and resided in nucleosome-depleted regions. Our data suggest that the chromatin-modifying activity conferred to EWS-FLI through chimerism is necessary and sufficient to alter the local chromatin landscape. The observation of widespread FAIRE enrichment of repetitive regions may also suggest that a favorable chromatin landscape permissive of EWS-FLI targeting may exist in the likely cell type of origin (mesenchymal stem cells), a question currently being explored.

Transcription factors themselves have long been believed to be undruggable for several reasons. First, designing small molecule inhibitors that block DNA-binding activity with any specificity is difficult due to sequence and structural properties shared among many transcription factors or transcription factor families. Second, in cases such as c-Myc, the transcription factor carries out many functions, only a subset of which are oncogenic, meaning the therapy would need high specificity toward cancerous cells to avoid potent off-target effects. The fact that EWS-FLI not only acts as a potent transcription factor but carries with it chromatin-modifying activity creates an exciting and novel therapeutic window. If, through the use of small molecules, we can revert the EWS-FLI-bound sites to a nucleosomal state, then perhaps its stunted potency would specifically inhibit cancer cell viability.

This is currently the main focus question of an ongoing study in our lab. We hypothesized that a drug-induced inhibition of chromatin-modifying activity would echo that of EWS-FLI silencing by siRNA, as we saw using FAIRE-qPCR in Figure 2.21C. We therefore designed a novel screen in which we could probe changes in chromatin accessibility, specifically at EWS-FLI-bound genomic loci but not other regions, after the administration of one of nearly 1,000 chemical probes designed to inhibit various epigenetic modifiers. Due to the scale of the experiment, traditional phenol-chloroform-based FAIRE would not be a suitable approach to measure chromatin accessibility. Instead, we modified the FAIRE procedure such that organic extraction was replaced with a simple column-based purification. The new method behaved almost identically to that of traditional phenol-chloroform-based FAIRE, but more importantly, it was amenable to robotics and could be performed in 96-well plates. Using column-based FAIRE-qPCR as a readout, this small molecule screen has yielded many interesting lead compounds and compound classes. One such compound has now been shown by our lab to inhibit growth in soft agar at low doses, a hallmark of anchorage-independent growth commonly exhibited by cancer cells, and at higher doses, drastically inhibit cell viability. Moreover, preliminary data show that these cellular phenotypic changes occur much later than the chromatin-based findings of the initial screen, suggestive of a high level of specificity toward EWS-FLI-bound targets rather than a simple (but potent) cytotoxic. Our ability to measure the direct effects of inhibition of epigenetic modifiers is an exciting and novel means to discover and develop efficacious cancer therapies. This study is also exciting from a basic research standpoint because lead compounds may shed light on specific EWS-FLI complex members, a common goal that has been difficult to study due to the inherently poor complex stability.

In addition to ascribing novel chromatin-modifying functions to fusion oncoproteins and inhibiting those effects therapeutically, we also investigated how mutations in chromatin regulators themselves alter the chromatin landscape, with the ultimate goal of understanding their oncogenicity. We utilized clear cell Renal Cell Carcinoma for this study, a cancer that carries several recurrent mutations in chromatin regulators, including modulators of histone methylation as well as nucleosome positioning. Despite their prevalence in human cancers, especially pediatric and hematological malignancies (as discussed in Chapter 1), the functional consequences of chromatin regulator mutations had not yet been investigated.

Again utilizing FAIRE, in combination with high-throughput genotyping and transcriptomics, we probed 49 primary human tissue samples, including some matched tumor and normal kidney, and associated variation in chromatin accessibility with mutations in *SETD2*, the histone H3 lysine-36 tri-methyltransferase. We observed widespread increases in chromatin accessibility especially in gene bodies typically harboring H3K36me3 in normal kidney tissue. Though we expected this chromatin-based effect to manifest itself in differences in transcript abundance, we instead noticed a far more widespread phenotype. These genes marked by H3K36me3 in normal kidney that exhibited alterations of chromatin in tumors, as well as many other genes, exhibited RNA processing defects, specifically the retention of introns. Since our data were generated from pools of total RNA, we sought to validate this phenotype in mature (polyA$^+$) RNA, and in a much larger cohort of primary tumor samples analyzed by The Cancer Genome Atlas. These data supported a more diverse model for transcriptional defects, including retention of introns, missplicing of exons, and usage of alternative transcriptional initiation or termination sites. These defects were widespread, affecting nearly 25% of all expressed genes, and defects were more common in highly transcribed genes.

Moreover, we found a surprising increase in chromatin accessibility immediately upstream (50 bp) of misspliced exons in *SETD2*-mutated tumors. This result suggests a mechanism by which the altered inclusion of the downstream exon is related to nucleosome positioning over the exon itself as well as the adjacent upstream nucleosome. Nucleosome positioning and histone modifications (including H3K36me3) are known to regulate multiple processes involved with splicing, including changes in the speed or pausing of RNA polymerase [187] [188] [189] [190] [191], and the ability for splicing machinery to appropriately recognize the splice donor and acceptor. Our finding also suggests that the positioning of this upstream nucleosome may be related to trimethylation of H3K36 on the exonic nucleosome. Although our results directly link *SETD2* mutation and H3K36 trimethylation to chromatin accessibility, studies that specifically examine nucleosome positioning and histone modification will be necessary to fully investigate this potential mechanism. Experiments utilizing micrococal nuclease (MNase) digestion to identify the exact positions of nucleosomes in and around misspliced exon starts would further our understanding of the molecular mechanisms of the RNA processing defects, but these experiments can not yet be easily performed in primary tumors. Instead, we will leverage a new cell line model developed by our group in which *SETD2* has been completely silenced through the use of TALENs. We have shown preliminarily that *SETD2*-knockout cells lack H3K36me3 and exhibit some of the same RNA processing defects. Future work with this model system will explore nucleosome positioning, spliceosome recruitment, and RNA polymerase binding and kinetics in the presence and absence of SETD2/H3K36me3.

Though our data associate *SETD2* mutations/H3K36me3 deficiency with aberrant RNA processing, exactly how this dysregulation contributes to tumorigenesis also remains unknown. Whether the defects in processing lead directly to the inhibition of tumor suppressors or the modulation or induction of oncogenes has not yet been investigated. Future work again utilizing cultured cells with or without *SETD2* perhaps in conjunction with a siRNA screen may inform us as to which key genes, when either directly silenced by siRNA or inhibited through *SETD2*-knockout-induced RNA processing defects, govern critical cellular processes.

Another means for better understanding the underlying tumorigenic behavior induced by SETD2/H3K36me3 loss is through a synthetic lethality screen. If we can identify compounds that specifically hinder the growth of cancer cells lacking SETD2, then perhaps they are helping to reverse the induction of crucial oncogenes or the silencing of specific tumor suppressors. This screen is a current focus of our lab, and we believe it will not only shed light on underlying biology, but also yield numerous lead compounds for the treatment of clear cell Renal Cell Carcinomas harboring *SETD2* mutation.

An additional important contribution is the ability to now probe chromatin accessibility in Formalin-Fixed Paraffin-Embedded (FFPE) tissues. This tissue archiving technique is by far the most common, as opposed to freezing tissue and storing at -80°C, and is traditionally used by pathologists in disease diagnosis by studying cellular morphology. Their massive availability coupled with clinical annotations makes these specimens an attractive source of biological material for use in numerous RNA- or DNA-based assays. The extended fixation in formalin critical to this archiving process, however, is known to damage nucleic acids in multiple ways. Despite the numerous technical barriers, we showed that both promoter-proximal and distal locations of nucleosome depletion were highly concordant among cultured cells, frozen tissue, and FFPE tissue in a highly controlled xenograft system. Moreover, we found that there is a lower limit for detection of accessible regions of chromatin from FFPE samples; FAIRE-seq library complexity degraded significantly with 5 x 10^5 cells but was robust to 1 x 10^6 cells. Ongoing work will investigate whether the type, location, or age of the FFPE tumor sample plays a role in our ability to identify biologically relevant regions of chromatin accessibility and tumor-type-specific differences in chromatin accessibility that coincide with differences in gene expression or genotype.

This approach will ultimately enable novel large-scale studies of rare diseases, allow us to understand the molecular consequences of therapeutic intervention, and both prospectively and retrospectively identify the best course of therapeutic action across hundreds to thousands of patients simultaneously. In addition, if data are obtained across many different forms of cancer, we can design clinical diagnostics based on differentially accessible regions of chromatin, similar to how differential gene expression is currently used as a diagnostic in systems such as breast cancer. Considerable efforts will be needed to not only identify the genomic regions with largest diagnostic and prognostic power, but also design the algorithms with which tumors can be assigned to a type and subtype *de novo* with extremely high specificity. Clinical usage of FAIRE from FFPE material would also likely require significant improvements to the methodology itself to permit its automation on a large scale. Future work will explore the usage of specially designed 96-well plates to more quickly and efficiently deparaffinize and rehydrate tissues prior to their lysis and downstream manipulations.

Together, these studies advance our understanding of chromatin dysregulation in cancer in multiple aspects. We now better understand the relationship between genetic alterations affecting chromatin organization and alterations in transcription, and moreover, RNA processing defects mediated by chromatin or other means may be a common phenotype of many cancers. We also discovered epigenetic alterations driven by a chimeric transcription factor, which led to preliminary yet exciting new therapeutic avenues that reverse these chromatin-modifying effects and inhibit cancer cell growth. Though much work will be needed, chromatin accessibility-based assays will be at the crux of many new powerful tools for cancer diagnosis and therapy.

BIBLIOGRAPHY

[1] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, pp. 376–80, 2005. 1

[2] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church, "Accurate multiplex polony sequencing of an evolved bacterial genome," *Science*, vol. 309, pp. 1728–32, 2005. 1

[3] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, pp. 823–37, 2007. 1, 52

[4] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, pp. 553–60, 2007. 1, 34

[5] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, "Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing," *Nat Methods*, vol. 4, pp. 651–7, 2007. 1

[6] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-dna interactions," *Science*, vol. 316, pp. 1497 – 1502, 2007. 1

[7] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, pp. 311–22, 2008. 1

[8] J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos, "Global mapping of protein-dna interactions in vivo by digital genomic footprinting," *Nature methods*, vol. 6, pp. 283–9, 2009. 1

[9] D. E. Schones, K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, "Dynamic regulation of nucleosome positioning in the human genome," *Cell*, vol. 132, pp. 887–98, 2008. 1

[10] P. G. Giresi and J. D. Lieb, "Isolation of active regulatory elements from eukaryotic chromatin using faire (formaldehyde assisted isolation of regulatory elements)," *Methods*, vol. 48, pp. 233–9, 2009. 1, 6, 48

[11] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch, "Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis," *Nucleic acids research*, vol. 33, pp. 5868–77, 2005. 1

[12] J. Zhao, T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky, and J. T. Lee, "Genome-wide identification of polycomb-associated rnas by rip-seq," *Molecular cell*, vol. 40, pp. 939–53, 2010. 1

[13] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nat Meth*, vol. 5, pp. 621–628, 2008. 1

[14] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, pp. 57–74, 2012. 1, 4, 85

[15] K. J. Gaulton, T. Nammo, L. Pasquali, J. M. Simon, P. G. Giresi, M. P. Fogarty, T. M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, T. Berney, E. Montanya, K. L. Mohlke, J. D. Lieb, and J. Ferrer, "A map of open chromatin in human pancreatic islets," *Nat Genet*, vol. 42, pp. 255–259, 2010. 1, 4, 6, 48, 82

[16] C. G. A. R. Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–8, 2008. 2, 4

[17] T. C. G. A. R. Network, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, pp. 519–25, 2012. 2, 4

[18] T. C. G. A. Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012. 2, 4

[19] T. C. G. A. Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, pp. 330–7, 2012. 2, 4

[20] T. C. G. A. R. Network, "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia," *The New England journal of medicine*, vol. 368, pp. 2059–74, 2013. 2, 4

[21] C. Kandoth, N. Schultz, A. D. Cherniack, R. Akbani, Y. Liu, H. Shen, A. G. Robertson, I. Pashtan, R. Shen, C. C. Benz, C. Yau, P. W. Laird, L. Ding, W. Zhang, G. B. Mills, R. Kucherlapati, E. R. Mardis, and D. A. Levine, "Integrated genomic characterization of endometrial carcinoma," *Nature*, vol. 497, pp. 67–73, 2013. 2, 4

[22] T. C. G. A. R. Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, pp. 43–9, 2013. 2, 4

[23] T. J. Pugh, S. D. Weeraratne, T. C. Archer, D. A. Pomeranz Krummel, D. Auclair, J. Bochicchio, M. O. Carneiro, S. L. Carter, K. Cibulskis, R. L. Erlich, H. Greulich, M. S. Lawrence, N. J. Lennon, A. McKenna, J. Meldrim, A. H. Ramos, M. G. Ross, C. Russ, E. Shefler, A. Sivachenko, B. Sogoloff, P. Stojanov, P. Tamayo, J. P. Mesirov, V. Amani, N. Teider, S. Sengupta, J. P. Francois, P. A. Northcott, M. D. Taylor, F. Yu, G. R. Crabtree, A. G. Kautzman, S. B. Gabriel, G. Getz, N. Jager, D. T. Jones, P. Lichter, S. M. Pfister, T. M. Roberts, M. Meyerson, S. L. Pomeroy, and Y. J. Cho, "Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations," *Nature*, vol. 488, pp. 106–10, 2012. 2, 4

[24] T. J. Pugh, O. Morozova, E. F. Attiyeh, S. Asgharzadeh, J. S. Wei, D. Auclair, S. L. Carter, K. Cibulskis, M. Hanna, A. Kiezun, J. Kim, M. S. Lawrence, L. Lichenstein, A. McKenna, C. S. Pedamallu, A. H. Ramos, E. Shefler, A. Sivachenko, C. Sougnez, C. Stewart, A. Ally, I. Birol, R. Chiu, R. D. Corbett, M. Hirst, S. D. Jackman, B. Kamoh, A. H. Khodabakshi, M. Krzywinski, A. Lo, R. A. Moore, K. L. Mungall, J. Qian, A. Tam, N. Thiessen, Y. Zhao, K. A. Cole, M. Diamond, S. J. Diskin, Y. P. Mosse, A. C. Wood, L. Ji, R. Sposto, T. Badgett, W. B. London, Y. Moyer, J. M. Gastier-Foster, M. A. Smith, J. M. Auvil, D. S. Gerhard, M. D. Hogarty, S. J. Jones, E. S. Lander, S. B. Gabriel, G. Getz, R. C. Seeger, J. Khan, M. A. Marra, M. Meyerson, and J. M. Maris, "The genetic landscape of high-risk neuroblastoma," *Nature genetics*, vol. 45, pp. 279–84, 2013. 2, 4

[25] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, pp. 214–8, 2013. 2, 4

[26] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinsk, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki,

A. N. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, pp. 415–21, 2013. 2, 4

[27] H. E. Varmus, "The molecular genetics of cellular oncogenes," *Annual review of genetics*, vol. 18, pp. 553–612, 1984. 2

[28] T. J. Giordano, R. Kuick, D. G. Thomas, D. E. Misek, M. Vinco, D. Sanders, Z. Zhu, R. Ciampi, M. Roh, K. Shedden, P. Gauger, G. Doherty, N. W. Thompson, S. Hanash, R. J. Koenig, and Y. E. Nikiforov, "Molecular classification of papillary thyroid carcinoma: distinct braf, ras, and ret/ptc mutation-specific gene expression profiles discovered by dna microarray analysis," *Oncogene*, vol. 24, pp. 6646–56, 2005. 2

[29] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, and L. A. Akslen, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747 – 752, 2000. 2

[30] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer Cell*, vol. 17, pp. 98–110, 2010. 2

[31] G. Duns, E. van den Berg, I. van Duivenbode, J. Osinga, H. Hollema, R. M. Hofstra, and K. Kok, "Histone methyltransferase gene setd2 is a novel tumor suppressor gene in clear cell renal cell carcinoma," *Cancer research*, vol. 70, pp. 4287–91, 2010. 3, 48

[32] I. Varela, P. Tarpey, K. Raine, D. Huang, C. K. Ong, P. Stephens, H. Davies, D. Jones, M. L. Lin, J. Teague, G. Bignell, A. Butler, J. Cho, G. L. Dalgliesh, D. Galappaththige, C. Greenman, C. Hardy, M. Jia, C. Latimer, K. W. Lau, J. Marshall, S. McLaren, A. Menzies, L. Mudie, L. Stebbings, D. A. Largaespada, L. F. Wessels, S. Richard, R. J. Kahnoski, J. Anema, D. A. Tuveson, P. A. Perez-Mancera, V. Mustonen, A. Fischer, D. J. Adams, A. Rust, W. Chan-on, C. Subimerb, K. Dykema, K. Furge, P. J. Campbell, B. T. Teh, M. R. Stratton, and P. A. Futreal, "Exome sequencing identifies frequent mutation of the swi/snf complex gene pbrm1 in renal carcinoma," *Nature*, vol. 469, pp. 539–42, 2011. 3, 4, 47

[33] G. van Haaften, G. L. Dalgliesh, H. Davies, L. Chen, G. Bignell, C. Greenman, S. Edkins, C. Hardy, S. O'Meara, J. Teague, A. Butler, J. Hinton, C. Latimer, J. Andrews, S. Barthorpe, D. Beare, G. Buck, P. J. Campbell, J. Cole, S. Forbes, M. Jia, D. Jones, C. Y. Kok, C. Leroy, M. L. Lin, D. J. McBride, M. Maddison, S. Maquire,

K. McLay, A. Menzies, T. Mironenko, L. Mulderrig, L. Mudie, E. Pleasance, R. Shepherd, R. Smith, L. Stebbings, P. Stephens, G. Tang, P. S. Tarpey, R. Turner, K. Turrell, J. Varian, S. West, S. Widaa, P. Wray, V. P. Collins, K. Ichimura, S. Law, J. Wong, S. T. Yuen, S. Y. Leung, G. Tonon, R. A. DePinho, Y. T. Tai, K. C. Anderson, R. J. Kahnoski, A. Massie, S. K. Khoo, B. T. Teh, M. R. Stratton, and P. A. Futreal, "Somatic mutations of the histone h3k27 demethylase gene utx in human cancer," *Nature genetics*, vol. 41, pp. 521–3, 2009. 3

[34] R. D. Morin, M. Mendez-Lago, A. J. Mungall, R. Goya, K. L. Mungall, R. D. Corbett, N. A. Johnson, T. M. Severson, R. Chiu, M. Field, S. Jackman, M. Krzywinski, D. W. Scott, D. L. Trinh, J. Tamura-Wells, S. Li, M. R. Firme, S. Rogic, M. Griffith, S. Chan, O. Yakovenko, I. M. Meyer, E. Y. Zhao, D. Smailus, M. Moksa, S. Chittaranjan, L. Rimsza, A. Brooks-Wilson, J. J. Spinelli, S. Ben-Neriah, B. Meissner, B. Woolcock, M. Boyle, H. McDonald, A. Tam, Y. Zhao, A. Delaney, T. Zeng, K. Tse, Y. Butterfield, I. Birol, R. Holt, J. Schein, D. E. Horsman, R. Moore, S. J. Jones, J. M. Connors, M. Hirst, R. D. Gascoyne, and M. A. Marra, "Frequent mutation of histone-modifying genes in non-hodgkin lymphoma," *Nature*, vol. 476, pp. 298–303, 2011. 3

[35] J. Schwartzentruber, A. Korshunov, X. Y. Liu, D. T. Jones, E. Pfaff, K. Jacob, D. Sturm, A. M. Fontebasso, D. A. Quang, M. Tonjes, V. Hovestadt, S. Albrecht, M. Kool, A. Nantel, C. Konermann, A. Lindroth, N. Jager, T. Rausch, M. Ryzhova, J. O. Korbel, T. Hielscher, P. Hauser, M. Garami, A. Klekner, L. Bognar, M. Ebinger, M. U. Schuhmann, W. Scheurlen, A. Pekrun, M. C. Fruhwald, W. Roggendorf, C. Kramm, M. Durken, J. Atkinson, P. Lepage, A. Montpetit, M. Zakrzewska, K. Zakrzewski, P. P. Liberski, Z. Dong, P. Siegel, A. E. Kulozik, M. Zapatka, A. Guha, D. Malkin, J. Felsberg, G. Reifenberger, A. von Deimling, K. Ichimura, V. P. Collins, H. Witt, T. Milde, O. Witt, C. Zhang, P. Castelo-Branco, P. Lichter, D. Faury, U. Tabori, C. Plass, J. Majewski, S. M. Pfister, and N. Jabado, "Driver mutations in histone h3.3 and chromatin remodelling genes in paediatric glioblastoma," *Nature*, vol. 482, pp. 226–31, 2012. 3

[36] A. Fujimoto, Y. Totoki, T. Abe, K. A. Boroevich, F. Hosoda, H. H. Nguyen, M. Aoki, N. Hosono, M. Kubo, F. Miya, Y. Arai, H. Takahashi, T. Shirakihara, M. Nagasaki, T. Shibuya, K. Nakano, K. Watanabe-Makino, H. Tanaka, H. Nakamura, J. Kusuda, H. Ojima, K. Shimada, T. Okusaka, M. Ueno, Y. Shigekawa, Y. Kawakami, K. Arihiro, H. Ohdan, K. Gotoh, O. Ishikawa, S. I. Ariizumi, M. Yamamoto, T. Yamada, K. Chayama, T. Kosuge, H. Yamaue, N. Kamatani, S. Miyano, H. Nakagama, Y. Nakamura, T. Tsunoda, T. Shibata, and H. Nakagawa, "Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators," *Nature genetics*, vol. 44, pp. 760–764, 2012. 3

[37] Z. J. Zang, I. Cutcutache, S. L. Poon, S. L. Zhang, J. R. McPherson, J. Tao, V. Rajasegaran, H. L. Heng, N. Deng, A. Gan, K. H. Lim, C. K. Ong, D. Huang, S. Y. Chin, I. B. Tan, C. C. Ng, W. Yu, Y. Wu, M. Lee, J. Wu, D. Poh, W. K. Wan, S. Y. Rha, J. So, M. Salto-Tellez, K. G. Yeoh, W. K. Wong, Y. J. Zhu, P. A. Futreal, B. Pang, Y. Ruan, A. M. Hillmer, D. Bertrand, N. Nagarajan, S. Rozen, B. T. Teh, and P. Tan, "Exome

sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes," *Nature genetics*, vol. 44, pp. 570–4, 2012. 3

[38] A. Dolnik, J. C. Engelmann, M. Scharfenberger-Schmeer, J. Mauch, S. Kelkenberg-Schade, B. Haldemann, T. Fries, J. Kronke, M. W. Kuhn, P. Paschka, S. Kayser, S. Wolf, V. I. Gaidzik, R. F. Schlenk, F. G. Rucker, H. Dohner, C. Lottaz, K. Dohner, and L. Bullinger, "Commonly altered genomic regions in acute myeloid leukemia are enriched for somatic mutations involved in chromatin remodeling and splicing," *Blood*, vol. 120, pp. e83–92, 2012. 3

[39] D. T. Jones, N. Jager, M. Kool, T. Zichner, B. Hutter, M. Sultan, Y. J. Cho, T. J. Pugh, V. Hovestadt, A. M. Stutz, T. Rausch, H. J. Warnatz, M. Ryzhova, S. Bender, D. Sturm, S. Pleier, H. Cin, E. Pfaff, L. Sieber, A. Wittmann, M. Remke, H. Witt, S. Hutter, T. Tzaridis, J. Weischenfeldt, B. Raeder, M. Avci, V. Amstislavskiy, M. Zapatka, U. D. Weber, Q. Wang, B. Lasitschka, C. C. Bartholomae, M. Schmidt, C. von Kalle, V. Ast, C. Lawerenz, J. Eils, R. Kabbe, V. Benes, P. van Sluis, J. Koster, R. Volckmann, D. Shih, M. J. Betts, R. B. Russell, S. Coco, G. P. Tonini, U. Schuller, V. Hans, N. Graf, Y. J. Kim, C. Monoranu, W. Roggendorf, A. Unterberg, C. Herold-Mende, T. Milde, A. E. Kulozik, A. von Deimling, O. Witt, E. Maass, J. Rossler, M. Ebinger, M. U. Schuhmann, M. C. Fruhwald, M. Hasselblatt, N. Jabado, S. Rutkowski, A. O. von Bueren, D. Williamson, S. C. Clifford, M. G. McCabe, V. P. Collins, S. Wolf, S. Wiemann, H. Lehrach, B. Brors, W. Scheurlen, J. Felsberg, G. Reifenberger, P. A. Northcott, M. D. Taylor, M. Meyerson, S. L. Pomeroy, M. L. Yaspo, J. O. Korbel, A. Korshunov, R. Eils, S. M. Pfister, and P. Lichter, "Dissecting the genomic complexity underlying medulloblastoma," *Nature*, vol. 488, pp. 100–5, 2012. 3

[40] M. Le Gallo, A. J. O'Hara, M. L. Rudd, M. E. Urick, N. F. Hansen, N. J. O'Neil, J. C. Price, S. Zhang, B. M. England, A. K. Godwin, D. C. Sgroi, P. Hieter, J. C. Mullikin, M. J. Merino, and D. W. Bell, "Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes," *Nature genetics*, vol. 44, pp. 1310–5, 2012. 3

[41] L. Giulino-Roth, K. Wang, T. Y. MacDonald, S. Mathew, Y. Tam, M. T. Cronin, G. Palmer, N. Lucena-Silva, F. Pedrosa, M. Pedrosa, J. Teruya-Feldstein, G. Bhagat, B. Alobeid, L. Leoncini, C. Bellan, E. Rogena, K. A. Pinkney, M. A. Rubin, R. C. Ribeiro, R. Yelensky, W. Tam, P. J. Stephens, and E. Cesarman, "Targeted genomic sequencing of pediatric burkitt lymphoma identifies recurrent alterations in antiapoptotic and chromatin-remodeling genes," *Blood*, vol. 120, pp. 5181–4, 2012. 3

[42] G. L. Dalgliesh, K. Furge, C. Greenman, L. Chen, G. Bignell, A. Butler, H. Davies, S. Edkins, C. Hardy, C. Latimer, J. Teague, J. Andrews, S. Barthorpe, D. Beare, G. Buck, P. J. Campbell, S. Forbes, M. Jia, D. Jones, H. Knott, C. Y. Kok, K. W. Lau, C. Leroy, M. L. Lin, D. J. McBride, M. Maddison, S. Maguire, K. McLay, A. Menzies, T. Mironenko, L. Mulderrig, L. Mudie, S. O'Meara, E. Pleasance, A. Rajasingham, R. Shepherd, R. Smith, L. Stebbings, P. Stephens, G. Tang, P. S. Tarpey, K. Turrell,

K. J. Dykema, S. K. Khoo, D. Petillo, B. Wondergem, J. Anema, R. J. Kahnoski, B. T. Teh, M. R. Stratton, and P. A. Futreal, “Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes,” *Nature*, vol. 463, pp. 360–3, 2010. 3, 47

[43] P. W. Lewis, M. M. Muller, M. S. Koletsky, F. Cordero, S. Lin, L. A. Banaszynski, B. A. Garcia, T. W. Muir, O. J. Becher, and C. D. Allis, “Inhibition of prc2 activity by a gain-of-function h3 mutation found in pediatric glioblastoma,” *Science*, vol. 340, pp. 857–61, 2013. 3

[44] K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond, “Crystal structure of the nucleosome core particle at 2.8a resolution,” *Nature*, vol. 389, pp. 251–260, 1997. 3

[45] P. J. Robinson, L. Fairall, V. A. Huynh, and D. Rhodes, “Em measurements define the dimensions of the ”30-nm” chromatin fiber: evidence for a compact, interdigitated structure,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 6506–11, 2006. 3

[46] S. John, P. J. Sabo, R. E. Thurman, M. H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos, “Chromatin accessibility pre-determines glucocorticoid receptor binding patterns,” *Nature genetics*, vol. 43, pp. 264–8, 2011. 3

[47] B. D. Strahl and C. D. Allis, “The language of covalent histone modifications,” *Nature*, vol. 403, pp. 41–5, 2000. 3

[48] T. Kouzarides, “Chromatin modifications and their function,” *Cell*, vol. 128, pp. 693–705, 2007. 3

[49] K. Ahmad and S. Henikoff, “The histone variant h3.3 marks active chromatin by replication-independent nucleosome assembly,” *Molecular cell*, vol. 9, pp. 1191–200, 2002. 3

[50] T. Gautier, D. W. Abbott, A. Molla, A. Verdel, J. Ausio, and S. Dimitrov, “Histone variant h2abbd confers lower stability to the nucleosome,” *EMBO reports*, vol. 5, pp. 715–20, 2004. 3

[51] B. Brower-Toland, D. A. Wacker, R. M. Fulbright, J. T. Lis, W. L. Kraus, and M. D. Wang, “Specific contributions of histone tails and their acetylation to the mechanical stability of nucleosomes,” *Journal of molecular biology*, vol. 346, pp. 135–46, 2005. 3

[52] M. Ptashne, “On the use of the word ’epigenetic’,” *Current biology : CB*, vol. 17, pp. R233–6, 2007. 4

[53] J. Ernst and M. Kellis, “Discovery and characterization of chromatin states for systematic annotation of the human genome,” *Nat Biotechnol*, vol. 28, pp. 817–25, 2010. 4, 34

[54] L. Song, Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B.-K. Lee, N. C. Sheffield, S. Grf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey, "Open chromatin defined by dnasei and faire identifies regulatory elements that shape cell-type identity," *Genome Research*, vol. 21, pp. 1757–1767, 2011. 4, 48, 82, 85

[55] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren, "Histone modifications at human enhancers reflect global cell-type-specific gene expression," *Nature*, vol. 459, pp. 108–12, 2009. 4

[56] A. A. Hakimi, Y. B. Chen, J. Wren, M. Gonen, O. Abdel-Wahab, A. Heguy, H. Liu, S. Takeda, S. K. Tickoo, V. E. Reuter, M. H. Voss, R. J. Motzer, J. A. Coleman, E. H. Cheng, P. Russo, and J. J. Hsieh, "Clinical and pathologic impact of select chromatin-modulating tumor suppressors in clear cell renal cell carcinoma," *European urology*, vol. 63, pp. 848–854, 2012. 4, 47

[57] M. Patel, J. M. Simon, M. D. Iglesia, S. B. Wu, A. W. McFadden, J. D. Lieb, and I. J. Davis, "Tumor-specific retargeting of an oncogenic transcription factor chimera results in dysregulation of chromatin and transcription," *Genome Research*, vol. 22, pp. 259–70, 2012. 4, 5, 82, 83, 89

[58] L. Holmfeldt, L. Wei, E. Diaz-Flores, M. Walsh, J. Zhang, L. Ding, D. Payne-Turner, M. Churchman, A. Andersson, S. C. Chen, K. McCastlain, J. Becksfort, J. Ma, G. Wu, S. N. Patel, S. L. Heatley, L. A. Phillips, G. Song, J. Easton, M. Parker, X. Chen, M. Rusch, K. Boggs, B. Vadodaria, E. Hedlund, C. Drenberg, S. Baker, D. Pei, C. Cheng, R. Huether, C. Lu, R. S. Fulton, L. L. Fulton, Y. Tabib, D. J. Dooling, K. Ochoa, M. Minden, I. D. Lewis, L. B. To, P. Marlton, A. W. Roberts, G. Raca, W. Stock, G. Neale, H. G. Drexler, R. A. Dickins, D. W. Ellison, S. A. Shurtleff, C. H. Pui, R. C. Ribeiro, M. Devidas, A. J. Carroll, N. A. Heerema, B. Wood, M. J. Borowitz, J. M. Gastier-Foster, S. C. Raimondi, E. R. Mardis, R. K. Wilson, J. R. Downing, S. P. Hunger, M. L. Loh, and C. G. Mullighan, "The genomic landscape of hypodiploid acute lymphoblastic leukemia," *Nature genetics*, vol. 45, pp. 242–52, 2013. 4

[59] R. Esgueva, S. Perner, J. L. C, V. Scheble, C. Stephan, M. Lein, F. R. Fritzsche, M. Dietel, G. Kristiansen, and M. A. Rubin, "Prevalence of tmprss2-erg and slc45a3-erg gene fusions in a large prostatectomy cohort," *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, vol. 23, pp. 539–46, 2010. 4

[60] K. M. Bernt and S. A. Armstrong, "Targeting epigenetic programs in mll-rearranged leukemias," *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, vol. 2011, pp. 354–60, 2011. 4

[61] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal, "Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer," *Nucleic acids research*, vol. 39, no. Database issue, pp. D945–50, 2011. 4

[62] W. Timp and A. P. Feinberg, "Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host," *Nature reviews. Cancer*, vol. 13, pp. 497–510, 2013. 5

[63] E. Pujadas and A. P. Feinberg, "Regulated noise in the epigenetic landscape of development and disease," *Cell*, vol. 148, pp. 1123–31, 2012. 5

[64] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins, "Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss)," *Genome Research*, vol. 16, pp. 123–131, 2006. 6

[65] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-resolution mapping andcharacterization of open chromatin across the genome," *Cell*, vol. 132, pp. 311–322, 2008. 6

[66] L. Song and G. E. Crawford, "Dnase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," *Cold Spring Harb Protoc*, vol. 2010, pp. pdb.prot5384–, 2010. 6

[67] M. A. Keene, V. Corces, K. Lowenhaupt, and S. C. Elgin, "Dnase i hypersensitive sites in drosophila chromatin occur at the 5' ends of regions of transcription," *Proceedings of the National Academy of Sciences*, vol. 78, pp. 143–146, 1981. 6

[68] J. D. McGhee, W. I. Wood, M. Dolan, J. D. Engel, and G. Felsenfeld, "A 200 base pair region at the 5' end of the chicken adult [beta]-globin gene is accessible to nuclease digestion," *Cell*, vol. 27, no. 1, Part 2, pp. 45–55, 1981. 6

[69] G. Felsenfeld and M. Groudine, "Controlling the double helix," *Nature*, vol. 421, pp. 448–453, 2003. 6

[70] D. S. Gross and W. T. Garrard, "Nuclease hypersensitive sites in chromatin," *Annual Review of Biochemistry*, vol. 57, pp. 159–197, 1988. 6

[71] J. Stalder, A. Larsen, J. D. Engel, M. Dolan, M. Groudine, and H. Weintraub, "Tissue-specific dna cleavages in the globin chromatin domain introduced by dnaase i," *Cell*, vol. 20, pp. 451–460, 1980. 6

[72] G. J. Hogan, C.-K. Lee, and J. D. Lieb, "Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters," *PLoS Genet*, vol. 2, p. e158, 2006. 6, 48, 82

[73] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, “Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin,” *Genome Research*, vol. 17, pp. 877–885, 2007. 6, 44, 48, 82

[74] J. M. Simon, P. G. Giresi, I. J. Davis, and J. D. Lieb, “Using formaldehyde-assisted isolation of regulatory elements (faire) to isolate active regulatory dna,” *Nature protocols*, vol. 7, pp. 256–67, 2012. 6, 48, 76, 83, 89

[75] J. M. Simon, P. G. Giresi, I. J. Davis, and J. D. Lieb, “A detailed protocol for formaldehyde-assisted isolation of regulatory elements (faire),” *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, vol. Chapter 21, p. Unit21 26, 2013. 6, 76, 83, 89

[76] N. Ponts, E. Y. Harris, J. Prudhomme, I. Wick, C. Eckhardt-Ludka, G. R. Hicks, G. Hardiman, S. Lonardi, and K. G. Le Roch, “Nucleosome landscape and control of transcription in the human malaria parasite,” *Genome Research*, vol. 20, pp. 228–238, 2010. 6, 82

[77] M. Louwers, R. Bader, M. Haring, R. van Driel, W. de Laat, and M. Stam, “Tissue- and expression level-specific chromatin looping at maize b1 epialleles,” *The Plant cell*, vol. 21, pp. 832–42, 2009. 6, 82

[78] I. B. Hilton, J. M. Simon, J. D. Lieb, I. J. Davis, B. Damania, and D. P. Dittmer, “The open chromatin landscape of kaposi’s sarcoma-associated herpesvirus,” *Journal of virology*, vol. xx, p. xx, 2013. 7, 82

[79] J. Taunton, C. A. Hassig, and S. L. Schreiber, “A mammalian histone deacetylase related to the yeast transcriptional regulator rpd3p,” *Science*, vol. 272, pp. 408–11, 1996. 7

[80] R. Furumai, Y. Komatsu, N. Nishino, S. Khochbin, M. Yoshida, and S. Horinouchi, “Potent histone deacetylase inhibitors built from trichostatin a and cyclic tetrapeptide antibiotics including trapoxin,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 87–92, 2001. 7

[81] A. Cihak, “Biological effects of 5-azacytidine in eukaryotes,” *Oncology*, vol. 30, pp. 405–22, 1974. 7

[82] S. R. Daigle, E. J. Olhava, C. A. Therkelsen, A. Basavapathruni, L. Jin, P. A. Boriack-Sjodin, C. J. Allain, C. R. Klaus, A. Raimondi, M. P. Scott, N. J. Waters, R. Chesworth, M. P. Moyer, R. A. Copeland, V. M. Richon, and R. M. Pollock, “Potent inhibition of dot1l as treatment of mll-fusion leukemia,” *Blood*, vol. 122, pp. 1017–25, 2013. 7

[83] S. K. Knutson, N. M. Warholic, T. J. Wigle, C. R. Klaus, C. J. Allain, A. Raimondi, M. Porter Scott, R. Chesworth, M. P. Moyer, R. A. Copeland, V. M. Richon, R. M. Pollock, K. W. Kuntz, and H. Keilhack, “Durable tumor regression in genetically altered malignant rhabdoid tumors by inhibition of methyltransferase ezh2,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 7922–7, 2013. 7

[84] P. Filippakopoulos, J. Qi, S. Picaud, Y. Shen, W. B. Smith, O. Fedorov, E. M. Morse, T. Keates, T. T. Hickman, I. Felletar, M. Philpott, S. Munro, M. R. McKeown, Y. Wang, A. L. Christie, N. West, M. J. Cameron, B. Schwartz, T. D. Heightman, N. La Thangue, C. A. French, O. Wiest, A. L. Kung, S. Knapp, and J. E. Bradner, "Selective inhibition of bet bromodomains," *Nature*, vol. 468, pp. 1067–73, 2010. 7

[85] A. Bernheim, "Cytogenomics of cancers: from chromosome to sequence," *Mol Oncol*, vol. 4, pp. 309–22, 2010. 9

[86] O. Delattre, J. Zucman, B. Plougastel, C. Desmaze, T. Melot, M. Peter, H. Kovar, I. Joubert, P. de Jong, G. Rouleau, and et al., "Gene fusion with an ets dna-binding domain caused by chromosome translocation in human tumours," *Nature*, vol. 359, pp. 162–5, 1992. 9

[87] P. H. Sorensen, S. L. Lessnick, D. Lopez-Terrada, X. F. Liu, T. J. Triche, and C. T. Denny, "A second ewing's sarcoma translocation, t(21;22), fuses the ews gene to another ets-family transcription factor, erg," *Nat Genet*, vol. 6, pp. 146–51, 1994. 9, 27

[88] I. S. Jeon, J. N. Davis, B. S. Braun, J. E. Sublett, M. F. Roussel, C. T. Denny, and D. N. Shapiro, "A variant ewing's sarcoma translocation (7;22) fuses the ews gene to the ets gene etv1," *Oncogene*, vol. 10, pp. 1229–34, 1995. 9

[89] M. Kinsey, R. Smith, and S. L. Lessnick, "Nr0b1 is required for the oncogenic phenotype mediated by ews/fli in ewing's sarcoma," *Mol Cancer Res*, vol. 4, pp. 851–9, 2006. 9, 16

[90] R. Smith, L. A. Owen, D. J. Trem, J. S. Wong, J. S. Whangbo, T. R. Golub, and S. L. Lessnick, "Expression profiling of ews/fli identifies nkx2.2 as a critical target gene in ewing's sarcoma," *Cancer Cell*, vol. 9, pp. 405–16, 2006. 9, 13, 19

[91] S. L. Lessnick, B. S. Braun, C. T. Denny, and W. A. May, "Multiple domains mediate transformation by the ewing's sarcoma ews/fli-1 fusion gene," *Oncogene*, vol. 10, pp. 423–31, 1995. 9

[92] S. Jaishankar, J. Zhang, M. F. Roussel, and S. J. Baker, "Transforming activity of ews/fli is not strictly dependent upon dna-binding activity," *Oncogene*, vol. 18, pp. 5592–7, 1999. 9

[93] S. L. Lessnick, C. S. Dacwag, and T. R. Golub, "The ewing's sarcoma oncoprotein ews/fli induces a p53-dependent growth arrest in primary human fibroblasts," *Cancer Cell*, vol. 1, pp. 393–401, 2002. 10

[94] Y. Miyagawa, H. Okita, H. Nakaijima, Y. Horiuchi, B. Sato, T. Taguchi, M. Toyoda, Y. U. Katagiri, J. Fujimoto, J. Hata, A. Umezawa, and N. Kiyokawa, "Inducible expression of chimeric ews/ets proteins confers ewing's family tumor-like phenotypes to human mesenchymal progenitor cells," *Mol Cell Biol*, vol. 28, pp. 2125–37, 2008. 10

[95] N. Riggi, M. L. Suva, D. Suva, L. Cironi, P. Provero, S. Tercier, J. M. Joseph, J. C. Stehle, K. Baumer, V. Kindler, and I. Stamenkovic, "Ews-fli-1 expression triggers a ewing's sarcoma initiation program in primary human mesenchymal stem cells," *Cancer Res*, vol. 68, pp. 2176–85, 2008. 10, 19

[96] V. N. Rao, T. Ohno, D. D. Prasad, G. Bhattacharya, and E. S. Reddy, "Analysis of the dna-binding and transcriptional activation functions of human fli-1 protein," *Oncogene*, vol. 8, pp. 2167–73, 1993. 10

[97] F. O. Bartel, T. Higuchi, and D. D. Spyropoulos, "Mouse models in the study of the ets family of transcription factors," *Oncogene*, vol. 19, pp. 6443–54, 2000. 10

[98] D. D. Spyropoulos, P. N. Pharr, K. R. Lavenburg, P. Jackers, T. S. Papas, M. Ogawa, and D. K. Watson, "Hemorrhage, impaired hematopoiesis, and lethality in mouse embryos carrying a targeted disruption of the fli1 transcription factor," *Mol Cell Biol*, vol. 20, pp. 5643–52, 2000. 10, 11

[99] F. Liu, M. Walmsley, A. Rodaway, and R. Patient, "Fli1 acts at the top of the transcriptional network driving blood and endothelial development," *Curr Biol*, vol. 18, pp. 1234–40, 2008. 10, 11

[100] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan, "Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer," *Science*, vol. 310, pp. 644–8, 2005. 10

[101] T. Ohno, M. Ouchida, L. Lee, Z. Gatalica, V. N. Rao, and E. S. Reddy, "The ews gene, involved in ewing family of tumors, malignant melanoma of soft parts and desmoplastic small round cell tumors, codes for an rna binding protein with novel regulatory domains," *Oncogene*, vol. 9, pp. 3087–97, 1994. 10

[102] H. Li, W. Watford, C. Li, A. Parmelee, M. A. Bryant, C. Deng, J. O'Shea, and S. B. Lee, "Ewing sarcoma gene ews is essential for meiosis and b lymphocyte development," *J Clin Invest*, vol. 117, pp. 1314–23, 2007. 10

[103] J. Zucman, O. Delattre, C. Desmaze, A. L. Epstein, G. Stenman, F. Speleman, C. D. Fletchers, A. Aurias, and G. Thomas, "Ews and atf-1 gene fusion induced by t(12;22) translocation in malignant melanoma of soft parts," *Nat Genet*, vol. 4, pp. 341–5, 1993. 10

[104] L. E. Benjamin, W. J. Fredericks, F. G. Barr, and r. Rauscher, F. J., "Fusion of the ews1 and wt1 genes as a result of the t(11;22)(p13;q12) translocation in desmoplastic small round cell tumors," *Med Pediatr Oncol*, vol. 27, pp. 434–9, 1996. 10

[105] M. P. Pusztaszeri, W. Seelentag, and F. T. Bosman, "Immunohistochemical expression of endothelial markers cd31, cd34, von willebrand factor, and fli-1 in normal human tissues," *J Histochem Cytochem*, vol. 54, pp. 385–95, 2006. 11

[106] T. A. Egelhofer, A. Minoda, S. Klugman, K. Lee, P. Kolasinska-Zwierz, A. A. Alekseyenko, M. S. Cheung, D. S. Day, S. Gadel, A. A. Gorchakov, T. Gu, P. V. Kharchenko, S. Kuan, I. Latorre, D. Linder-Basso, Y. Luu, Q. Ngo, M. Perry, A. Rechtsteiner, N. C. Riddle, Y. B. Schwartz, G. A. Shanower, A. Vielle, J. Ahringer, S. C. Elgin, M. I. Kuroda, V. Pirrotta, B. Ren, S. Strome, P. J. Park, G. H. Karpen, R. D. Hawkins, and J. D. Lieb, "An assessment of histone-modification antibody quality," *Nat Struct Mol Biol*, vol. 18, pp. 91–93, 2010. 11

[107] W. A. May, S. L. Lessnick, B. S. Braun, M. Klemsz, B. C. Lewis, L. B. Lunsford, R. Hromas, and C. T. Denny, "The ewing's sarcoma ews/fli-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than fli-1," *Mol Cell Biol*, vol. 13, pp. 7393–8, 1993. 13

[108] W. A. May, M. L. Gishizky, S. L. Lessnick, L. B. Lunsford, B. C. Lewis, O. Delattre, J. Zucman, G. Thomas, and C. T. Denny, "Ewing sarcoma 11;22 translocation produces a chimeric transcription factor that requires the dna-binding domain encoded by fli1 for transformation," *Proc Natl Acad Sci U S A*, vol. 90, pp. 5752–6, 1993. 13

[109] K. Tanaka, T. Iwakuma, K. Harimaya, H. Sato, and Y. Iwamoto, "Ews-fli1 antisense oligodeoxynucleotide inhibits proliferation of human ewing's sarcoma and primitive neuroectodermal tumor cells," *J Clin Invest*, vol. 99, pp. 239–47, 1997. 13

[110] N. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb, "Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions," *Genome Biol*, vol. 12, p. R67, 2011. 13, 45, 78

[111] K. Gangwal, S. Sankar, P. C. Hollenhorst, M. Kinsey, S. C. Haroldsen, A. A. Shah, K. M. Boucher, W. S. Watkins, L. B. Jorde, B. J. Graves, and S. L. Lessnick, "Microsatellites as ews/fli response elements in ewing's sarcoma," *Proc Natl Acad Sci U S A*, vol. 105, pp. 10149–54, 2008. 13, 16, 24, 27, 30, 42

[112] N. Guillon, F. Tirode, V. Boeva, A. Zynovyev, E. Barillot, and O. Delattre, "The oncogenic ews-fli1 protein binds in vivo ggaa microsatellite sequences with potential transcriptional activation function," *PLoS One*, vol. 4, p. e4932, 2009. 13, 24, 30

[113] G. H. Wei, G. Badis, M. F. Berger, T. Kivioja, K. Palin, M. Enge, M. Bonke, A. Jolma, M. Varjosalo, A. R. Gehrke, J. Yan, S. Talukder, M. Turunen, M. Taipale, H. G. Stunnenberg, E. Ukkonen, T. R. Hughes, M. L. Bulyk, and J. Taipale, "Genome-wide analysis of ets-family dna-binding in vitro and in vivo," *EMBO J*, vol. 29, pp. 2147–60, 2010. 13, 24, 27, 31

[114] K. Gangwal and S. L. Lessnick, "Microsatellites are ews/fli response elements: genomic "junk" is ews/fli's treasure," *Cell Cycle*, vol. 7, pp. 3127–32, 2008. 16, 27

[115] E. Garcia-Aragoncillo, J. Carrillo, E. Lalli, N. Agra, G. Gomez-Lopez, A. Pestana, and J. Alonso, "Dax1, a direct target of ews/fli1 oncoprotein, is a principal regulator of cell-cycle progression in ewing's tumor cells," *Oncogene*, vol. 27, pp. 6034–43, 2008. 16, 43

[116] N. Riggi, L. Cironi, P. Provero, M. L. Suva, K. Kaloulis, C. Garcia-Echeverria, F. Hoffmann, A. Trumpp, and I. Stamenkovic, "Development of ewing's sarcoma from primary bone marrow-derived mesenchymal progenitor cells," *Cancer Res*, vol. 65, pp. 11459–68, 2005. 19

[117] M. Kauer, J. Ban, R. Kofler, B. Walker, S. Davis, P. Meltzer, and H. Kovar, "A molecular function map of ewing's sarcoma," *PLoS One*, vol. 4, p. e5415, 2009. 19

[118] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano, "Great improves functional interpretation of cis-regulatory regions," *Nat Biotechnol*, vol. 28, pp. 495–501, 2010. 22, 50, 55, 77

[119] A. A. Sharov and M. S. Ko, "Exhaustive search for over-represented dna sequence motifs with cisfinder," *DNA Res*, vol. 16, pp. 261–73, 2009. 24, 46

[120] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "Transfac: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, pp. 374–8, 2003. 24

[121] D. C. Shing, D. J. McMullan, P. Roberts, K. Smith, S. F. Chin, J. Nicholson, R. M. Tillman, P. Ramani, C. Cullinane, and N. Coleman, "Fus/erg gene fusions in ewing's tumors," *Cancer Res*, vol. 63, pp. 4568–76, 2003. 27

[122] T. E. P. Consortium, "The encode (encyclopedia of dna elements) project," *Science*, vol. 306, pp. 636–40, 2004. 30, 33

[123] P. C. Hollenhorst, K. J. Chandler, R. L. Poulsen, W. E. Johnson, N. A. Speck, and B. J. Graves, "Dna specificity determinants associate with distinct transcription factor functions," *PLoS Genet*, vol. 5, p. e1000778, 2009. 31, 42

[124] X. Xie, P. Rigor, and P. Baldi, "Motifmap: a human genome-wide map of candidate regulatory motif sites," *Bioinformatics*, vol. 25, pp. 167–74, 2009. 32

[125] C. Cillo, A. Faiella, M. Cantile, and E. Boncinelli, "Homeobox genes and cancer," *Exp Cell Res*, vol. 248, pp. 1–9, 1999. 32

[126] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander, "A bivalent chromatin structure marks key developmental genes in embryonic stem cells," *Cell*, vol. 125, pp. 315–26, 2006. 33

[127] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander, "Genome-scale dna methylation maps of pluripotent and differentiated cells," *Nature*, vol. 454, pp. 766–70, 2008. 34

[128] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, “The dna-encoded nucleosome organization of a eukaryotic genome,” *Nature*, vol. 458, pp. 362–6, 2009. 36, 40

[129] K. Gangwal, D. Close, C. A. Enriquez, C. P. Hill, and S. L. Lessnick, “Emergent properties of ews/fli regulation via ggaa microsatellites in ewing’s sarcoma,” *Genes Cancer*, vol. 1, pp. 177–187, 2010. 42

[130] A. G. Bassuk and J. M. Leiden, “A direct physical association between ets and ap-1 transcription factors in normal human t cells,” *Immunity*, vol. 3, pp. 223–37, 1995. 42

[131] S. Rao, A. Matsumura, J. Yoon, and M. C. Simon, “Spi-b activates transcription via a unique proline, serine, and threonine domain and exhibits dna binding affinity differences from pu.1,” *J Biol Chem*, vol. 274, pp. 11115–24, 1999. 42

[132] A. Verger, E. Buisine, S. Carrere, R. Wintjens, A. Flourens, J. Coll, D. Stehelin, and M. Duterque-Coquillaud, “Identification of amino acid residues in the ets transcription factor erg that mediate erg-jun/fos-dna ternary complex formation,” *J Biol Chem*, vol. 276, pp. 17181–9, 2001. 42

[133] S. Kim, C. T. Denny, and R. Wisdom, “Cooperative dna binding with ap-1 proteins is required for transformation by ews-ets fusion proteins,” *Mol Cell Biol*, vol. 26, pp. 2467–78, 2006. 42

[134] M. J. Stankiewicz and J. D. Crispino, “Ets2 and erg promote megakaryopoiesis and synergize with alterations in gata-1 to immortalize hematopoietic progenitor cells,” *Blood*, vol. 113, pp. 3337–47, 2009. 42

[135] C. W. Garvie, J. Hagman, and C. Wolberger, “Structural studies of ets-1/pax5 complex formation on dna,” *Mol Cell*, vol. 8, pp. 1267–76, 2001. 42

[136] D. T. Ting, D. Lipson, S. Paul, B. W. Brannigan, S. Akhavanfard, E. J. Coffman, G. Contino, V. Deshpande, A. J. Iafrate, S. Letovsky, M. N. Rivera, N. Bardeesy, S. Maheswaran, and D. A. Haber, “Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers,” *Science*, vol. 331, pp. 593–596, 2011. 42

[137] L. A. Cirillo, C. E. McPherson, P. Bossard, K. Stevens, S. Cherian, E. Y. Shim, K. L. Clark, S. K. Burley, and K. S. Zaret, “Binding of the winged-helix transcription factor hnf3 to a linker histone site on the nucleosome,” *EMBO J*, vol. 17, pp. 244–54, 1998. 42

[138] L. A. Cirillo, F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, and K. S. Zaret, “Opening of compacted chromatin by early developmental transcription factors hnf3 (foxa) and gata-4,” *Mol Cell*, vol. 9, pp. 279–89, 2002. 42

[139] J. S. Carroll, X. S. Liu, A. S. Brodsky, W. Li, C. A. Meyer, A. J. Szary, J. Eeckhoute, W. Shao, E. V. Hestermann, T. R. Geistlinger, E. A. Fox, P. A. Silver, and M. Brown,

"Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein foxa1," *Cell*, vol. 122, pp. 33–43, 2005. 42

[140] J. S. Carroll, C. A. Meyer, J. Song, W. Li, T. R. Geistlinger, J. Eeckhoute, A. S. Brodsky, E. K. Keeton, K. C. Fertuck, G. F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E. A. Fox, P. A. Silver, T. R. Gingeras, X. S. Liu, and M. Brown, "Genome-wide analysis of estrogen receptor binding sites," *Nat Genet*, vol. 38, pp. 1289–1297, 2006. 42

[141] H. H. He, C. A. Meyer, H. Shin, S. T. Bailey, G. Wei, Q. Wang, Y. Zhang, K. Xu, M. Ni, M. Lupien, P. Mieczkowski, J. D. Lieb, K. Zhao, M. Brown, and X. S. Liu, "Nucleosome dynamics define transcriptional enhancers," *Nat Genet*, vol. 42, pp. 343–7, 2010. 42

[142] A. Hurtado, K. A. Holmes, C. S. Ross-Innes, D. Schmidt, and J. S. Carroll, "Foxa1 is a key determinant of estrogen receptor function and endocrine response," *Nat Genet*, vol. 43, pp. 27–33, 2011. 42

[143] F. Cattaneo and G. Nucifora, "Evi1 recruits the histone methyltransferase suv39h1 for transcription repression," *J Cell Biochem*, vol. 105, pp. 344–52, 2008. 42

[144] S. Frietze, H. O'Geen, K. R. Blahnik, V. X. Jin, and P. J. Farnham, "Znf274 recruits the histone methyltransferase setdb1 to the 3' ends of znf genes," *PLoS One*, vol. 5, p. e15082, 2010. 42

[145] N. Rezai-Zadeh, X. Zhang, F. Namour, G. Fejer, Y. D. Wen, Y. L. Yao, I. Gyory, K. Wright, and E. Seto, "Targeted recruitment of a histone h4-specific methyltransferase by the transcription factor yy1," *Genes Dev*, vol. 17, pp. 1019–29, 2003. 42

[146] L. Yang, L. Xia, D. Y. Wu, H. Wang, H. A. Chansky, W. H. Schubach, D. D. Hickstein, and Y. Zhang, "Molecular cloning of eset, a novel histone h3-specific methyltransferase that interacts with erg transcription factor," *Oncogene*, vol. 21, pp. 148–52, 2002. 43

[147] L. Yang, Q. Mei, A. Zielinska-Kwiatkowska, Y. Matsui, M. L. Blackburn, D. Benedetti, A. A. Krumm, J. Taborsky, G. J., and H. A. Chansky, "An erg (ets-related gene)-associated histone methyltransferase interacts with histone deacetylases 1/2 and transcription co-repressors msin3a/b," *Biochem J*, vol. 369, no. Pt 3, pp. 651–7, 2003. 43

[148] K. Scotlandi, S. Benini, M. Sarti, M. Serra, P. L. Lollini, D. Maurici, P. Picci, M. C. Manara, and N. Baldini, "Insulin-like growth factor i receptor-mediated circuit in ewing's sarcoma/peripheral neuroectodermal tumor: a possible therapeutic target," *Cancer Res*, vol. 56, pp. 4570–4, 1996. 43

[149] D. Herrero-Martin, D. Osuna, J. L. Ordonez, V. Sevillano, A. S. Martins, C. Mackintosh, M. Campos, J. Madoz-Gurpide, A. P. Otero-Motta, G. Caballero, A. T. Amaral, D. H. Wai, Y. Braun, M. Eisenacher, K. L. Schaefer, C. Poremba, and E. de Alava, "Stable interference of ews-fli1 in an ewing sarcoma cell line impairs igf-1/igf-1r signalling and reveals topk as a new target," *Br J Cancer*, vol. 101, pp. 80–90, 2009. 43

[150] D. A. Rubinson, C. P. Dillon, A. V. Kwiatkowski, C. Sievers, L. Yang, J. Kopinja, D. L. Rooney, M. Zhang, M. M. Ihrig, M. T. McManus, F. B. Gertler, M. L. Scott, and L. Van Parijs, "A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by rna interference," *Nat Genet*, vol. 33, pp. 401–6, 2003. 44

[151] I. J. Davis, J. J. Kim, F. Ozsolak, H. R. Widlund, O. Rozenblatt-Rosen, S. R. Granter, J. Du, J. A. Fletcher, C. T. Denny, S. L. Lessnick, W. M. Linehan, A. L. Kung, and D. E. Fisher, "Oncogenic mitf dysregulation in clear cell sarcoma: defining the mit family of human cancers," *Cancer Cell*, vol. 9, pp. 473–84, 2006. 44

[152] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome Biol*, vol. 10, p. R25, 2009. 44, 89

[153] H. Shin, T. Liu, A. K. Manrai, and X. S. Liu, "Ceas: cis-regulatory element annotation system," *Bioinformatics*, vol. 25, pp. 2605–6, 2009. 45, 78

[154] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–2, 2010. 45, 76, 89

[155] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biol*, vol. 11, no. 8, p. R86, 2010. 45, 78

[156] A. J. Saldanha, "Java treeview–extensible visualization of microarray data," *Bioinformatics*, vol. 20, pp. 3246–3248, 2004. 46, 76

[157] S. Pena-Llopis, S. Vega-Rubin-de Celis, A. Liao, N. Leng, A. Pavia-Jimenez, S. Wang, T. Yamasaki, L. Zhrebker, S. Sivanand, P. Spence, L. Kinch, T. Hambuch, S. Jain, Y. Lotan, V. Margulis, A. I. Sagalowsky, P. B. Summerour, W. Kabbani, S. W. Wong, N. Grishin, M. Laurent, X. J. Xie, C. D. Haudenschild, M. T. Ross, D. R. Bentley, P. Kapur, and J. Brugarolas, "Bap1 loss defines a new class of renal cell carcinoma," *Nature genetics*, vol. 44, pp. 751–9, 2012. 47

[158] R. J. Ryan and B. E. Bernstein, "Molecular biology. genetic events that shape the cancer epigenome," *Science*, vol. 336, pp. 1513–4, 2012. 47

[159] P. Kapur, S. Pena-Llopis, A. Christie, L. Zhrebker, A. Pavia-Jimenez, W. K. Rathmell, X. J. Xie, and J. Brugarolas, "Effects on survival of bap1 and pbrm1 mutations in sporadic clear-cell renal-cell carcinoma: a retrospective analysis with independent validation," *The lancet oncology*, vol. 14, pp. 159–67, 2013. 47

[160] W. Y. Kim and W. G. Kaelin, "Role of vhl gene mutation in human cancer," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 22, pp. 4991–5004, 2004. 47

[161] G. Bratslavsky, S. Sudarshan, L. Neckers, and W. M. Linehan, “Pseudohypoxic pathways in renal cell carcinoma,” *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 13, pp. 4667–71, 2007. 47

[162] M. L. Nickerson, E. Jaeger, Y. Shi, J. A. Durocher, S. Mahurkar, D. Zaridze, V. Matveev, V. Janout, H. Kollarova, V. Bencko, M. Navratilova, N. Szeszenia-Dabrowska, D. Mates, A. Mukeria, I. Holcatova, L. S. Schmidt, J. R. Toro, S. Karami, R. Hung, G. F. Gerard, W. M. Linehan, M. Merino, B. Zbar, P. Boffetta, P. Brennan, N. Rothman, W.-H. Chow, F. M. Waldman, and L. E. Moore, “Improved identification of von hippel-lindau gene alterations in clear cell renal tumors,” *Clinical Cancer Research*, vol. 14, pp. 4726–4734, 2008. 47

[163] E. Jonasch, P. A. Futreal, I. J. Davis, S. T. Bailey, W. Y. Kim, J. Brugarolas, A. J. Giaccia, G. Kurban, A. Pause, J. Frydman, A. J. Zurita, B. I. Rini, P. Sharma, M. B. Atkins, C. L. Walker, and W. K. Rathmell, “State of the science: an update on renal cell carcinoma,” *Molecular cancer research : MCR*, vol. 10, pp. 859–80, 2012. 47, 48

[164] J. D. Gordan, P. Lal, V. R. Dondeti, R. Letrero, K. N. Parekh, C. E. Oquendo, R. A. Greenberg, K. T. Flaherty, W. K. Rathmell, B. Keith, M. C. Simon, and K. L. Nathanson, “Hif-alpha effects on c-myc distinguish two subtypes of sporadic vhl-deficient clear cell renal carcinoma,” *Cancer Cell*, vol. 14, pp. 435–46, 2008. 48

[165] M. E. Gore and J. M. Larkin, “Challenges and opportunities for converting renal cell carcinoma into a chronic disease with targeted therapies,” *British journal of cancer*, vol. 104, pp. 399–406, 2011. 48

[166] P. L. Nagy, M. L. Cleary, P. O. Brown, and J. D. Lieb, “Genomewide demarcation of rna polymerase ii transcription units revealed by physical fractionation of chromatin,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 6364–6369, 2003. 48, 82

[167] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos, “The accessible chromatin landscape of the human genome,” *Nature*, vol. 489, pp. 75–82, 2012. 48

[168] G. V. Rayasam, O. Wendling, P. O. Angrand, M. Mark, K. Niederreither, L. Song, T. Lerouge, G. L. Hager, P. Chambon, and R. Losson, “Nsd1 is essential for early post-implantation development and has a catalytically active set domain,” *The EMBO journal*, vol. 22, pp. 3153–63, 2003. 48

[169] X. J. Sun, J. Wei, X. Y. Wu, M. Hu, L. Wang, H. H. Wang, Q. H. Zhang, S. J. Chen, Q. H. Huang, and Z. Chen, “Identification and characterization of a novel human histone h3 lysine 36-specific methyltransferase,” *The Journal of biological chemistry*, vol. 280, pp. 35261–71, 2005. 48

[170] M. A. Brown, r. Sims, R. J., P. D. Gottlieb, and P. W. Tucker, “Identification and characterization of smyd2: a split set/mynd domain-containing histone h3 lysine 36-specific methyltransferase that interacts with the sin3 histone deacetylase complex,” *Molecular cancer*, vol. 5, p. 26, 2006. 48

[171] J. W. Edmunds, L. C. Mahadevan, and A. L. Clayton, “Dynamic histone h3 methylation during gene induction: Hypb/setd2 mediates all h3k36 trimethylation,” *The EMBO journal*, vol. 27, pp. 406–20, 2008. 48, 67, 74

[172] S. M. Yoh, J. S. Lucas, and K. A. Jones, “The iws1:spt6:ctd complex controls cotranscriptional mrna biosynthesis and hypb/setd2-mediated histone h3k36 methylation,” *Genes & development*, vol. 22, pp. 3422–34, 2008. 48

[173] P. Kolasinska-Zwierz, T. Down, I. Latorre, T. Liu, X. S. Liu, and J. Ahringer, “Differential chromatin marking of introns and expressed exons by h3k36me3,” *Nature genetics*, vol. 41, pp. 376–81, 2009. 48, 52, 67

[174] S. Schwartz, E. Meshorer, and G. Ast, “Chromatin organization marks exon-intron structure,” *Nature structural & molecular biology*, vol. 16, pp. 990–5, 2009. 48, 67

[175] R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, and T. Misteli, “Regulation of alternative splicing by histone modifications,” *Science*, vol. 327, pp. 996–1000, 2010. 48, 60

[176] M. M. Pradeepa, H. G. Sutherland, J. Ule, G. R. Grimes, and W. A. Bickmore, “Psip1/ledgf p52 binds methylated histone h3k36 and splicing factors and contributes to the regulation of alternative splicing,” *PLoS genetics*, vol. 8, p. e1002717, 2012. 48, 60

[177] S. Carvalho, A. C. Raposo, F. B. Martins, A. R. Grosso, S. C. Sridhara, J. Rino, M. Carmo-Fonseca, and S. F. de Almeida, “Histone methyltransferase setd2 coordinates fact recruitment with nucleosome dynamics during transcription,” *Nucleic acids research*, vol. 41, pp. 2881–2893, 2013. 48, 69

[178] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, and C. Swanton, “Intratumor heterogeneity and branched evolution revealed by multiregion sequencing,” *The New England journal of medicine*, vol. 366, pp. 883–92, 2012. 49, 57

[179] A. M. Fontebasso, J. Schwartzentruber, D. A. Khuong-Quang, X. Y. Liu, D. Sturm, A. Korshunov, D. T. Jones, H. Witt, M. Kool, S. Albrecht, A. Fleming, D. Hadjadj, S. Busche, P. Lepage, A. Montpetit, A. Staffa, N. Gerges, M. Zakrzewska, K. Zakrzewski, P. P. Liberski, P. Hauser, M. Garami, A. Klekner, L. Bognar, G. Zadeh, D. Faury, S. M. Pfister, N. Jabado, and J. Majewski, "Mutations in setd2 and genes affecting histone h3k36 methylation target hemispheric high-grade gliomas," *Acta neuropathologica*, vol. 125, pp. 659–669, 2013. 49

[180] J. Schodel, S. Oikonomopoulos, J. Ragoussis, C. W. Pugh, P. J. Ratcliffe, and D. R. Mole, "High-resolution genome-wide mapping of hif-binding sites by chip-seq," *Blood*, vol. 117, pp. e207–17, 2011. 50, 54, 76

[181] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities," *Molecular cell*, vol. 38, pp. 576–89, 2010. 50, 77

[182] A. Dhayalan, A. Rajavelu, P. Rathert, R. Tamas, R. Z. Jurkowska, S. Ragozin, and A. Jeltsch, "The dnmt3a pwwp domain reads histone 3 lysine 36 trimethylation and guides dna methylation," *The Journal of biological chemistry*, vol. 285, pp. 26114–20, 2010. 60

[183] S. F. de Almeida, A. R. Grosso, F. Koch, R. Fenouil, S. Carvalho, J. Andrade, H. Levezinho, M. Gut, D. Eick, I. Gut, J. C. Andrau, P. Ferrier, and M. Carmo-Fonseca, "Splicing enhances recruitment of methyltransferase hypb/setd2 and methylation of histone h3 lys36," *Nature structural & molecular biology*, vol. 18, pp. 977–83, 2011. 60

[184] D. Singh, C. F. Orellana, Y. Hu, C. D. Jones, Y. Liu, D. Y. Chiang, J. Liu, and J. F. Prins, "Fdm: a graph-based statistical method to detect differential transcription using rna-seq data," *Bioinformatics*, vol. 27, pp. 2633–40, 2011. 62, 80

[185] M. J. Carrozza, B. Li, L. Florens, T. Suganuma, S. K. Swanson, K. K. Lee, W. J. Shia, S. Anderson, J. Yates, M. P. Washburn, and J. L. Workman, "Histone h3 methylation by set2 directs deacetylation of coding regions by rpd3s to suppress spurious intragenic transcription," *Cell*, vol. 123, pp. 581–92, 2005. 69, 74

[186] C. R. Lickwar, B. Rao, A. A. Shabalin, A. B. Nobel, B. D. Strahl, and J. D. Lieb, "The set2/rpd3s pathway suppresses cryptic transcription without regard to gene length or transcription frequency," *PLoS One*, vol. 4, p. e4886, 2009. 69, 74

[187] S. Kadener, P. Cramer, G. Nogues, D. Cazalla, M. de la Mata, J. P. Fededa, S. E. Werbajh, A. Srebrow, and A. R. Kornblihtt, "Antagonistic effects of t-ag and vp16 reveal a role for rna pol ii elongation on alternative splicing," *The EMBO journal*, vol. 20, pp. 5759–68, 2001. 74, 94

[188] K. J. Howe, C. M. Kane, and J. Ares, M., "Perturbation of transcription elongation influences the fidelity of internal exon inclusion in saccharomyces cerevisiae," *RNA*, vol. 9, pp. 993–1006, 2003. 74, 94

[189] E. Batsche, M. Yaniv, and C. Muchardt, “The human swi/snf subunit brm is a regulator of alternative splicing,” *Nature structural & molecular biology*, vol. 13, pp. 22–9, 2006. 74, 94

[190] A. R. Kornblihtt, “Coupling transcription and alternative splicing,” *Advances in experimental medicine and biology*, vol. 623, pp. 175–89, 2007. 74, 94

[191] M. J. Munoz, M. S. Perez Santangelo, M. P. Paronetto, M. de la Mata, F. Pelisch, S. Boireau, K. Glover-Cutter, C. Ben-Dov, M. Blaustein, J. J. Lozano, G. Bird, D. Bentley, E. Bertrand, and A. R. Kornblihtt, “Dna damage regulates alternative splicing through inhibition of rna polymerase ii elongation,” *Cell*, vol. 137, pp. 708–20, 2009. 74, 94

[192] T. Lassmann, Y. Hayashizaki, and C. O. Daub, “Tagdust–a program to eliminate artifacts from next generation sequencing data,” *Bioinformatics*, vol. 25, pp. 2839–40, 2009. 76, 89

[193] H. Li and R. Durbin, “Fast and accurate short read alignment with burrows-wheeler transform,” *Bioinformatics*, vol. 25, pp. 1754–60, 2009. 76, 77

[194] J. Feng, T. Liu, B. Qin, Y. Zhang, and X. S. Liu, “Identifying chip-seq enrichment using macs,” *Nature protocols*, vol. 7, pp. 1728–40, 2012. 76

[195] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, “The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data,” *Genome Research*, vol. 20, pp. 1297–303, 2010. 77

[196] C. Trapnell, L. Pachter, and S. L. Salzberg, “Tophat: discovering splice junctions with rna-seq,” *Bioinformatics*, vol. 25, pp. 1105–11, 2009. 79

[197] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu, “Mapsplice: accurate mapping of rna-seq reads for splice junction discovery,” *Nucleic acids research*, vol. 38, p. e178, 2010. 80

[198] N. Blow, “Tissue preparation: Tissue issues,” *Nature*, vol. 448, no. 7156, pp. 959–63, 2007. Blow, Nathan England Nature. 2007 Aug 23;448(7156):959-63. 81

[199] R. C. Grafstrom, J. Fornace, A. J., H. Autrup, J. F. Lechner, and C. C. Harris, “Formaldehyde damage to dna and inhibition of dna repair in human bronchial cells,” *Science*, vol. 220, no. 4593, pp. 216–8, 1983. Grafstrom, R C Fornace, A J Jr Autrup, H Lechner, J F Harris, C C New York, N.Y. Science. 1983 Apr 8;220(4593):216-8. 81

[200] M. Fanelli, S. Amatori, I. Barozzi, M. Soncini, R. Dal Zuffo, G. Bucci, M. Capra, M. Quarto, G. I. Dellino, C. Mercurio, M. Alcalay, G. Viale, P. G. Pelicci, and S. Minucci, “Pathology tissue-chromatin immunoprecipitation, coupled with high-throughput

sequencing, allows the epigenetic profiling of patient samples," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 21535–40, 2010. 81

[201] M. Fanelli, S. Amatori, I. Barozzi, and S. Minucci, "Chromatin immunoprecipitation and high-throughput sequencing from paraffin-embedded pathology tissue," *Nature protocols*, vol. 6, pp. 1905–19, 2011. 81

[202] J. M. Simon, K. E. Hacker, D. Singh, A. R. Brannon, J. S. Parker, M. Weiser, T. H. Ho, P. F. Kuan, E. Jonasch, T. S. Furey, J. F. Prins, J. D. Lieb, W. K. Rathmell, and I. J. Davis, "Variation in chromatin accessibility in human kidney cancer links h3k36 methyltransferase loss with widespread rna processing defects," *Genome Research*, vol. xx, pp. yy–yy, 2013. Simon, Jeremy M Hacker, Kathryn E Singh, Darshan Brannon, A Rose Parker, Joel S Weiser, Matthew Ho, Thai H Kuan, Pei-Fen Jonasch, Eric Furey, Terrence S Prins, Jan F Lieb, Jason D Rathmell, W Kimryn Davis, Ian J Genome Res. 2013 Oct 24. 82

[203] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. Edwards, G. R. Bignell, M. R. Stratton, and P. A. Futreal, "Bowtie: An ultrafast memory-efficient short read aligner identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nat Genet*, vol. 40, pp. 722 – 729, 2008. 89