Comprehensive and Integrative Analysis of the KMT2D Regulome

by

James T. Topham

B.Sc., Simon Fraser University, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES (Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2017

© James T. Topham, 2017

Abstract

Lysine (K)-specific methyltransferase 2D (KMT2D) is a critical component of epigenetic regulation through its role in mono-methylation of lysine 4 of histone H3 (H3K4me1). KMT2D is among the most frequently mutated genes in many forms of cancer, with particularly high occurrence of mutation in lymphoid malignancies. Despite being the recurrent target of somatic alteration across many cancer types, the consequences of KMT2D mutation, and their relevance to tumorigenesis, remain unclear. To expand on the current understanding of KMT2D loss, I performed comprehensive and integrative bioinformatics analyses of the epigenetic and transcriptome landscapes of isogenic KMT2D-mutant HEK293A cell lines. Analysis of ChIPsequencing data from KMT2D-mutant cells showed genome-wide alterations in the distribution of H3K4me1, with loss of H3K4me1 occurring at active and poised enhancer regions. Interestingly, epigenetic disruption of enhancers in KMT2D-mutant cells was not sufficient for inducing transcriptional alteration of nearby genes, indicating a possible requirement for additional co-factors to be present in order to observe the consequences of KMT2D-dependent enhancer loss. Genes associated with KMT2D-dependent enhancers were enriched for members of the TGF-beta and retinoic acid (RA) signaling networks, highlighting transcriptional response to these pathways as candidate processes in which functional KMT2D-dependent enhancers may be required. Given the roles of both TGF-beta and RA signaling in cancer, identification of the convergence between the KMT2D regulome and these signaling axes provides a potential means by which *KMT2D* mutations may contribute to tumorigenesis.

Lay Summary

Cancer cells often show an accumulation of mutations, which can disrupt the genes of a cell. One particular gene, known as *KMT2D*, is one of the most frequently mutated genes in lymphoma, as well as other types of cancer. For this reason, studying the consequences of mutations in *KMT2D* may allow us to generate biological insight into many different forms of cancer. KMT2D is known to play a role in the regulation of other genes, by activating enhancer regions that control their expression. To explore the impact of *KMT2D* mutations, I identified enhancer alterations in *KMT2D*-mutant cells, as well as the types of genes these enhancers regulated. Through my research, I uncovered evidence indicating that *KMT2D* mutation is associated with deficiencies in TGF-beta and retinoic acid signaling. Disruption of these pathways is known to occur in cancer, and is perhaps a mechanism by which *KMT2D* mutation contributes to the disease.

Preface

Investigation into the epigenomic and transcriptome alterations associated with *KMT2D* mutation was conceived by Dr. Marco Marra. Ryan Huff and Diane Trinh generated the isogenic KMT2D-mutant and wildtype HEK293A cell lines. James Topham, Dr. Alessia Gagliardi and Diane Trinh conceived strategies regarding the characterization of epigenome and transcriptome data sets. James Topham performed all bioinformatics analyses included in this thesis project. Dr. Emilia Lim, Elizabeth Chun, Rodrigo Goya and Dr. Davide Pellacani all provided technical advice regarding several of the bioinformatics analyses.

Table of Contents

Abstract	ii
Lay Summary	iii
Preface	iv
Table of Contents	v
List of Figures	ix
List of Abbreviations	xi
List of Genes	xiv
Acknowledgements	xvii
Dedication	xviii
Chapter 1: Introduction	1
1.1 Overview	1
1.2 On the study of cancer	1
1.3 Cancer as a disease of the genome	3
1.3.1 Somatic mutations act to facilitate tumor evolution	4
1.3.2 Oncogenes and tumor suppressors	5
1.3.3 Genes frequently mutated across cancers	7
1.4 Cancer as a disease of the epigenome	9
1.4.1 Epigenetic dysregulation in cancer	11
1.5 KMT2D	
1.5.1 KMT2 family proteins	
1.5.2 The KMT2D-containing complex	14

1.5.3	KMT2D and enhancers	15
1.5.4	KMT2D and Kabuki Syndrome	16
1.5.5	Incidence of KMT2D mutation in cancer	16
1.5.6	Consequences of KMT2D mutation	17
1.6 Th	esis roadmap and chapter summaries	20
Chapter 2	: Materials and methods	.29
Chapter 3	: Results	.35
3.1 KM	MT2D mutation is associated with alterations in the epigenetic regulatory landscape.	35
3.1.1	Genome-wide profiling of H3K4me1 enrichment supports global loss of H3K4me	:1
in KM	IT2D-deficient HEK293 cells	36
3.1.2	A subset of H3K4me1 peaks detected in KMT2D-wildtype cells are consistently lo	ost
in <i>KM</i>	TT2D-mutant cells	39
3.1.3	KMT2D-dependent H3K4me1 loss occurs at regions co-occupied by H3K27ac bu	.t
not H3	3K4me3	42
3.1.4	KMT2D deficiency is associated with a genome-wide decrease in the number of	
active	enhancers	44
3.2 KM	MT2D mutation is associated with transcriptional alterations	46
3.2.1	KMT2D mutation is associated with decreased transcript abundance	47
3.2.2	KMT2D-dependent genes converge on pathways related to structural organization	l
and ce	ellular adhesion	48
3.3 Inte	egrative analysis of the KMT2D regulome reveals convergence between KMT2D-	
depende	nt enhancers and the TGF-beta and RA signaling networks	49

	3.3.1	A subset of KMT2D-dependent genes are proximally associated with KMT2D-	
	depe	ndent enhancers	50
	3.3.2	Alteration of KMT2D-dependent enhancers alone is not sufficient for modulating	
	trans	cription of nearby genes	52
	3.3.3	Genes associated with KMT2D-dependent active enhancers are enriched for genes	
	up-re	egulated upon RA treatment	54
Ch	apter 4	4: Discussion	76
۷	l.1 <i>K</i>	MT2D mutation is associated with alterations in the epigenetic regulatory landscape.	76
۷	1.2 K	MT2D mutation is associated with transcriptional alterations	79
۷	1.3 In	ntegrative analysis of the KMT2D regulome reveals convergence between KMT2D-	
Ċ	lepend	ent enhancers and the TGF-beta and RA signaling networks	83
۷	1.4 C	oncluding remarks	87
Bib	oliogra	phy	89
Ap	pendic	ces1	05
A	Append	dix A Differentially expressed genes identified in <i>KMT2D</i> -mutant HEK293 cell lines	
		1	05
A	Append	dix B mSigDB gene sets significantly enriched among genes up-regulated in KMT2D-	
r	nutant	cell lines1	13
A	Append	dix C Top 200 mSigDB gene sets most significantly enriched among genes down-	
r	egulate	ed in <i>KMT2D</i> -mutant cell lines	14
		dix D GO terms significantly enriched among genes down-regulated in <i>KMT2D</i> -muta	
	ell line		20

Appendix E KEGG pathways significantly enriched among genes down-regu	lated in <i>KMT2D</i> -
mutant cell lines	130
Appendix F Curated gene sets significantly enriched among genes associated	with KMT2D-
dependent active enhancers	131

List of Figures

Figure 1.1	Top 375 most frequently mutated genes across 32 cancer types	26
Figure 1.2	Frequency, distribution and types of KMT2D mutations across Kabuki	
syndrome a	nd 34 cancer types.	28
Figure 3.1	Genomic coverage of peaks achieves saturation for each ChIP-seq library	at
their respec	tive sequencing depths.	56
Figure 3.2	KMT2D mutation is associated with a global loss of H3K4me1.	57
Figure 3.3	Regions that gain or lose H3K4me1 in KMT2D-mutant cells are enriched for	or
CTCF and	AP-1 DNA binding motifs, respectively.	59
Figure 3.4	H3K4me1 peaks gained or lost in KMT2D-mutant cell lines occur	
predomina	ntly at promoter and enhancer regions, respectively	61
Figure 3.5	A subset of H3K4me1 peaks gained in KMT2D-mutant cell lines overlap w	ith
TSS regions	S•	62
Figure 3.6	Poised and active enhancers detected in wildtype HEK293 cells are distal t	0
promoter r	egions.	64
Figure 3.7	Mutation of KMT2D is associated with loss of histone modifications H3K4r	ne1
and H3K27	me3/ac at poised and active enhancers, respectively	66
Figure 3.8	A subset of KMT2D-dependent poised and active enhancers lose	
H3K27me3	/ac (respectively), despite retaining H3K4me1, in KMT2D-mutant cells	67
Figure 3.9	KMT2D mutation is associated with decreased transcript abundance	68
Figure 3.10	KMT2D-dependent genes are enriched for pathways related to cytoskelet	al
organizatio	n and cellular adhesion.	69
Figure 3.11	A subset of DE genes are within 100 kb of a KMT2D-dependent enhancer	· 71

Figure 3.12	Genes associated with KMT2D-dependent active enhancers include targets	of
of the TGF-l	beta and retinoic acid signaling pathways	73
Figure 3.13	RA target genes associated with KMT2D-dependent active enhancers are	
poised for tr	anscriptional activation.	75

List of Abbreviations

H3K4me1 Histone 3 lysine 4 mono-methylation

DNA Deoxyribonucleic acid
PCR Polymerase chain reaction
NGS Next-generation sequencing
TCGA The Cancer Genome Atlas
SNV Single nucleotide variant
WGS Whole-genome sequencing

GOF Gain of function
LOF Loss of function
TS Tumor suppressor

H3K27me3 Histone 3 lysine 27 tri-methylation

WES Whole-exome sequencing

H2A Histone 2A H2B Histone 2B H3 Histone 3 H4 Histone 4

ChIP-seq Chromatin immuno-precipitation sequencing

H3K4me3 Histone 3 lysine 4 tri-methylation

ChIA-PET Chromatin interaction analysis by paired-end tag sequencing

H3K27ac Histone 3 lysine 27 acetylation

RNA Ribonucleic acid

RING Really interesting new gene hg19 human genome build 19

CDS Coding sequence

kDA kilodalton ZNF Zinc finger

SET Su(var)3-9, enhancer of zeste and trithorax

PEV Position-effect variegation

COMPASS Complex proteins associated with Set1

ASCOM ASC2-containing complex

GC Germinal center

DLBCL Diffuse large B-cell lymphoma

ESC Embryonic stem cell
KS Kabuki syndrome
MLL Mixed lineage leukemia
MCL Mantle cell lymphoma
MED Medulloblastoma

NMZL Nodal marginal zone lymphoma

BRFE Breast fibrous epitheal

SKCM Skin cutaneous melanoma
ESCA Esophageal carcinoma
PAAD Pancreatic adenocarcinoma
STAD Stomach adenocarcinoma

UCEC Uterine corpus endometrial carcinoma

shRNA Small hairpin RNA

RNA-seq Ribonucleic acid sequencing cAMP Cyclic adenosine mono-phosphate

mRNA Messenger ribonucleic acid

TF Transcription factor Pol II DNA polymerase II

HEK293A Human embryonic kidney 293A TGF-beta Transforming growth factor beta

RA Retinoic acid

UCSC University of California Santa Cruz

INDEL Insertion/deletion
IGR Intergenic region
UTR Untranslated region
TSS Transcription start site

mSigDB Molecular signatures database GSEA Gene set enrichment analysis

GO Gene ontology

KEGG Kyoto encyclopedia of genes and genomes

MHC Multiple hypothesis correction

BH Benjamini-Hochberg

GREAT Genomic regions enrichment of annotations tool

H3K9me3 Histone 3 lysine 9 tri-methylation

bp Base pair

AP-1 Activator complex-1

VSMC Vascular smooth muscle cells
DE Differentially expressed

FC Fold change

ECM Extra-cellular matrix

TCC Tethered chromatin capture
DEA Differential expression analysis

RPKM Reads per million mapped reads per thousand bp

WT Wildtype

RPM Reads per million mapped reads
EMT Epithelial-mesenchymal transition

RAR Retinoic acid receptor

MEF Mouse embryonic fibroblast

MDS eRNA Myelodysplastic syndrome Enhancer ribonucleic acid

List of Genes

KMT2D Lysine (K)-specific methyltransferase 2DHER2 Receptor tyrosine-protein kinase erbB-2

TP53 Tumor protein 53

EZH2 Enhancer of zeste homolog 2

PIK3CA Phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit alpha

APC Adenomatous polyposis coli

TTN Titin MUC16 Mucin 16

CSMD3 CUB and sushi multiple domains protein 3

RYR2 Ryanodine receptor 2

ARID1A AT-rich interaction domain 1A PTEN Phosphatase and tensin homolog

KMT2C Lysine (K)-specific methyltransferase 2C

EP400 E1A binding protein p400

MLH1 Mutl homolog 1 BRCA1 Breast cancer 1

IGF2 Insulin-like growth factor 2

ALR ALL1-related protein

MLL2 Mixed lineage leukemia 2

MLL4 Mixed lineage leukemia 4

KMT2E
 KMT2A
 Lysine (K)-specific methyltransferase 2E
 KMT2A
 Lysine (K)-specific methyltransferase 2A
 KMT2B
 Lysine (K)-specific methyltransferase 2B
 KMT2C
 Lysine (K)-specific methyltransferase 2C

UTX Ubiquitously transcribed tetratricopeptide repeat X chromosome

PTIP PAX interacting protein

PA1 Plasminogen activator inhibitor-1NCOA6 Nuclear receptor coactivator 6

WDR5RBBP5RB binding protein 5hDPY30DPY-like protein 30

ASH2 Absent small or homeotic-like 2

KDM6ALysine demethylase 6ALAMB3Laminin subunit beta 3LOXL1Lysyl oxidase like 1

GPR56 G protein coupled receptor 56

VTN Vitronectin
PCDH7 Protocadherin 7

CRIP1 Cysteine rich protein 1

FHL1 Four and a half LIM domains 1 S100A4 S100 calcium binding protein A4

PKIA cAMP-dependent protein kinase inhibitor alpha

RPS6KA1 Ribosomal protein S6 kinase A1

PPP3CA Protein phosphatase 3 catalytic subunit alpha

IL7 Interluekin 7

IL7R Interluekin 7 receptor

GJA1 Gap junction protein alpha 1

SOX4 SRY-box 4

P300 E1A-associated proten p300 TNFAIP3 TNF alpha induced protein 3

SOCS3 Suppressor of cytokine signaling 3
TNFRSF14 TNF receptor superfamily member 14

FOS c-FOS proto-oncogene

JUN Jun proto-oncogene

CTCF CCCTC-binding factor

MAPK8 Mitogen-activated protein kinase 8

JNK c-Jun NH2-terminal kinase

SSR4 Signal sequence receptor subunit 4 TOR1AIP1 Torsin 1A interacting protein 1

CTTN Cortactin

INSIG1 Insulin induced gene 1

SLC6A6 Solute carrier family 6 member 6

GATA6 GATA binding protein 6

AIMP2 Aminoacyl TRNA synthetase complex interacting multifunctional protein 2

STX4 Syntaxin 4

BMP2 Bone morphogenic protein 2 ZNF532 Zinc finger protein 532

PARP14 Poly(ADP-ribose) polymerase family member 14

SLC2A3 Solute carrier family 6 member 6

GYG2 Glycogenin 2

SFRP2 Secreted frizzled related protein 2

PGR Progesterone receptor

SLC35F1 Solute carrier family 6 member 6 BRSK2 BR serine/threonine kinase 2

PCDH10 Protocadherin 10

S100A16 S100 calcium binding protein A16 VEGFA Vascular endothelial growth factor A

CUTL1 Cut-like 1

FGF12 Fibroblast growth factor 12 EGR1 Early growth response 1

ST6GALNAC5 ST6 N-acetylgalactosaminide alpha-2,6-sialytransferase 5

TAC1 Tachykinin

RB1 Retinoblastoma-associated protein 1

SMAD2 SMAD family member 2 TCF21 Transcription factor 21

ESRRA Estrogen related receptor alpha RARG Retinoic acid receptor gamma

CREBBP CREB binding protein

Acknowledgements

I would first like to thank Dr. Marco Marra for his incredible guidance and support towards my scientific development. Thank you Dr. Emilia Lim for your amazing mentorship and positive impact on the trajectory of my career as a scientist. Thank you Dr. Alessia Gagliardi, Elizabeth Chun and Dr. Suganthi Chittaranjan for your immense support during my time in the Marra lab. I would also like to thank all other members of the Marra lab, along with Lulu Crisostomo, for their continual encouragement, advice and support. Thank you to my committee members Dr. Sara Mostafavi and Dr. Matthew Lorincz for their guidance, as well as Dr. Paul Pavlidis, Dr. Martin Hirst and Dr. Steven Jones for the additional mentorship they provided throughout my time as a graduate student. Lastly, I would like to thank the Canadian Institutes of Health Research (CIHR) for project funding.

Dedication

To Tom Melnyk.

Chapter 1: Introduction

1.1 Overview

Cancer manifests as a common and often incurable disease, and represents a fundamental motivation in the overall context of my thesis work. Therefore, I begin by introducing a brief history of the study of cancer in Section 1.2, describing several milestones in cancer research that have contributed to our understanding of the pathological attributes of tumors. I then go on to introduce cancer as a disease of the genome in Section 1.3, with reflections on an experiment I performed to assay the most frequently mutated genes across multiple cancer types. From this experiment emerged the observation that genes involved in epigenetic regulation are represented among the top most frequently mutated genes across cancers, which leads to an introduction of cancer as a disease of the epigenome in Section 1.4. Given the lack of knowledge regarding the consequences of mutations in many epigenetic regulatory genes in cancer, together with the high frequency of mutations in key epigenetic regulator *KMT2D* across cancers, *KMT2D* is highlighted as a model to study cancer epigenetics. This then leads to a comprehensive introduction into KMT2D biology in Section 1.5. Finally, Section 1.6 outlines the specific hypotheses and aims addressed in the Results section of my thesis.

1.2 On the study of cancer

With a pathological profile that in many ways parallels the natural process of ageing¹, cancer represents a disease that occurs frequently among human populations². Consistent with a high occurrence of cancer among humans, thereby imparting a large impact on human society, cancer has been documented throughout much of human history³. Some of the earliest descriptions of

cancer can be dated as far back as ancient Egyptian texts such as the Edwin Smith Papyrus, in which nodules of the breast were described as "bulging tumors" for which "there [was] no treatment". Later, Greek physicians would use terms such as *carcinoma* and *cancer* (translated as "crab") to describe the swollen masses with surrounding vasculature resembling the legs of a crab. By the 19th century, advancements in surgical medicine, concomitant with the discovery and use of anaesthetic agents, enabled some of the first routine surgical operations for treating cancer, involving removal of tumors from cancer patients^{3,6}.

To this day, there remains no reliable cure for many forms of human cancer, positioning cancer research as a pivotal inquisition into the biology of tumors, as a means to establish better therapeutic measures. While early cancer research was largely limited to morphological observations of tumor samples⁷, the field greatly benefited from several 20th century scientific advances in the field of genomics, including the development of Sanger DNA sequencing in 1977⁸ and polymerase chain reaction (PCR) amplification of DNA fragments in 1985⁹. Subsequent mapping of the human genome, completed in 2001¹⁰, would mark the next milestone in cancer genomics research by providing a comprehensive reference sequence that cancer genomes could be compared against. More recent advancements in whole-genome interrogation through next-generation sequencing (NGS) methods have now driven a multitude of studies on many forms of human cancers, perhaps demonstrated best through the efforts of The Cancer Genome Atlas (TCGA) project¹¹, which has produced publicly available NGS data for thousands of tumor samples and a multitude of publications.

Years of detailed study into the biology of human tumors have lead to the concept of canonical *hallmarks* of cancer, chosen due to their consistency across the many forms of malignancy. Hanahan and Weinberg outlined six hallmarks of cancer in 2000¹², which were claimed to be acquired by nearly all forms of cancer at some point during the developmental trajectory of the disease. These consisted of: evasion of apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals, tissue invasion and metastasis, limitless replicative potential and sustained angiogenesis. The same authors went on to describe four additional hallmarks in 2011, adding deregulation of cellular energetics, avoiding immune destruction, tumor-promoting inflammation and genome instability to the list of cancer hallmarks¹³.

Despite limitations imposed by the complexity of the human genome and how it is regulated, our understanding of tumor development continues to build as new technologies for genome interrogation become available, such as profiling individual tumor cells with single-cell sequencing to characterize the complex heterogeneity of tumors¹⁴. Genomic technologies therefore represent a key component to contemporary cancer research, and continue to generate an impetus for advancements in our knowledge of tumor biology.

1.3 Cancer as a disease of the genome

Experimental observations indicating an association between human cancer and gross somatic alterations of the genome preceded the use of genome-wide sequencing technologies, and can be seen in many early studies including those related to DNA copy number alterations (such as amplifications encompassing the *HER2* gene in breast cancer¹⁵) and gene fusions (with a classic example being the *BCR-ABL* fusion in chronic myologenous leukemia¹⁶). With the recent

widespread application of genome sequencing technologies to profile tumor genomes, it has become clear that tumors are often characterized by a constellation of somatic DNA alterations, including changes in copy number, translocation events, insertion/deletions and single nucleotide variants (SNVs). SNVs represent one of the most common alterations in cancer, as tumors often bear a hypermutation phenotype, as revealed through whole-genome sequencing (WGS) of several cancer types including lung squamous cell carcinoma and melanoma, which showed, per individual tumor, an average of over 30,000 unique SNVs of any type¹⁷. The function and consequence of somatic mutation in cancer have been explored extensively, and are the focus of the following section.

1.3.1 Somatic mutations act to facilitate tumor evolution

Somatic mutations are DNA alterations that occur in cells not belonging to the germline, and are therefore not inherited by subsequent generations. The acquisition of somatic mutations can lead to fitness benefits for individual cells relative to surrounding cells within the same organism. In this way, single cells can be thought of as individuals competing for resources (including oxygen, nutrients and space) within a particular environment, such as a specific area or tissue of the body. Enhanced capabilities to compete for resources among a group of cells can manifest as early tumorigenesis, as mutations enhancing cellular fitness can positively affect hallmarks of cancer such as cell growth, replicative potential, and angiogenesis¹². Acquisition of such mutations is frequently observed in tumors demonstrating hypermutation¹⁷, in which extensive collections of somatic mutations provide enhanced opportunity for tumor cells to prosper. Nonsynonymous SNVs often have a functional impact on the genes they affect, as they alter the amino acid sequence of the encoded protein. Meanwhile, the impact of synonymous SNVs is less

obvious, as they do not change the amino acid sequence of the protein product of the gene.

Codon usage bias represents the process by which individual tri-nucleotide sequences of DNA, which redundantly encode the same amino acid, are utilized at different frequencies throughout the genome of an organism¹⁸. Differential codon usage has been shown to impact translation efficiency of genes¹⁹, and thereby provides a functional role of synonymous SNVs. By conferring cellular attributes aligning to the hallmarks of cancer, the ability for somatic mutations to underpin tumor biology is of significant research interest, and is discussed in the following section.

1.3.2 Oncogenes and tumor suppressors

H.J. Muller (1932) conceptualized several different forms of mutation, each being capable of impacting gene function in unique ways that converge on either gain or loss of gene function (GOF and LOF, respectively)²⁰. LOF mutations may phenotypically manifest as either amorphic or hypomorphic, in which gene expression is completely abrogated or partially reduced, respectively. Frequently observed in tumor cells, LOF mutations are often found to reside in tumor suppressor (TS) genes; defined as those that normally function to safeguard cells from tumorigenesis²¹. One example of a TS relevant in cancer is that of tumor protein 53 (*TP53*), which has been found to be one of the most frequent targets of somatic mutation across many human cancers^{22,23}. TP53 is known to be involved in the response to several cellular stresses such as DNA damage and hypoxia, with responses including protective measures such as apoptosis, cell cycle arrest and senescence²⁴. In certain contexts, LOF mutations in *TP53* may drive cancer progression through various means, for example, by attenuating apoptosis mediated by *TP53* expression and thus allowing tumor cells to evade cellular senescence in prostate cancer²⁵.

Analogous to LOF mutations, copy number deletions, in which the number of allelic copies of a gene is reduced (relative to a diploid model), represent an orthogonal mechanism of abrogating gene function. As such, copy number deletions in tumor suppressor genes, such as *TP53*, have been shown to drive tumor progression by means similar to that of deleterious mutations²⁶.

Despite several examples of TP53 acting as a tumor suppressor in many contexts, it is important to know that certain mutations in *TP53* can also confer increases in TP53 levels in the context of high-grade serous ovarian carcinoma²⁷, highlighting the importance of context-specificity when considering the impact of somatic alteration of certain genes.

SNVs may also impart gain-of-function (GOF) effects on genes, where mutations confer effects advantageous to the functioning/activity of a gene such as increased expression, protein activity and stability. GOF mutations often occur at specific (hotspot) locations along the gene body, as there are theoretically fewer positions where SNVs have the ability to enhance gene activity (compared to LOF mutations, which could affect the protein at many locations along the length of the gene). While hypermorphic mutations simply elicit GOF through increases in gene product activity, neomorphic and antimorphic mutations are unique GOF mutations in that they act to impart novel gene function which, in the case of antimorphic mutation, can elicit an effect that contradicts that of the genes from which they are derived²⁸. In contrast to LOF mutations, which often target TS genes, GOF mutations are often found to affect genes with functions that can act to facilitate tumorigenesis, particularly when over-activated. An example of a hypermorphic GOF mutation in cancer is the Y641 mutation in the enhancer of zeste homolog 2 (*EZH2*) gene, observed in lymphomas²⁹, which acts to increase EZH2-mediated histone 3 lysine 27 trimethylation (H3K27me3)³⁰. Copy number amplifications represent an additional mechanism by

which gene function can be positively altered, and are commonly observed in many tumor genomes³¹.

The means by which somatic mutations contribute to malignant transformation and progression is an important concept in tumor biology. In cancer, driver mutations are broadly defined as those directly contributing to tumorigenesis, with examples including LOF and GOF alterations in tumor suppressors and oncogenes, respectively. In contrast, passenger mutations are those that may not directly promote tumorigenesis, but can persist across tumor cell lineages through indirect selection, such as the hitchhiker effect in which passenger mutations are in close enough proximity to driver alterations that they are not separated from the driver locus during recombination³². Distinguishing between driver and passenger mutations has been achieved for a subset of mutations that are highly recurrent across cancer types, under the rationale that driver mutations are more likely to re-occur across distinct cancer types relative to passenger mutations²¹.

1.3.3 Genes frequently mutated across cancers

Many forms of cancer demonstrate significant inter-tumor heterogeneity, in which the profiles of somatic mutation differ across tumors. However, a subset of genes can be identified as targets of somatic mutation across multiple forms of malignancy, given their significant relevance in cancer as tumor suppressors or oncogenes. In this way, identification of genes recurrently mutated across distinct cancer types can highlight the widespread relevance of known driver genes, such as *TP53*, *PIK3CA* and *APC*²¹. Furthermore, studies that assay the most frequently mutated genes across cancers can also provide insight into whether such genes tend to converge

on specific pathways or functions. To this end, I conducted a brief experiment to identify the most frequently mutated genes across cancers using whole-exome sequencing (WES) data for 8,598 tumor samples (encompassing 32 distinct cancer types) obtained from TCGA, using the types of mutations provided by TCGA (see Methods). The top 375 most frequently mutated genes in TCGA WES data are depicted in Figure 1.1. Among the top most frequently mutated genes across these cancer types are very large genes such as titin (TTN; coding sequence (CDS) 304,814 bp in size), mucin 16 (*MUC16*; 132,499 bp), CUB and sushi multiple domains 3 (CSMD3; 1,214,540 bp), and ryanodine receptor 2 (RYR2; 791,784 bp). While some very large genes may have roles in tumorigenesis, it is important to note that this stratification of genes by the number of mutations identified is biased towards larger genes, as these genes are more likely to be targets of mutation by random chance given their large size (and thereby are likely to represent targets of passenger alterations). Consistent with the notion of TS gene TP53 being relevant in many cancer types^{22,23}, TP53 is the second most frequent target of somatic mutation highlighted in my analysis. What also emerges is the trend in which many of the most frequently mutated genes in cancer are known tumor suppressors. These include, but are not limited to, (K) methyltransferase 2D (KMT2D³³), AT-rich interaction domain 1A (ARID1A³⁴), adenomatous polyposis coli (APC^{35}) and phosphatase and tensin homolog $(PTEN^{36})$.

Interestingly, many genes with functions in epigenetic regulation reside among the most frequently mutated genes in TCGA WES data (Figure 1.1), including lysine (K) methyltransferase 2C (*KMT2C*) and 2D (*KMT2D*), both epigenetic writers known to establish H3K4me1. Others include E1A binding protein 400 (*EP400*), functioning in nucleosome remodeling, as well as *ARID1A*, a member of the SWI-SNF complex, also involved in

nucleosome remodeling. The importance of alterations in the epigenetic landscapes of cells has recently emerged as a substantial factor in the progression of many forms of cancer^{37,38}, and is the focus of the following section.

1.4 Cancer as a disease of the epigenome

Epigenetic regulation encompasses chemical modifications to DNA and histones that do not involve alteration of the underlying nucleotide sequence, with many forms of epigenetic regulation playing important roles in modulating transcriptional activity. DNA methylation is a common example of epigenetic regulation in mammalian cells, with one example involving the addition of a methyl group to cytosine residues in a CpG dinucleotide context³⁹. The effect of DNA methylation on transcription can be repressive, with increased methylation patterns associated with decreased expression of genes, as the addition of methyl groups to DNA can interfere with binding of transcription factors required in transcription⁴⁰. However, this is not always the case as DNA methylation is context specific, with increased methylation at promoters being associated with decreased gene expression⁴¹ and increased methylation in gene bodies being associated with increases in gene expression⁴².

The nucleosome represents a fundamental unit of DNA packaging, in which DNA is wrapped around a histone core consisting of two copies of each of four histone protein subunits (H2A, H2B, H3 and H4). Modification of histone proteins represents a common mechanism of epigenetic regulation, where several covalent modifications such as methylation, acetylation, ubiquitylation, sumoylation and phosphorylation are established at specific tail residues of histones, with many having consequences related to chromatin structure⁴³. ChIP-sequencing

(ChIP-seq) technology enables high-resolution identification of various histone modifications genome-wide. Briefly, this approach often involves protein-DNA complex crosslinking, application of antibodies specific to a histone modification of interest, chromatin fragmentation and purification. Crosslinks are then reversed, and the remaining DNA is subjected to sequencing⁴⁴.

The impact of histone modifications on transcriptional regulation can be dependent on both the type of histone modification and the histone residue affected. For instance, tri-methylation of lysine 27 on histone H3 (H3K27me3) at promoter regions is often associated with transcriptional repression⁴⁵, while mono-methylation of lysine 4 of histone H3 (H3K4me1) and absence of H3K27me3 at promoters can be associated with transcriptional activation⁴⁶. Distinct, overlapping patterns of multiple histone modifications have been shown to represent functional domains across the genome, each with various consequences in regards to transcriptional regulation. For instance, the presence of both active mark H3K4me3 and repressive mark H3K27me3 at gene promoters defines a bivalent state that has been shown to enable transcriptional plasticity, in which transcription is poised for rapid activation⁴⁷.

While promoter regions reside proximal to exonic regions, enhancer regions are *cis* regulatory elements that can modulate transcriptional activity of genes from a greater distance. Using chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), the distances spanned by enhancer-gene interactions have been found to range between 1-100 kb⁴⁸. Enhancers can be classified into three commonly observed states based on the presence or absence of three histone modifications: H3K4me1, H3K27me3 and H3K27ac. Active enhancers are decorated with

activating marks H3K4me1 and H3K27ac, and act to facilitate transcription of target genes. Meanwhile, bivalent (or 'poised') enhancers (H3K4me1 and H3K27me3) and primed enhancers (H3K4me1 only) are said to limit target gene expression to very low levels, as they await activation through gain of H3K27ac (in the case of primed enhancers) or replacement of H3K27me3 with H3K27ac (poised enhancers)⁴⁹.

The field of epigenetics offers insights into different chemical modifications to structures such as DNA, histones and even RNA⁵⁰. Despite an emerging accumulation of research, there remain many gaps in our knowledge of the different types of epigenetic modifications, in terms of the functional domains they inhabit across the genome, their impact on transcriptional regulation and how their dysregulation may be relevant to tumorigenesis.

1.4.1 Epigenetic dysregulation in cancer

Epigenetic regulatory mechanisms are fundamental to the maintenance and timing of developmental programs in human cells, and deleterious effects can result when the epigenome is dysregulated³⁷. Epigenetic patterns controlling gene expression represent heritable cellular attributes, as they are transmitted to daughter cells during cell division⁵¹. In this way, abnormalities in the normal epigenetic landscape of cells can be propagated across cell lineages, and result in metastable patterns capable of facilitating or predisposing malignancy. In many forms of cancer, epigenetic alterations have been attributed to disease progression^{37,38}, and represent a unique opportunity to better understand malignant transformation and general tumor biology.

As previously mentioned, the dysregulation of TS genes and oncogenes represents a fundamental mechanism of tumorigenesis, with many cancers bearing somatic mutations in these genes that alter their function or expression in ways that can promote a malignant phenotype. In addition to somatic mutation, the aberrant modulation of TS gene and oncogene transcription can be achieved through alterations in the epigenetic regulation of these genes. For example, abnormal accumulation of histone modifications associated with densely packaged (therefore inaccessible to transcription factors and other elements that normally promote transcription) heterochromatin can result in reduced expression of TS genes residing in the vicinity of such alterations. Such an event has been observed for the TS gene MLH1 in colorectal cancer cell lines, in which enrichment of H3K9 methylation, a repressive heterochromatin-associated mark, at the MLH1 promoter is associated with transcriptional silencing of MLH1⁵². Another example of epigenetic dysregulation relevant to cancer is the abnormal formation of open, transcriptionally active chromatin (euchromatin) in areas of the genome that would normally contain compact, transcriptionally silent chromatin (heterochromatin). This phenomenon has been shown to occur in mice with conditional knockout of *Brca1*, a TS gene frequently affected by LOF mutation in ovarian and breast cancers^{53,54}. Brca1-deficient mice showed up-regulation of the *Igf2* oncogene via alleviation of heterochromatin-induced repression through loss of H2A ubiquitination, a repressive histone modification that was found to be performed by the RING finger domain of functionally active Brca1⁵⁵.

While significant insight into the effect of deleterious mutations has been generated for *BRCA1*, the functional consequences of mutation remain relatively unknown for many other genes, including those known to primarily function in epigenetic regulation. This then leads to a large

gap in our understanding of tumor biology, as many epigenetic regulators are frequently mutated across many types of cancer (Figure 1.1). The epigenetic regulator *KMT2D* is a TS gene that is one of the most frequent targets of somatic mutation across 32 distinct cancer types (Figure 1.1). Despite their significant presence in many cancers, the consequences of *KMT2D* mutations, and their relevance in tumorigenesis, are poorly understood. *KMT2D* mutation therefore presents a model in which to study the mutation of epigenetic regulatory genes in cancer.

1.5 KMT2D

According to reference human genome build 19 (hg19), lysine (K)-specific methyltransferase 2D (*KMT2D*) encompasses 42,921 bp (CDS) on chromosome 12q13.12. Also occasionally referred to as *ALR* or *MLL2/4*, *KMT2D* encodes a relatively large protein product (5,537 amino acids in size; measuring 593 kDa) that contains PHD, HMG-box, LXXLL and SET domains⁵⁶. As *KMT2D* is a central focus in my thesis, this section serves to outline KMT2D biology and the relevance of *KMT2D* mutation in disease, along with an overview of what is currently known regarding the molecular consequences of *KMT2D* mutation.

1.5.1 KMT2 family proteins

KMT2D belongs to the highly conserved⁵⁷ KMT2 family of seven proteins that each contain a catalytic SET (Su(var)3-9, Enhancer of zester and Trithorax) domain which, with the exception of KMT2E⁵⁸, confers intrinsic histone methyltransferase activity. The discovery of genes encoding proteins with SET domains emerged from early investigations into heterochromatin-mediated transcriptional regulation in *Drosophila melanogaster*. Position-effect variegation (PEV) is defined as the abnormal silencing of genes resulting from the juxtaposition of

euchromatin and heterochromatin. Early PEV screens performed in *Drosophila* identified suppressors and enhancers (Su(var)s and E(var)s, respectively) of PEV, which often included genes with a 130-140 amino acid motif coined the SET domain⁵⁹, named after three SET-containing *Drosophila* genes: *Su(var)3-9*, *Enhancer of zeste* and *trx*⁶⁰. *Trx* is the ancestral homolog of mammalian genes *KMT2A/B*, while a similar gene in *Drosophila*, *trr*, is homologous to the mammalian genes *KMT2C/D*.

1.5.2 The KMT2D-containing complex

Identified in early experiments using yeast (*S. cerevisiae*), the COMPASS (<u>Complex of Proteins As</u>sociated with <u>Set1</u>) complex was the first protein complex discovered to be capable of H3K4 methylase activity, conferred by a SET-containing, *trx*-related yeast protein Set1⁶¹. Since then, six COMPASS-like complexes have been identified in humans, one of which contains either KMT2C or KMT2D as a catalytic methyltransferase component⁶². This COMPASS-like complex, containing KMT2C/D, has also referred to as the ASC-2 complex (ASCOM)⁶³ and will be further referred to here as the KMT2D complex. Along with KMT2D, the KMT2D complex has also been found to contain histone demethylase UTX, as well as PTIP, PA1, NCOA6, WDR5, RBBP5, hDPY30 and ASH2⁶².

That the KMT2D complex can incorporate either KMT2C or KMT2D as the catalytic component is compatible with the notion that the two methyltransferases share some level of functional redundancy. However, KMT2C/D have been found to be functionally redundant in only a handful of experimental contexts, which often converge on the function of the KMT2D complex in receptor transactivation^{64–67}. In contrast, particularly when focused on the ability of the

KMT2D complex to establish genome-wide H3K4me1, KMT2C/D are found to be only partially redundant, as individual abrogation of either KMT2C or KMT2D results in H3K4me1 being lost at some but not all genomic locations^{68–71}.

1.5.3 KMT2D and enhancers

Several studies have demonstrated the importance of KMT2D in maintaining genome-wide levels of H3K4me1. For instance, genome-wide H3K4me1 levels have been compared between wildtype and KMT2D deficient cells in multiple cellular contexts including colorectal cancer cell line HCT116, germinal center (GC) B cells, DLBCL cell lines and embryonic stem cells (ESCs)^{33,68–70,72,73}. H3K4me1 is thought to mark enhancer regions of the genome, residing at active (marked by H3K4me1 and H3K27ac), poised (H3K4me1 and H3K27me3) and primed (H3K4me1 only) enhancers⁴⁹. KMT2D was first found to localize at mammalian enhancer regions in HCT116 cell lines by Guo and collaborators⁷², with the association between KMT2D and enhancer regions being further supported by evidence generated in several subsequent studies^{68,74}. Furthermore, the role of KMT2D in establishing H3K4me1 at enhancer regions has been demonstrated in several studies^{68,72,74}, thereby highlighting enhancer-specific H3K4 monomethylation as a major function of KMT2D.

Enhancers are critical in orchestrating the precise timing of transcriptional events during cellular differentiation and development, with aberrancies in the establishment and maintenance of enhancers being implicated in disease^{75,76}. KMT2D malfunction therefore presents an opportunity to study the effects of enhancer dysregulation in disease. The relevance of KMT2D in human disease is therefore discussed in the following section.

1.5.4 KMT2D and Kabuki Syndrome

Kabuki syndrome (KS) is a developmental syndrome in humans that is generally characterized by malformation and mental retardation, along with variable phenotypic attributes including distinctive facial features, skeletal abnormalities and inhibited body growth⁷⁷. The primary cause of KS is mutation of *KMT2D*, with as many as 74% of KS patients bearing *KMT2D* mutations, many of which confer LOF^{78,79}. Alongside *KMT2D*, LOF mutations in KMT2C/D complex subunit lysine (K)-specific demethylase 6A (*KDM6A*, also known as *UTX*) have also been shown to be associated with KS⁸⁰. While the mechanism by which KMT2D deficiency could cause KS remains unclear, the dysregulation of *HOX* family genes in KS is perhaps relevant in development of the disease, given the importance of *HOX* gene expression during mammalian development⁸¹ and the ability of KMT2D to regulate expression of *HOX* genes⁸², though this remains to be elucidated.

1.5.5 Incidence of *KMT2D* mutation in cancer

Early interest into the role of KMT2 family proteins in cancer can perhaps be attributed to historic discoveries of *KMT2A* rearrangements in mixed lineage leukemia (MLL)⁸³. In MLL tumors with KMT2A rearrangements, the methyltransferase function of *KMT2A* is replaced with that of it's many possible fusion partners⁸⁴, and results in dysregulation of genes normally regulated by KMT2A such as *HOX* family genes⁸⁵. *KMT2A/B/C/D* are all common targets of somatic alterations in cancer (Figure 1.1), with varying frequencies of mutations in *KMT2A/B/C/D* depending on the type of cancer⁸⁶. Mutations in *KMT2D* are most prevalent in forms of non-Hodgkin lymphoma (NHL), such as FL and DLBCL, in which 89% and 32% of

patients have KMT2D mutations, respectively⁸⁷. Other types of cancer in which KMT2D is frequently mutated include mantle cell lymphoma (MCL; 14% of patients⁸⁸), medulloblastoma (MED; 23%⁸⁹), nodal marginal zone lymphoma (NMZL; 25%⁹⁰) and breast fibroepithelial (BRFE; 13%⁹¹) tumors. To further investigate the frequencies of somatic KMT2D mutations across cancer types, I leveraged WES data from TCGA to comprehensively compile all somatic KMT2D mutations across 32 cancer types, in a method similar to that performed by Ryan Huff⁹² (Figure 1.2). This experiment revealed that *KMT2D* mutation also occurs at high prevalence (>=20% of samples) in skin cutaneous melanoma (SKCM; 26% of samples), esophageal carcinoma (ESCA; 26%), pancreatic adenocarcinoma (PAAD; 24%), stomach adenocarcinoma (STAD; 22%) and uterine corpus endometrial carcinoma (UCEC; 20%)¹¹. Moreover, across all cancer types with KMT2D-mutant patient samples, KMT2D mutations tend to be evenly distributed along the KMT2D locus and are often LOF mutations (Figure 1.2). Given the high frequency of KMT2D mutation across malignancies, developing insights into the consequences of KMT2D mutation may greatly benefit our ability to develop treatments applicable to a wide range of cancer types.

1.5.6 Consequences of *KMT2D* mutation

Early studies into the consequences of *KMT2D* mutation were focused on identifying genes with transcriptional patterns dependent on the presence of functional KMT2D^{93,94}. In 2007, Issaeva *et al.* compared transcriptomes (using microarrays for mRNA quantification) of wildtype and KMT2D knockdown (via shRNA) HeLa cell lines, revealing dysregulation of several genes involved in cellular adhesion/cytoskeletal organization (*LAMB3*, *LOXL1*, *GPR56*, *VTN* and *PCDH7*) and cell growth (*CRIP1* and *FHL1*)⁹³. Guo *et al.* performed a subsequent study in 2012,

using *KMT2D*-knockout (via Cre recombinase technology) HCT116 cell lines⁹⁴. Following RNA-seq, transcriptome comparisons between *KMT2D*-mutant and wildtype HCT116 revealed dysregulation of *S100A*-family genes (*S100A2/3/4/5/14/16*) and genes involved in diverse signaling pathways including cAMP signaling (*PDE12*, *PKIA/B*, *RPS6KA1*, *PPP3CA*), B-cell development (*IL7*, *IL7R*) and WNT signaling (*GJA1*, *SOX4/9*)⁹⁴. The number of KMT2D-dependent genes identified in both studies is low (4/318; 1.3%), which may reflect differences in methodology (microarray versus RNA-seq for mRNA profiling; HCT116 versus HeLa cell lines) or perhaps indicates context-specificity of KMT2D-dependent genes. Although lists of candidate KMT2D-dependent genes were produced by these two studies, the mechanism by which KMT2D loss could result in the transcriptional dysregulation of these genes remained unclear, as no histone marks, particularly H3K4me1, were profiled in these studies.

With the emergence of technology capable of profiling the genome-wide distributions of histone modifications, investigation into the epigenetic landscapes of *KMT2D*-mutant cells was logical considering the role of KMT2D as a methyltransferase. In 2013, Guo *et al.* demonstrated significant reduction of genome-wide H3K4me1 levels in *KMT2D*-knockout HCT116 cells using ChIP-sequencing technology⁷². The global loss of H3K4me1 in KMT2D-deficient cells has since been reproduced in multiple experimental contexts^{33,68–70,73}. A study by Lee *et al.* (2013)⁶⁸ demonstrated an association between *KMT2D* deletion and alterations in enhancer regions, where *KMT2D*-mutant mouse adipocytes showed defects in the H3K27ac-mediated activation of enhancers⁶⁸, indicating a role for KMT2D in H3K4me1-mediated enhancer priming⁴⁹. The authors also showed transcriptional dysregulation in *KMT2D*-mutant mouse adipocytes during differentiation, and went on to propose a model of enhancer-mediated transcriptional activation

of genes, which involves (1) transcription factor (TF) binding, (2) enhancer priming by KMT2D, (3) enhancer activation of H3K27 acetyltransferase p300 and (4) Pol II recruitment, promoter looping and transcriptional activation of target gene(s)⁶⁸. Enhancer alterations have also been identified in *KMT2D*-mutant cells by Wang *et al.* (2016)⁷⁰, who showed concomitant H3K4me1 and H3K27ac loss in 29% of active enhancers (defined by presence of H3K4me1/H3K27ac and absence of H3K4me3 in wildtype cells) in *KMT2D*-knockout ESCs, which were associated with dysregulation of cell-fate associated genes during differentiation⁷⁰. Given the possibility of context-specific roles of KMT2D, however, it has yet to be determined whether enhancer-mediated transcriptional alterations in *KMT2D*-mutant cells are reproducible in other cellular contexts.

Two recent studies performed in B cells have provided insight into the potential role of *KMT2D* mutation in B-cell lymphomas^{33,73}. Ortega *et al.* (2015)³³ demonstrated an association between KMT2D deficiency and delays in germinal center involution, impediment of B-cell differentiation and dysregulation of TS genes *TNFAIP3*, *SOCS3* and *TNFRSF14* in B cells³³. Zhang *et al.* (2015)⁷³ further demonstrated the ability for KMT2D deficiency to facilitate lymphomagenesis in B cells⁷³. Taken together, these studies provide a potential role for *KMT2D* mutation in lymphomas, although, while both demonstrate reductions in bulk H3K4me1 in *KMT2D*-deficient cells, the presence of enhancer alterations, and their relevance in lymphomagenesis, were not investigated, leaving a significant gap in the precise epigenetic mechanisms by which *KMT2D* mutation could contribute to tumorigenesis.

In 2015, an isogenic KMT2D-mutant model cell line system was generated by Ryan Huff using adherent human embryonic kidney 293A (HEK293A) cell lines⁹². KMT2D was inactivated using zinc finger nuclease (ZFN) technology targeted to exon 39 of KMT2D (Chr12: 49427414-49427455 (hg19 build)). Although ZFN technology can introduce off-target mutations⁹⁵, it is reasonable to assume that the isogenic KMT2D-mutant HEK293A cell lines contain far fewer mutations compared to the heterogeneous cancer cell lines in which KMT2D function has been previously studied^{72,93,94}. The KMT2D-mutant HEK293A model cell lines therefore represent an experimentally tractable system in which to study the consequences of KMT2D mutation, as transcriptional and epigenetic patterns detected in KMT2D-mutant cells are less likely to be confounded by mutations in genes other than KMT2D. In my thesis work, I seek to address several knowledge gaps in KMT2D biology that I previously alluded to throughout Chapter 1. To generate further knowledge in our understanding of the consequences of KMT2D mutation, I perform a comprehensive analysis of the transcriptome and epigenomic landscapes of KMT2Dmutant HEK293A cell lines. To determine whether the effect of KMT2D mutation on enhancer landscapes, as shown by previous studies, is context specific, I investigated how enhancers are altered in KMT2D-mutant HEK293A cell lines relative to wildtype cells. Finally, to provide insight into the epigenetic mechanism by which KMT2D mutation may contribute to oncogenesis, I positioned my findings regarding KMT2D-dependent epigenetic patterns in the context of pathways relevant to cancer.

1.6 Thesis roadmap and chapter summaries

The overarching hypotheses of this thesis were (a) that *KMT2D* mutation would be associated with alterations in the epigenetic and transcriptome landscapes, and (b) that epigenetic regions

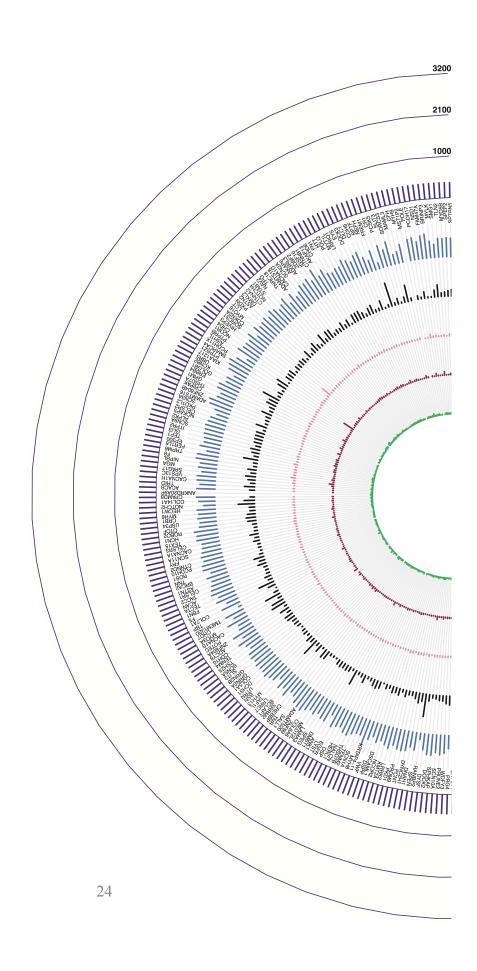
altered in *KMT2D*-mutant cells would be proximally associated with dysregulated genes. As such, my research aims encompassed characterization of KMT2D-dependent histone modifications and gene expression patterns in *KMT2D*-mutant cell lines, followed by integrative analyses to investigate the proximal relationship between these epigenetic and transcriptional alterations. Chapter 2 describes materials and methods used toward the experimental testing of my hypotheses. Chapter 3 then describes the specific aims used to address each of my hypotheses.

The research I describe in Chapter 3.1 pertains to the identification of epigenetic alterations associated with *KMT2D* mutation in an isogenic HEK293 model cell line system. Given the role of KMT2D in establishing H3K4me1 and the relationship between H3K4me1 and enhancer regions, I hypothesized that loss-of-function *KMT2D* mutation would be associated with (a) genome-wide decreases in H3K4me1 levels and (b) alterations of histone modifications present at enhancer regions in *KMT2D*-mutant cells. In this section, I identify a genome-wide decrease in H3K4me1 levels in *KMT2D*-mutant cells relative to wildtype, and provide a comprehensive profiling of both H3K4me1 peaks gained, lost or retained in *KMT2D*-mutant cells relative to wildtype. In addition, I also demonstrate significant changes in the presence of various histone modifications at enhancer regions in *KMT2D*-mutant cells, indicating an importance of KMT2D in the maintenance of both poised and active enhancers. These observations thus provided evidence consistent with the notion that *KMT2D* mutation is associated with alteration of the epigenetic landscape of isogenic HEK293A cell lines.

Research described in Chapter 3.2 was performed with the aim of determining whether *KMT2D* mutation was associated with transcriptional alterations in the HEK293 model system. To this end, I performed differential expression analysis to identify genes demonstrating changes in transcript abundance between *KMT2D*-mutant and wildtype cell lines. With this list of genes, I then performed comprehensive enrichment analyses to identify pathways common among KMT2D-dependent genes. I revealed enrichment of pathways associated with extracellular matrix organization and cell adhesion, which was consistent with the role of KMT2D in regulating the expression of genes involved in these pathways.

Finally, the aim of the work I describe in Chapter 3.3 was to integrate ChIP and RNA-sequencing data to explore associations between KMT2D-dependent enhancers and genes. I was able to identify a subset of KMT2D-dependent genes that were located in the vicinity of a KMT2D-dependent enhancer, which included the known KMT2D-target gene \$100A4\$. I noted that the majority of KMT2D-dependent enhancers were not in the vicinity of a KMT2D-dependent gene, and genes associated with KMT2D-dependent enhancers were not dysregulated in \$KMT2D\$-mutant cells. This result was consistent with the notion that alteration of KMT2D-dependent enhancers alone is not sufficient for invoking expression changes in the majority of nearby genes. Interestingly, genes associated with KMT2D-dependent enhancers were enriched for members of the TGF-beta and retinoic acid (RA) signaling networks, highlighting possible convergences between the regulatory axes of KMT2D and these pathways, mediated though KMT2D-dependent active enhancer regions.

A discussion of notable findings and ramifications of my research is included in Chapter 4. Overall, my research generates further knowledge regarding the consequences of *KMT2D* mutation. In addition, I also highlighted epigenetic alterations associated with *KMT2D* mutation in HEK293A cell lines, thereby providing an additional cell system in which KMT2D is required for maintaining histone modifications at enhancer regions. Finally, I provided insight into a potential mechanism by which KMT2D-dependent epigenetic dysregulation may contribute to oncogenesis, by highlighting deficient TGF-beta and RA signaling as potential consequences of loss of active enhancers in *KMT2D*-mutant cells.



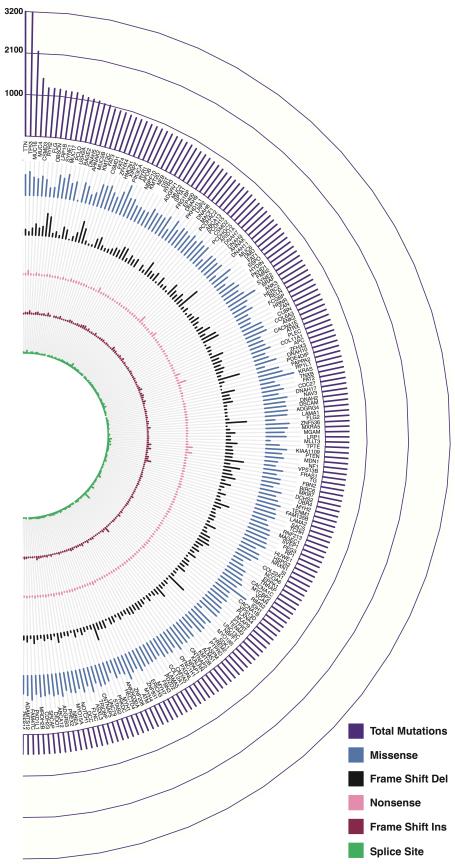


Figure 1.1 Top 375 most frequently mutated genes across 32 cancer types.

(This figure spans two pages). Somatic mutations were compiled from WES data of 8,366 cancer patients, encompassing 32 types of cancer (provided by TCGA). All types of somatic mutation assayed by the TCGA WES mutation-calling platform were considered in the total mutation count of each gene (see Methods). Genes were sorted in descending order by the total number of samples with at least one somatic mutations called in each gene, with the top 375 most frequently mutated genes shown. Purple bars (outer) indicate the total number of mutations across all samples. Each inner bar reflects the proportion of mutations belonging to each of five common mutation types (outwards to inwards): missense (blue), frameshift deletion (black), nonsense (pink), frameshift insertion (red) and splice site (green) mutations.

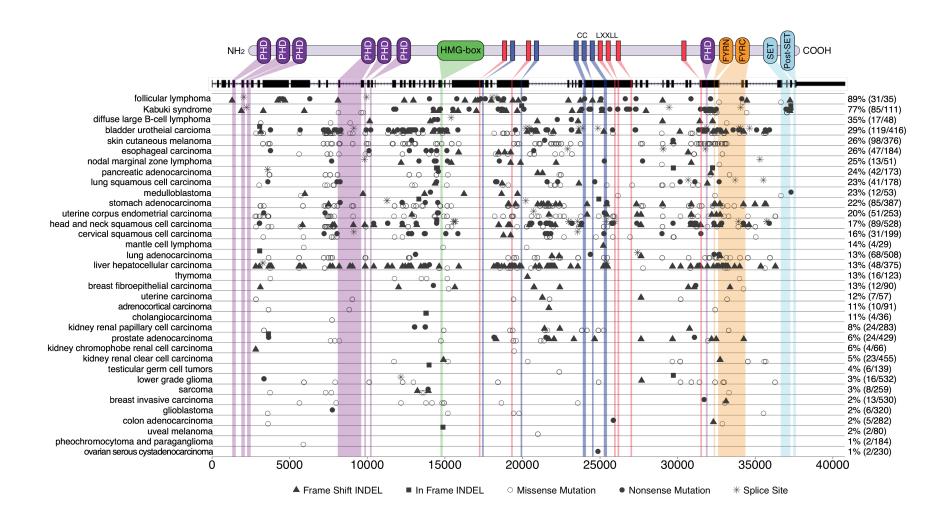


Figure 1.2 Frequency, distribution and types of *KMT2D* mutations across Kabuki syndrome and 34 cancer types.

The KMT2D protein structure is shown (top), together with the exon architecture of *KMT2D* obtained from the UCSC genome browser⁹⁶. Disease type acronyms are displayed on the left, while the percentage of cases with one of either in-frame/frameshift INDEL, missense, nonsense and splice site mutations shown on the right. The x-axis depicts the position, in bp, along the *KMT2D* gene, from 5' to 3' ends (left to right). This figured was adapted from a similar analysis performed by Ryan Huff⁹². Given that events other than SNVs could impact *KMT2D* function (such as copy number losses and DNA hypermethylation) in tumors, the proportion of samples with *KMT2D* alterations is likely larger than depicted here.

Chapter 2: Materials and methods

TCGA somatic mutation analysis

WES data (TCGA level 3) was collected from the TCGA data portal (https://tcga-data.nci.nih.gov/) for all available samples (n = 9,393) on November 6th, 2015. Entrez IDs were mapped to Ensembl gene IDs using the Ensembl BioMart database⁹⁷. Samples were mapped to cancer type using the available metadata from TCGA, accessed November 13th, 2015. Samples were filtered for those with (a) at least one mutation in a gene with a matched gene Ensembl ID, (b) a mapped cancer type given the available metadata, (c) processed by either HGSC, Broad, UCSC, BCGSC, WUSTL or M.D. Anderson, and (d) processed using reference genome build hg19. Application of these filtering criteria resulted in a final count of 8,510 samples across 32 cancer types. When identifying the most frequently mutated genes, all mutation types annotated by the TCGA mutation-calling platform were considered, which included missense, nonsense, nonstop, splice site, silent, RNA, translation start site, frameshift insertion/deletion, in-frame insertion/deletion, intergenic region (IGR), intron, 3' UTR, 5' flank and 5' UTR mutations. The number of cases in which a gene was found mutated was defined as the number of TCGA samples that had at least one mutation (of any type) in that gene.

Plotting

All plots were generated within R⁹⁸ v.3.2.0, using the ggplot2⁹⁹ v2.2.0 package for all bar, pie, scatter, line and boxplots, and the ggbio¹⁰⁰ v1.22.3 package for circular plots. Packages colorspace¹⁰¹ v1.3.2 and RColorBrewer¹⁰² v1.1.2 were used for selection of colors. ChIP-seq tracks were visualized using IGV¹⁰³ v2.3.72.

ChIP-sequencing library processing

ChIP-sequencing reads were aligned to human reference genome GRCh37-lite using BWA¹⁰⁴ v.0.7.15. BAM files were sorted and indexed using Samtools¹⁰⁵ v0.1.18. For peak calling, duplicated reads were flagged and removed using Picard v.1.114. Normalization of raw read coverage was performed by calculating reads per million (RPM) mapped reads. BIGWIG files were generated using BAM2WIG v1.0.0 (Bilenky *et al.*, unpublished) and UCSCs wigToBigWig¹⁰⁶.

Saturation analysis of ChIP-sequencing data

Saturation analysis involves multiple iterations in which the total number of sequenced reads are bootstrapped at different proportions. At each iteration, peaks are called using the bootstrapped subsample of reads, and the number of peaks are counted. I performed saturation analysis using the built-in function provided by FindER (-fractionS parameter), for fractions of the total number of reads equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9, 0.925, 0.95, 0.975 and 0.999 (for a total of 20 iterations). The number of peaks called at each iteration is sensitive to the peak widths. When comparing two peak calling results, one result could contain fewer peaks despite having the same number of base pairs covered by peaks, if small peaks located close to one another are merged. For this reason, I chose to evaluate each iteration of my saturation analysis by the number of base pairs covered by peaks, rather than the total number of peaks. When very few peaks are able to be identified (for example, at iterations including a small proportion of sequenced reads), the number of base pairs covered by peaks can be sensitive to peaks with significance values very close to the default threshold for peak significance (q < 0.05), resulting in very jagged saturation curves. For this reason, I filtered the

peaks I was considering for those of relatively high significance, by selecting a threshold significance threshold based on the distribution significance levels for all peaks detected across all iterations of the saturation analysis. This threshold was determined to be $q = 7.9e^{-6}$, based on the distribution of all q values.

ChIP-sequencing analysis

Genomic regions significantly enriched for read coverage from ChIP-seq libraries were determined using FindER v.1.0.0b (Bilenky *et al.*, unpublished) with default parameters (min. read quality = 5, FDR = 0.05, lower limit for PET fragment length = 75, upper limit = 1028, min. size of enriched region = 300, max. gap size for which neighbouring regions are merged = 100). Information regarding FindER is available through the Canadian Epigenetics, Environment and Health Research Consortium Network website¹⁰⁷. Raw read coverage within genomic regions as well as overlap between genomic regions were calculated using Bedtools¹⁰⁸ v2.17.0.

H3K4me1 analysis

For *KMT2D*-mutant and wildtype HEK293 cell lines, H3K4me1 peaks within each library were merged if they were located within < 450bp of each other, and peaks smaller than 600bp were excluded from further analysis, as per Pellacani *et al.* (2016)¹⁰⁹. H3K4me1 peaks were assessed for overlap with peaks of each of the five other histone marks in wildtype HEK293A by considering two peaks to be overlapping if they shared at least 150 bp of overlap¹⁰⁹. Annotated TSS regions were obtained from Refseq (February 2, 2017), and the distance between TSSs and H3K4me1 peaks was determined using Bedops¹¹⁰ v2.4.20.

Enhancer analysis

Active enhancers were defined as H3K4me1 peaks that overlapped (150 bp minimum) with an H3K27ac peak but not an H3K27me3 peak¹⁰⁹. Similarly, poised enhancers were defined as H3K4me1 peaks overlapping with H3K27me3 peaks but not H3K27ac peaks¹⁰⁹. For both poised and active enhancers, those overlapping with an H3K4me3 peak were considered to be promoters and excluded from further analysis, along with active enhancers overlapping with the heterochromatin mark H3K9me3.

Motif analysis

DNA sequence motif discovery was performed individually on H3K4me1 peaks retained, gained and lost in the *KMT2D*-mutant cell lines with respect to wildtype using Homer2¹¹¹ v.4.9.1, with a region size (-s) of 1000 (as recommended for histone methylation peaks in the Homer documentation (http://homer.ucsd.edu/homer/ngs/peakMotifs.html) and otherwise default parameters (min. mer size = 6, max. mer size = 10, find motifs of length = 10, gaps = 0, number of seeds = 10).

RNA-sequencing

RNA-sequencing reads were aligned to human reference genome GRCh37-lite using JAGuaR¹¹² v.2.03. Feature counts (number of reads mapping to each gene) were generated using Subread¹¹³ v.1.4.6-p5, using default parameters: meta-feature level, paired-end = true, strand specific = false, multi-mapping reads = not counted, multi-overlapping reads = not counted, chimeric reads = not counted, and both ends mapped = required.

Differential Expression Analysis

Ensembl IDs were mapped to gene IDs using the Ensembl BioMart database⁹⁷. Differential expression analysis was conducted on raw read counts using DESeq2¹¹⁴ v.1.8.2 in R v.3.2.0, using ~condition as the design formula (where ~condition specified *KMT2D* mutation status (knockout/wildtype)). This resulted in a list containing all genes assayed, ranked by their adjusted p-value calculated by DESeq2.

Gene Set Enrichment Analyses

Throughout my thesis, I refer to gene set enrichment analysis (GSEA) as the general method by which pre-determined sets of genes are investigated in terms of being over-represented among an input list of genes, rather than the software named "GSEA" developed by the Broad Institute 115. The molecular signatures database (mSigDB) provides an extensive collection of annotated gene sets for which to test for enrichment among a set of genes 116. GSEA was performed on the 74 genes with TSSs that gained H3K4me1 and H3K27ac in *KMT2D*-mutant cell lines (along with presence of H3K4me3 in wildtype and *KMT2D*-mutant cell lines) using all 13,361 gene sets provided by mSigDB, using hypergeometric tests followed by Benjamini-Hochberg (BH) multiple test correction (all within R v.3.2.0).

GSEA was performed separately on up-regulated (adjusted p < 0.05 and fold change > 1.5) and down-regulated (adj. p < 0.05 and fold change < -1.5) genes, using the same methodology as described above. Up and down-regulated genes were separately tested for enrichment of (a) all mSigDB gene sets (13,361), (b) GO terms (3,562) and (c) KEGG pathways (148), each downloaded from the mSigDB website (software.broadinstitute.org/gsea/msigdb/). Multiple

hypothesis correction (MHC) of p values was performed individually for each of the three collections of gene sets, for up-regulated and down-regulated genes, using BH multiple test correction.

GSEA was performed on genes associated with KMT2D-dependent enhancers with the same methodology described above, but using a filtered list of gene sets. Out of the 4,731 curated gene sets available from mSigDB (accessed May 10th, 2017), sets were filtered using keywords "BOUND", "TARGET" and "RESPONSE", to produce a final list of 896 gene sets, in an attempt to focus more on gene sets pertaining to specific pathways and transcriptional coregulators.

Identification of genes associated with enhancers

The Genomic Regions Enrichment of Annotations Tool (GREAT¹¹⁷) was used to identify genes associated with KMT2D-dependent poised and active enhancers, using human genome assembly GRCh37 and the "basal plus extension" association rule (default) with a maximum distance of 100 kb. 100 kb was chosen as the maximum distance at which a TSS could be considered to be associated with an enhancer due to results of a previous study showing that this window captures the majority of true enhancer-promoter interactions⁴⁸.

Chapter 3: Results

3.1 KMT2D mutation is associated with alterations in the epigenetic regulatory landscape

The role of KMT2D in histone modification is attributed to the presence of the catalytic SET domain within KMT2D that confers intrinsic methyltransferase activity⁶⁰. Mono-methylation of lysine (K)-4 of histone H3 (H3K4me1) is a histone modification found throughout the mammalian genome, and is known to be associated with distal *cis*-regulatory regions⁴⁹. Previous studies have demonstrated an association between KMT2D deficiency and decreased H3K4me1 in multiple contexts, highlighting KMT2D as a major H3K4 mono-methyltransferase in mammalian cells^{33,68,70,72,73}. Enhancers represent *cis* regulatory regions capable of controlling gene expression, and can be categorized into functionally unique states based on the presence of different histone modifications, including primed (H3K4me1 only), active (H3K4me1 and H3K27ac) and poised (H3K4me1 and H3K27me3)⁴⁹. Previous studies have identified a role of KMT2D in commissioning active enhancers, with loss of KMT2D being associated with decreases in the number of active enhancers^{68,70}. To further explore the association between KMT2D mutation and epigenetic alterations, I analyzed ChIP-sequencing data from KMT2D wildtype and mutant HEK293 cell lines, which included enhancer marks H3K4me1, H3K27ac and H3K27me3. To further characterize the effect of KMT2D mutation on histone modification, I also incorporated into my analysis H3K9me3 and H3K4me3 marks, which are associated with closed chromatin regions and gene promoters, respectively.

3.1.1 Genome-wide profiling of H3K4me1 enrichment supports global loss of H3K4me1 in KMT2D-deficient HEK293 cells

Global H3K4me1 levels measured using western blots indicated the presence of KMT2D-dependent H3K4 mono-methylation in the same HEK293A cell lines I was performing my analyses on 92, supporting the notion that KMT2D plays a major role establishing H3K4me1 in this context. To further investigate H3K4me1 depletion in KMT2D-deficient HEK293 cells, and to determine genomic regions in which H3K4me1 is consistently lost upon *KMT2D* mutation, I characterized genome-wide changes in H3K4me1 enrichment across wildtype and *KMT2D*-mutant HEK293 cell lines.

As with other sequencing technologies, localized enrichment of ChIP-seq reads can be biased based on DNA sequence and accessibility, resulting in a decreased signal-to-noise ratio 118. In an effort to deconvolute signal from noise, genomic regions with localized enrichment of sequencing reads, relative to input control, can be determined from ChIP-seq data 119. Identification of such regions, referred to as *peaks*, represents a common first step in ChIP-sequencing analysis pipelines. For this reason, I used a peak-calling algorithm to identify peaks (see Methods) in each ChIP-sequencing library analyzed (15 in total, including H3K4me1/3, H3K27ac/me3 and H3K9me3 from each of the cell lines). FindER (Bilenky *et al.*, unpublished) was chosen as the peak-calling algorithm due to the broad range of histone modifications included in my analysis, as the algorithm provides a mechanism to identify enrichment of histone modifications that show localized profiles (such as H3K4me3), dispersed profiles (H3K27me3), or a combination of both (H3K4me1) 107. The five histone modifications included in my analyses were chosen on the basis of having either direct association with the known function of KMT2D

(H3K4me1) or known association with enhancer regions (H3K27ac/me3), while promoter and heterochromatin-related marks (H3K4me3 and H3K9me3, respectively) were included to further filter for true enhancer regions (for instance, active enhancer regions should not contain either H3K4me3 nor H3K9me3⁴⁹).

Sequencing depth (defined as the total number of mapped sequencing reads) can represent a limiting factor in regards to the accuracy of ChIP-seq experiments when considering the number of true binding sites able to be detected. This can be explained by the fact that the number of reads sequenced required to capture the majority of binding sites should theoretically increase as the number of true binding sites increases¹¹⁸. To demonstrate the degree of saturation in regards to the number of peaks called given the depth of sequencing of the ChIP-seq data I analyzed, I performed a saturation analysis (see Methods). The total number of peaks identified in a ChIPseq library is dependent on the distribution of peak sizes, and, when used as a metric to compare libraries, can indicate comparatively fewer peaks for a library (for instance, if small peaks nearby one another are merged) even though the total area of the genome covered by peaks is larger. For this reason, I used the total number of bp covered by peaks as the dependent variable in the saturation analysis (represented on the y-axis of Figure 3.1). Furthermore, rather than considering all significant (q < 0.05) peaks, I filtered peaks for those of high significance, by using a threshold q-value of 7.9e⁻⁶ (chosen on the basis of representing a point between the bimodal distribution of q-values for all peaks). In this way, I focused my analysis on the saturation of peaks most likely to be true positives, rather than including peaks that may have only marginally passed the default significance threshold of q = 0.05. For each of the 15 ChIP-seq libraries (H3K4me1/3, H3K27ac/me3 and H3K9me3 across wildtype and KMT2D-mutant cell

lines D320/D372), saturation of the genomic coverage of significant peaks was achieved, and is indicated by the plateau observed as the proportion of reads used to call peaks approaches 100% (Figure 3.1). Given that the ChIP-seq data I was analyzing demonstrated saturation of peak discovery given the read depths at which each cell line was sequenced, I next performed analyses to identify differences in the global number of histone marks between *KMT2D*-mutant and wildtype HEK293 cell lines.

To investigate KMT2D-dependent changes in the global levels of each histone mark, I compared the number of peaks identified in KMT2D-mutant (D320 and D372) and wildtype cell lines. To ensure that changes in the number of peaks represented a global increase/decrease in the number of peaks, rather than a shift in the distribution of peak widths (for instance, if many small peaks were merged to form large aggregated peaks in one library), I also calculated the total number of base pairs (bp) of the genome covered by all peaks for each library. Data from both KMT2Dmutant cell lines showed decreases in the total number of H3K4me1 peaks, with only 94,452 (D320) and 112,398 (D372) peaks identified compared to 144,171 in wildtype (34.5% and 22.0% loss, respectively; Figure 3.2). Furthermore, the number of bp covered by H3K4me1 peaks was reduced in both KMT2D-mutant cell lines, with 5.78x10⁷ (D320) and 6.29x10⁷ (D372) bp covered compared to 9.09×10^7 in wildtype (36.4% and 30.1% loss, respectively). Taken together, global decreases in both the number of H3K4me1 peaks and the total number of bp covered by peaks were consistent with the notion that KMT2D loss is associated with a global decrease of H3K4me1. In contrast, none of the other histone modifications analyzed showed a consistent change in both mutant lines relative to wildtype (Figure 3.2).

3.1.2 A subset of H3K4me1 peaks detected in *KMT2D*-wildtype cells are consistently lost in *KMT2D*-mutant cells

Decreases in the genome-wide occupancy of H3K4me1 peaks in both KMT2D-mutant cell lines were compatible with the notion that there exist KMT2D-dependent H3K4me1 peaks in HEK293 cells. To investigate KMT2D-associated changes in H3K4me1 distribution in more detail, I next compared the presence/absence of H3K4me1 peaks in each KMT2D-wildtype/mutant cell line. The union of all H3K4me1 peaks (identified in at least one of the cell lines) was identified, and encompassed a total of 232,820 peaks. These peaks were then filtered using criteria as described by others¹⁰⁹ (see Methods). In brief, peaks within 450 bp of another were merged, and peaks smaller than 600 bp in size were excluded. This resulted in a total of 72,587 H3K4me1 peaks, which were then investigated in terms of presence/absence in each of the three samples (Figure 3.3A). 11,535/72,587 (15.9%) peaks were present in wildtype HEK293 cell lines and lost in both KMT2D-mutant cell lines, indicating that a proportion of KMT2D-dependent loss of H3K4me1 occurs at the same genomic regions in both of the KMT2D-mutant cell lines. 25,243/72,587 peaks (34.8%) were unchanged between wildtype and both KMT2D-mutant cell lines, supporting the notion that a subset of H3K4me1 can be maintained by a separate methyltransferase (presumably KMT2C) in the absence of KMT2D^{33,68,72,73}. Interestingly, 8,570/72,587 peaks (11.8%) were gained in both KMT2D-mutant cell lines relative to wildtype, indicating that KMT2D mutation is also associated with a partial redistribution of H3K4me1. The gain of H3K4me1 in KMT2D-mutant cells has also been observed in a recent study involving KMT2D knockout in mouse ESCs⁷¹, although characterization of H3K4me1 peaks gained in the absence of KMT2D was not performed by the authors of this study. Finally, the remaining 27,239/72,587 (37.5%) H3K4me1 peaks did not show consistent occupancies in the KMT2D-mutant cell lines

(either gained or lost in one mutant cell line (relative to wildtype) but not both). As these H3K4me1 peaks are more likely to represent differences between the mutant cell lines, rather than KMT2D-dependent alterations, these H3K4me1 peaks were excluded from future analyses.

To further characterize H3K4me1 peaks lost, retained, or gained in both KMT2D-mutant cells relative to KMT2D-wildtype cells, I assessed normalized read coverage profiles, chromosome distributions, distances to nearby genes and enriched sequence motifs for each of the three groups of peaks. Normalized coverage profiles are used for averaging signal across a group of peaks in order to visualize the average shape and density of peaks. To provide a visualization of the groups of H3K4me1 peaks lost, retained or gained in KMT2D-mutant cells, I constructed coverage profiles using normalized H3K4me1 ChIP-seq signal, demonstrating average changes in each group of peaks between KMT2D-mutant and wildtype cell lines (Figure 3.3B). The distribution of H3K4me1 peaks (in terms of the number of peaks per bp of chromosome) was relatively uniform across all chromosomes for H3K4me1 peaks retained or lost in KMT2Dmutant cells, while H3K4me1 peaks gained in KMT2D-mutant cells showed a bias toward chromosomes 17 and 19 (Figure 3.3C). The H3K4me1 peaks gained on chromosomes 17 and 19 were uniformly distributed across each chromosome (that is, there was no localized aggregation of H3K4me1 peaks gained on either chromosome 17 or 19). The distance between each H3K4me1 peak and its nearest TSS was calculated, and the distribution of distances was assessed for each group of KMT2D-dependent H3K4me1 peaks. The distribution of distances to nearest TSS regions showed a tendency for all H3K4me1 peaks (retained, gained and lost) to reside 10-50 kb from TSSs, indicating that the H3K4me1 peaks being investigated were located in the vicinity of genes (Figure 3.3D). DNA sequence motif analysis of H3K4me1 peaks (see

Methods) showed significant enrichment (adj. $p = 1e^{-214}$) of motifs corresponding to binding sites of transcription factors FOS and JUN (activator protein (AP)-1 complex members) for H3K4me1 peaks retained in KMT2D-mutant cells (Figure 3.3E). Interestingly, enrichment of AP-1 DNA binding motifs was even more significant (adj. $p = 1e^{-280}$) among the H3K4me1 peaks lost in KMT2D-mutant cells, indicating that the AP-1 complex may interact with the KMT2Dcontaining complex at areas of the genome H3K4-monomethylated by KMT2D. Furthermore, previous studies demonstrating overlap between AP-1 complex binding and enhancer regions are consistent with the notion that AP-1 functions as a co-regulator of enhancer-mediated regulation of gene expression 120,121. Finally, H3K4me1 peaks that were gained in KMT2D-mutant cells did not show significant (p < 0.05) enrichment of AP-1/FOS/JUN binding motifs, and instead showed enrichment of various proteins including transcriptional regulator CTCF (adj. $p = 1e^{-53}$). Interestingly, CTCF motifs were not significantly (p < 0.05) enriched among the H3K4me1 peaks lost in KMT2D-mutant cells, and showed relatively lower enrichment (p = $1e^{-41}$) among the H3K4me1 peaks retained in KMT2D-mutant cells, indicating that the potential overlap between H3K4me1 and CTCF activity is unique to the H3K4me1 gained in the absence of KMT2D.

To further characterize H3K4me1 changes associated with *KMT2D* mutation, I continued my analyses with a focus on the regions of H3K4me1 gained, lost or retained in *KMT2D*-mutant cells. Given that effects of KMT2D deficiency on H3K4me1 have been shown to localize to enhancer regions rather than promoters⁶⁸, I next investigated the context in which KMT2D-dependent H3K4me1 peaks are located, in terms of promoter and enhancer-related marks that co-occupy each region in wildtype cells.

3.1.3 KMT2D-dependent H3K4me1 loss occurs at regions co-occupied by H3K27ac but not H3K4me3

Areas of the genome can be categorized into functionally distinct states based on different combinations of histone modifications. For example, active enhancer regions are marked with H3K4me1 and H3K27ac⁴⁹. To further characterize KMT2D-dependent H3K4me1 changes, I investigated the context in which H3K4me1 changes occurred, in terms of which histone modifications were present at these regions in wildtype cells. This was performed for H3K4me1 peaks found to be gained, lost or retained in both KMT2D-mutant cell lines with respect to wildtype cells. 3,090/8,570 (36.1%) of the KMT2D-dependent H3K4me1 peaks gained in both KMT2D-mutant cell lines overlapped with at least one of H3K27ac/me3 or H3K4me3 in the wildtype cell line, while this was true for 1,244/11,535 (10.8%) of the KMT2D-dependent H3K4me1 peaks lost in both mutant cell lines and 4,233/25,243 (16.8%) of the H3K4me1 peaks that were unchanged between KMT2D-mutant and wildtype cell lines (Figure 3.4A). The fact that over 60% of H3K4me1 peaks (either gained, lost or retained) did not overlap with any of the enhancer and promoter marks in wildtype cells aligns with other studies that have shown the majority of enhancers reside in the primed state (H3K4me1 only)^{74,122}. H3K4me1 peaks gained in the KMT2D-mutant cell lines were found to predominantly occur at promoter-like regions in the wildtype cells, as the majority (65.5%) of peaks shared overlap with H3K4me3. In contrast, H3K4me1 peaks lost in the KMT2D-mutant cell lines were found to predominantly overlap with active enhancer mark H3K27ac (756/1,244; 60.1%) or poised enhancer mark H3K27me3 (279/1,244; 22.4%), while only 285/1,244 (22.9%) overlapped with promoter mark H3K4me3 (Figure 3.4B). Taken together, these results suggest that KMT2D-dependent gain of H3K4me1

takes place in promoter regions in wildtype cells, while KMT2D-dependent loss of H3K4me1 occurs at enhancer regions.

The gain of H3K4me1 in the absence of KMT2D has only been documented in one recent study, in which mouse ESCs were found to gain 9,228 H3K4me1 peaks upon knockout of KMT2D⁷¹. While the gain of H3K4me1 in absence of KMT2D was mentioned, the authors of this study did not perform further analyses to profile these regions. I therefore sought to further characterize the regions gaining H3K4me1 in the absence of KMT2D, by investigating their overlap with TSS regions as well as KMT2D-dependent changes in the occupancy of overlapping H3K4me3 and H3K27ac peaks (see Methods). 1,506/8,570 (17.6%) H3K4me1 peaks gained in the absence of KMT2D overlapped with a TSS region, while 74/1,056 (4.9%) of these peaks also showed gain of the active promoter mark H3K27ac in both KMT2D-mutant cell lines (Figure 3.5). This result indicated that loss of KMT2D might have a positive impact on the transcription of some genes through KMT2D-independent gain of H3K4me1 (and subsequent gain of the activating mark H3K27ac) at their TSSs. To explore the function of the 74 genes that gain H3K4me1 and H3K27ac at their TSS regions in KMT2D-mutant cell lines, I performed a gene set enrichment analysis (GSEA) using a comprehensive list of available gene sets (see Methods). The 74 genes were found to be significantly (adj. p value < 0.05) enriched for a group of genes observed to be down-regulated in vascular smooth muscle cells (VSMC) by MAPK8/JNK1 by a previous study¹²³, and included SSR4, TOR1AIP1, CTTN, INSIG1, SLC6A6, GATA6, AIMP2 and STX4 (Figure 3.5).

The fact that the majority of KMT2D-dependent H3K4me1 peaks lost in *KMT2D*-mutant cells converged onto wildtype regions with enhancer marks indicated that the enhancer landscape in HEK293 cell lines is altered as a consequence of *KMT2D* mutation, and is consistent with the association of KMT2D with enhancers as shown by others^{68,70,72}. To explore the effect of *KMT2D* mutation on enhancer landscapes in further detail, I next profiled active and poised enhancer regions in wildtype cells and assessed how the distribution of various histone marks, including H3K4me1, were altered at enhancer regions in *KMT2D*-mutant cells.

3.1.4 KMT2D deficiency is associated with a genome-wide decrease in the number of active enhancers

To comprehensively investigate the effect of *KMT2D* mutation on the enhancer landscape in HEK293 cell lines, I first profiled active (containing H3K4me1 and H3K27ac) and poised (containing H3K4me1 and H3K27me3) enhancer regions in the wildtype sample, using criterion described by others¹⁰⁹ (see Methods). Briefly, H3K4me1 peaks were assessed for >150 bp overlap with H3K27ac and H3K27me3. H3K4me1 peaks overlapping with H3K27ac but not H3K27me3 were classified as active enhancers, while those overlapping with H3K27me3 but not H3K27ac were classified as poised enhancers. Lastly, H3K4me1 peaks overlapping with both H3K27ac and H3K27me3 were excluded from further analysis. Application of this strategy to classify enhancer regions resulted in the identification of 586 poised enhancers and 2,229 active enhancers (shown in Figure 3.6A).

Enhancer regions are capable of regulating gene expression in *cis* through long distance interactions with TSSs, with the majority of interactions taking place within a distance < 100 kb

(median distance of ~15 kb)⁴⁸. While multiple TSSs may exist within this 100 kb window, 40% of enhancers have been shown to regulate their nearest TSS¹²⁴. To determine what proportion of enhancer regions identified in wildtype HEK293 cell lines (based on the presence/absence of histone modifications) had a nearest TSS within a 100 kb window, I investigated the distribution of distances between each enhancer region and its nearest TSS. 563/586 (96.1%) poised enhancers and 2,129/2,229 (95.6%) active enhancers were located at least 1 kb away from their nearest TSSs (Figure 3.6B). Moreover, 195/586 (33.3%) and 835/2,229 (37.5%) poised and active enhancers (respectively) were within 10-50 kb of their nearest gene, in agreement with results of previous studies regarding enhancer-TSS distances^{48,124}.

To explore how these wildtype poised/active enhancer regions were altered in *KMT2D*-mutant cells, I next compared the distributions of the H3K4me1 and H3K27ac/me3 enhancer marks between wildtype and *KMT2D*-mutant cell lines. Consistent with my previous results demonstrating loss of H3K4me1 at enhancer regions, 185/586 (31.6%) and 627/2,229 (28.1%) of poised and active enhancers (respectively) showed consistent loss of H3K4me1 in both *KMT2D*-mutant cell lines relative to wildtype. Loss of H3K4me1 was coupled with loss of H3K27ac in 526/2,229 (23.6%) of active enhancers (Figure 3.7), which was consistent with the notion that KMT2D-dependent H3K4me1 may be necessary for establishment of H3K27ac at a subset of enhancers. A similar effect was also observed for a subset (133/586; 22.7%) of poised enhancers, in which loss of H3K4me1 was coupled with loss of H3K27me3. Interestingly, I observed loss of H3K27ac at 625/2,229 (28.0%) active enhancers that retained H3K4me1, indicating that, in absence of functional KMT2D, presence of H3K4me1 is insufficient for establishing H3K27ac at a subset of active enhancers. This was also seen for H3K27me3 at

111/586 (18.9%) poised enhancers, which demonstrated loss of H3K27me3 despite retaining H3K4me1. Coverage profiles of active and poised enhancers showing consistent loss of only H3K4me1 or both H3K4me1 and H3K27ac (active enhancers) or H3K27me3 (poised enhancers) are depicted in Figure 3.8.

Overall, *KMT2D* mutation was found to be associated with consistent changes (occurring in both *KMT2D*-mutant cell lines) in at least one of H3K4me1, H3K27ac or H3K27me3 in 189/586 (32.3%) of wildtype poised enhancers and 798/2,229 (35.8%) of active enhancers, indicating that *KMT2D* mutation can remodel the enhancer landscape of HEK293 cell lines. Given the widespread remodeling of enhancer regions in *KMT2D*-mutant cells, and the role of enhancers in transcriptional regulation, I next sought to investigate transcriptome alterations associated with *KMT2D* mutation through analysis of RNA-sequencing data.

3.2 KMT2D mutation is associated with transcriptional alterations

Previous studies of transcriptome alterations associated with *KMT2D* mutation have raised the possibility of a context-specific role for KMT2D in the regulation of a diverse set of biological pathways. Such pathways include cell adhesion and structural organization in the colorectal cancer cell line HCT116^{72,94}, fat/muscle cell differentiation in mouse embryos⁶⁸, and pathways relevant to cancer (such as the cell cycle and p53 signaling) in follicular lymphoma samples³³. To assess the impact of *KMT2D* mutation on the transcriptome, I compared the transcriptomes of *KMT2D*-mutant and wildtype HEK293 cell lines by analyzing RNA-sequencing data that was generated from the same wildtype and two *KMT2D*-mutant cell lines that I profiled using ChIP-sequencing analysis (described in Chapter 3.1).

3.2.1 KMT2D mutation is associated with decreased transcript abundance

To identify gene expression alterations associated with KMT2D mutation, I conducted a differential expression analysis (DEA) to compare mRNA transcript abundances between wildtype and KMT2D-mutant cell lines (see Methods). 265 genes were differentially abundant/expressed (DE; adjusted p < 0.05 and absolute fold change (FC) > 1.5; Appendix A) in *KMT2D*-mutant samples, 220 (83.0%) of which were down-regulated (FC < -1.5; Figure 3.9). Genes showing the largest changes in mRNA abundance in KMT2D-mutant cells included members of diverse families such as bone morphogenic protein 2 (BMP2; FC = -36.8), zinc finger protein 532 (ZNF532; FC = -39.4), poly(ADP-ribose) polymerase family member 14 (PARP14; FC = -13.9), solute carrier family 2 member 3 (SLC2A3; FC = -9.8) and glycogenin 2 (GYG2; FC = 13.0). 45/265 (17%) of genes were up-regulated, and included secreted frizzled related protein 2 (SFRP2; FC = 4.9), progesterone receptor (PGR; FC = 3.7), solute carrier family 35 member F1 (SLC35F1; FC = 2.6), brain-specific serine/threonine kinase 2 (BRSK2; FC = 2.3) and protocadherin 10 (*PCDH10*; FC = 2.6). In addition, known KMT2D-target S100family genes S100A16 and S100A494 were both significantly down-regulated in KMT2D-mutant HEK293 cell lines (adjusted p = $5.3e^{-13}$ and $7.7e^{-4}$; FC = -2.2 and -1.3, respectively), along with laminin subunit beta 3 (LAMB3) (adjusted p = 2.2e-6, FC = -3.11). To expand the scope of my transcriptome analysis from individual genes to molecular pathways, I next performed a GSEA to test for enrichment of a comprehensive list of pathways among the up and down-regulated genes.

3.2.2 KMT2D-dependent genes converge on pathways related to structural organization and cellular adhesion

To test for enrichment of various gene sets among the list of genes DE in KMT2D-mutant HEK293 cell lines, I implemented a high-throughput pipeline to perform self-contained hypergeometric tests. This approach allowed me to test for over-representation of annotated gene sets among up and down-regulated genes separately (see Methods). To comprehensively test for enrichment of a large variety of gene sets, I performed GSEA using 13,361 gene sets provided by the molecular signatures data base (mSigDB) as well as 3,562 GO terms and 148 KEGG pathways. Up-regulated genes were significantly (adj. p < 0.05) enriched for 5 mSigDB pathways (Appendix B) including genes found to be up-regulated upon stimulation of vascular endothelial growth factor (VEGFA)¹²⁵, genes up-regulated in mouse T-reg cells induced with ovalbumin¹²⁶, genes with TSSs containing two motifs with unknown functions (AACTTT and YNTTTNNNANGCARM) as well as genes with TSSs containing the CCAATAATCGAT motif, which matches annotation for transcriptional repressor cut-like 1 (CUTL1). These up-regulated genes with TSS motifs matching CUTL1 binding sites were FGF12, EGR1, ST6GALNAC5 and TAC1. Up-regulated genes were not significantly enriched for any GO terms or KEGG pathways. Down-regulated genes were significantly (adj. p < 0.05) enriched for 982 mSigDB gene sets (Appendix C), with many of the top enriched gene sets related to extracellular matrix (ECM) organization, such as genes encoding ECM-associated proteins 127 and genes involved in epithelial-mesenchymal transition (Figure 3.10). Other significantly enriched mSigDB gene sets included genes up-regulated in fibroblasts upon knockdown of retinoblastoma-1 (RB1)¹²⁸ and genes in the mSigDB "cancer" and "lung" modules. Down-regulated genes were also significantly enriched for 313 GO terms (Appendix D), with the top-most enriched GO terms

also converging on ECM and adhesion pathways. Finally, down-regulated genes were enriched for two KEGG pathways (Appendix E): ECM receptor interaction and focal adhesion. The reoccurrence of ECM and adhesion-related gene sets among enrichment analyses using orthogonal collections of gene sets (mSigDB, GO and KEGG) provides evidence supporting the role of KMT2D in regulating genes involved in ECM structure and organization in HEK293 cell lines, and is supported by several previous studies of KMT2D-dependent transcriptome alteration ^{93,94}.

To investigate whether the KMT2D-dependent epigenomic alterations (Chapter 3.1) were associated with the transcriptomic alterations identified in *KMT2D*-mutant cell lines, I next sought to combine both the ChIP-seq and RNA-seq data in an integrative analysis.

3.3 Integrative analysis of the KMT2D regulome reveals convergence between KMT2D-dependent enhancers and the TGF-beta and RA signaling networks

The process of identifying direct associations between genes and distal regulatory elements (such as enhancer regions) is experimentally challenging, with contemporary molecular techniques such as tethered chromatin capture (TCC) and ChIA-PET experimentally challenging ¹²⁹. Given the coordinates of enhancer regions and gene TSSs (compiled from ChIP and RNA-seq data, respectively), deriving an association between the two data types often relies on assumptions regarding the distances at which enhancers are thought to interact with TSSs. Such strategies include relating enhancer regions to the locations of genes of interest ¹³⁰, with a proximity of 100 kb being accepted as a threshold distance at which the majority of enhancers regulate their target(s) ⁴⁸. Within a 100 kb distance, the gene(s) most likely to be regulated by an enhancer

(enhancer associated genes) can be identified using local measurements¹⁰⁹, such as the Genomic Regions Enrichment of Annotations Tool (GREAT)¹¹⁷, which considers the positions of nearby gene regulatory domains, such as TSSs.

3.3.1 A subset of KMT2D-dependent genes are proximally associated with KMT2D-dependent enhancers

KMT2D mutation was found to be associated with epigenetic alteration in poised and active enhancer regions (Chapter 3.1), as well as transcriptional changes in 262 genes (Chapter 3.2). To investigate whether KMT2D-dependent epigenetic and transcriptome alterations were proximally associated with one another, I calculated the distance between KMT2D-dependent genes and KMT2D-dependent enhancers. 156/220 (71%) down-regulated genes and 32/45 (71%) of upregulated genes had an annotated Refseq TSS region and were included in this analysis, and a distance of 100 kb was used as a threshold for associating a gene with an enhancer⁴⁸. 32/156 (21%) of down-regulated genes were within 100 kb of a KMT2D-dependent active enhancer, indicating that alterations in a subset of KMT2D-dependent active enhancers are likely associated with significant (adj. p < 0.05) alterations in the transcription of a nearby gene (Figure 3.11A). This was not observed for KMT2D-dependent poised enhancer regions, in which only 9/156 (5.8%) of down-regulated genes were within 100 kb of a KMT2D-dependent poised enhancer. Proximal association between KMT2D-dependent enhancers and DE genes was also not observed for up-regulated genes, as only 4/32 (12%) and 1/32 (3.1%) of up-regulated genes were within a 100 kb distance of a KMT2D-dependent active or poised enhancer, respectively. The observation that many DE genes were not in the vicinity (< 100 kb distance) of KMT2Ddependent enhancers could perhaps be explained by (a) long-distance (> 100 kb) enhancer-gene

interactions taking place or (b) that *KMT2D* mutation causes transcriptional alterations through a mechanism independent of enhancer alteration. In the latter case, it may be that KMT2D binds the TSSs of a subset of KMT2D-dependent genes, and that the loss of this interaction in *KMT2D*-mutant cells is sufficient to alter transcript abundance. Due to the large size of the KMT2D protein (532 kDa), ChIP-seq of KMT2D is experimentally challenging, and was not able to be performed on the *KMT2D*-mutant and wildtype HEK293 cell lines during the time of my thesis.

The role of KMT2D in regulating the S100A-family of genes was revealed by Guo *et al.* $(2012)^{94}$, who demonstrated decreases in the expression of S100A2/3/4/5/14/16 in KMT2D-mutant HCT116 cells⁹⁴, although the mechanism by which KMT2D loss may have facilitated down-regulation of S100A genes was not explored. In the data I analyzed, S100A4 was found to be significantly (adj. p < 0.05) down-regulated in KMT2D-mutant cells, and had two nearby (< 100 kb) KMT2D-dependent active enhancers, which lost both H3K4me1 and H3K27ac in KMT2D-mutant cell lines (Figure 3.11C). This result indicated that perhaps loss of H3K4me1 and H3K27ac at active enhancer regions causes decreases in S100A4 transcription in KMT2D-mutant cells.

The observation that many of the KMT2D-dependent genes were not within the threshold distance (100 kb) of KMT2D-dependent enhancers indicated that epigenetic enhancer alterations may not be sufficient for invoking significant (adj. p < 0.05) changes in expression of nearby genes in *KMT2D*-mutant cells. To investigate what the potential consequences of epigenetic

alterations in KMT2D-dependent enhancers might be, I turned my investigation to genes proximally associated with KMT2D-dependent enhancers.

3.3.2 Alteration of KMT2D-dependent enhancers alone is not sufficient for modulating transcription of nearby genes

To further characterize KMT2D-dependent enhancers in terms of their relationship with genes DE in KMT2D-mutant cells, I first determined the proportion of KMT2D-dependent enhancers within 100 kb of a DE gene. I also included 1,151 enhancer regions with coordinates randomly permuted across the genome as a control in this analysis. 43/1,151 (3.7%) and 4/1,151 (0.35%) of KMT2D-dependent active enhancers were within 100 kb of a down or up-regulated gene, respectively, consistent with the notion that alterations in the majority of KMT2D-dependent active enhancers are not associated with nearby transcriptional alterations large enough to surpass significance thresholds used in DEA (adj. p < 0.05). This was also true for KMT2Ddependent poised enhancers, of which 9/244 (3.7%) and 1/244 (0.41%) were within 100 kb of a down or up-regulated gene, respectively (Figure 3.12A). While DE genes generally were not in the vicinity of KMT2D-dependent enhancers, it was possible that genes regulated by KMT2Ddependent enhancers were dysregulated in KMT2D-mutant cells but did not meet the DEA significance threshold. For this reason, I shifted the focus of my investigation from DE genes to genes proximally associated with KMT2D-dependent enhancers (and therefore likely to be regulated by these regions¹¹⁷).

The reliance on the assumption that enhancers regulate their nearest TSS has proven successful in several previous studies^{131,132}. However, this assumption has several limitations such as the

tendency to overlook true binding sites¹¹⁷ and introduce a bias toward large TSS regions (in which larger regions have a higher likelihood of being nearby any element by random chance ¹³³). To overcome these limitations, tools such as the GREAT algorithm¹¹⁷ are often implemented to identify genes associated with regulatory regions by leveraging both predictive and experimentally determined regulatory domains. For this reason, I utilized GREAT to identify genes proximally associated with KMT2D-dependent enhancers (see Methods). 1,049 and 226 genes were associated with KMT2D-dependent active and poised enhancers, respectively. To determine the extent to which these genes were dysregulated in KMT2D-mutant cells, I calculated the fold change of each gene (KMT2D-mutant versus wildtype) and compared fold change distributions between DE genes, genes associated with KMT2D-dependent active and poised enhancers, and all other genes (Figure 3.12B). 883/1,049 (84.2%) and 190/226 (84.1%) of genes associated with KMT2D-dependent active and poised enhancers (respectively) were assayed in the RNA-seq data, and were included in this analysis. While the fold change distributions of genes associated with KMT2D-dependent active and poised enhancers were significantly greater compared to all genes ($p = 1.2e^{-11}$ and $1.7e^{-4}$, respectively), this effect was very small compared to that of DE genes ($p = 6.6e^{-172}$), indicating that alteration of KMT2Ddependent enhancers alone is generally insufficient for inducing transcriptional dysregulation of nearby genes. This observation is consistent with results published in a recent study, which demonstrated that KMT2D is dispensable for maintaining expression levels of adjacent genes to KMT2D-dependent enhancers in mouse ESCs⁷⁰. The authors went on to show that genes adjacent to KMT2D-dependent enhancers were dysregulated only during differentiation, suggesting that KMT2D-dependent enhancer alteration resulted in dysregulation of adjacent genes only in the presence of other developmental regulatory factors/processes. I therefore

hypothesized that genes associated with KMT2D-dependent enhancers required a specific context or stimuli in order to exhibit dysregulation in *KMT2D*-mutant HEK293 cell lines.

To provide insight into what specific stimuli may be required in order to observe the transcriptional consequences of KMT2D-dependent enhancer alteration, I performed GSEA on the genes associated with KMT2D-dependent active and poised enhancers using 896 curated gene sets from mSigDB (Figure 3.12C). Interestingly, genes associated with KMT2D-dependent active enhancers were significantly enriched for targets of the retinoic acid (RA) signaling pathway, including targets of retinoic acid receptor gamma (RARG; adj. $p = 7.6e^{-4}$) and PML-RARA fusion protein (adj. $p = 2.7e^{-3}$). This result is consistent with the notion of the KMT2D regulome converging on genes that are involved in RA signaling 92,94, and indicates that perhaps during response to RA treatment, RA signaling genes associated with KMT2D-dependent active enhancers would be dysregulated in KMT2D-mutant cells compared to wildtype. In addition to the RA signaling network, genes associated with KMT2D-dependent active enhancers were also enriched for gene sets related to TGF-beta signaling, such as SMAD2/3 and TGFB1 targets (adj. $p = 2.1e^{-6}$ and $4.7e^{-4}$, respectively). Genes associated with KMT2D-dependent poised enhancers were only significantly enriched for four gene sets, which included TP53, TCF21, ESRRA and MLL-AF4 targets (adj. p = 0.026, 0.028, 0.028 and 0.028, respectively).

3.3.3 Genes associated with KMT2D-dependent active enhancers are enriched for genes up-regulated upon RA treatment

To further characterize the extent of overlap between the KMT2D regulome and the RA signaling network, I performed a second GSEA on genes associated with KMT2D-dependent

active enhancers, using 29 gene sets related to RA signaling. Interestingly, genes associated with KMT2D-dependent active enhancers were enriched for genes up-regulated in response to RA treatment but not genes that are down-regulated upon RA treatment (Figure 3.13A), indicating that KMT2D-dependent active enhancers function to induce transcription of these RA signaling genes in the presence of RA. I therefore hypothesized that RA target genes associated with KMT2D-dependent active enhancers are expressed at relatively low levels in wildtype HEK293 cell lines, which would enable their expression levels to be increased upon RA treatment. To address this hypothesis, normalized expression levels (RPKM) were calculated for each gene, and genes were separated into three groups based on whether they were RA target genes associated with a KMT2D-dependent active enhancer (n = 244), housekeeping genes (n = 295) assumed to be maintained at relatively high expression levels, and all other genes (n = 16,642; Figure 3.13B). While RA target genes associated with KMT2D-dependent active enhancers were expressed at higher levels compared to all other genes ($p = 1.5e^{-45}$), the difference was not as great compared to that of housekeeping genes ($p = 4.3e^{-138}$), indicating that RA target genes associated with KMT2D-dependent active enhancers are maintained at low levels, such that they may be poised for increased transcription upon RA treatment.

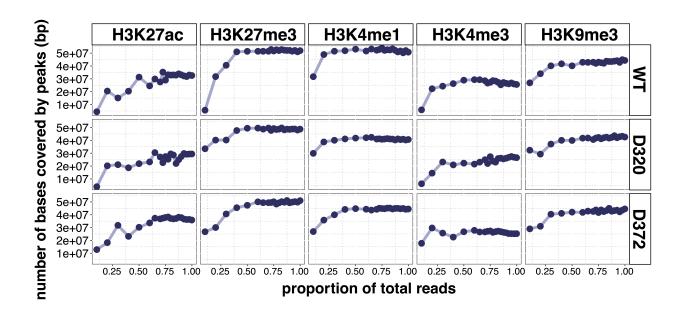


Figure 3.1 Genomic coverage of peaks achieves saturation for each ChIP-seq library at their respective sequencing depths.

Results of saturation analysis performed using each of the five ChIP-seq libraries (left to right). The number of bases covered by highly significant peaks (negative $\log_{10} q$ value > 5.1) saturates as the proportion of reads sampled approaches the total number of reads sequenced for each library, as seen by the plateau in the curve.

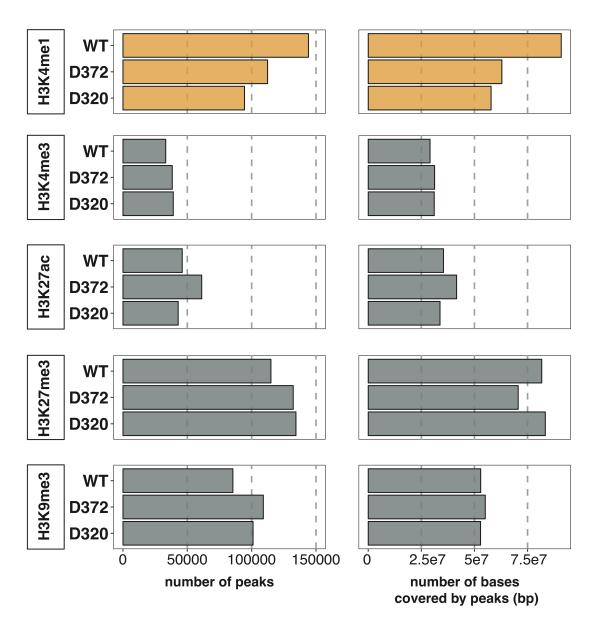


Figure 3.2 KMT2D mutation is associated with a global loss of H3K4me1.

Number of significant (q < 0.05) ChIP-seq peaks (left) and number of bases covered by significant peaks (right) is shown for H3K4me1/3, H3K27ac/me3 and H3K9me3 libraries for *KMT2D*-wildtype (WT) and mutant (D372, D320) HEK293A cell lines. *KMT2D*-mutant cell lines show a decrease in both the number of significant peaks detected and the number of bases covered by such peaks, and this extent of loss is not seen for the other histone marks assayed.

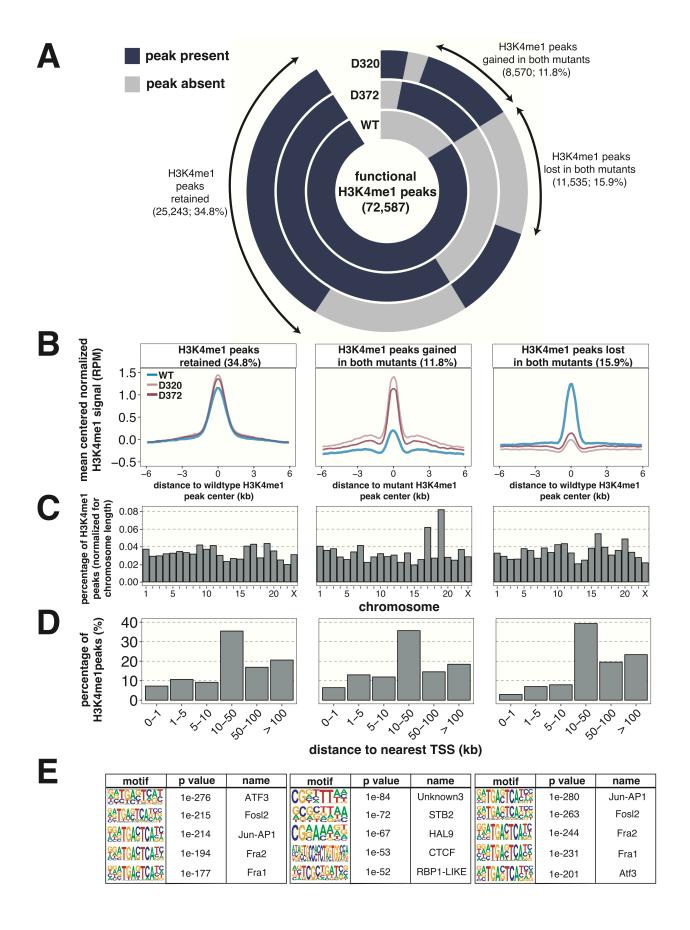
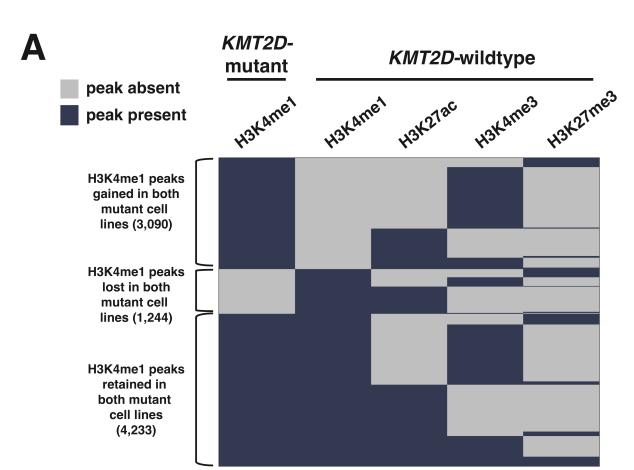
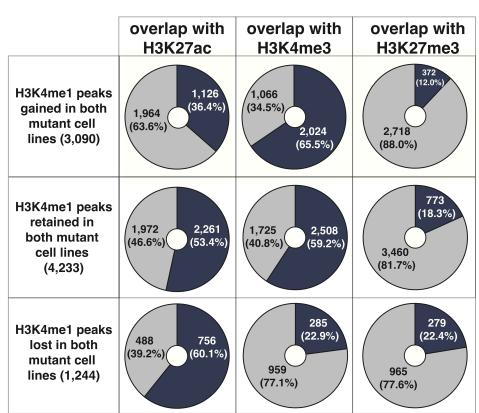


Figure 3.3 Regions that gain or lose H3K4me1 in *KMT2D*-mutant cells are enriched for CTCF and AP-1 DNA binding motifs, respectively.

(A) The union of all H3K4me1 peaks (n = 72,587) detected in wildtype and KMT2D-mutant HEK293 cell lines (D320 and D372) were assessed for presence/absence in each cell line. Peaks are represented radially, with black indicating presence of a peak that overlaps with the region of the original peak by at least 150 bp, while grey indicates that this criteria was not met. 11,535 (15.9%) H3K4me1 peaks were present in the wildtype cell line and lost in both KMT2D-mutant cell lines, while 8,570 (11.8%) peaks were absent in the wildtype cell line and gained in both KMT2D-mutant cell lines and 25,243 (34.8%) peaks remained unchanged (present in wildtype and both mutant cell lines). (B) Mean-centered normalized (RPM) ChIP-seq H3K4me1 coverage profiles in wildtype, D320 and D372, for H3K4me1 peaks that were either retained (left), gained (middle) or lost (right) in both KMT2D-mutant cell lines relative to wildtype. Coverage is centered on wildtype H3K4me1 peak centers when peaks were retained or lost, and centered on D320 H3K4me1 peak centers when peaks were gained. (C) Chromosomal distribution of H3K4me1 peaks either retained, gained or lost in KMT2D-mutant cells. The percentage (%) of peaks, normalized for chromosome size, is depicted on the y-axis. (D) Distribution of distances of each H3K4me1 peak to the nearest TSS. (E) Top five significantly enriched (adjusted p < 0.05) motifs for H3K4me1 peaks that were either retained, gained or lost in KMT2D-mutant cells.





B

Figure 3.4 H3K4me1 peaks gained or lost in *KMT2D*-mutant cell lines occur predominantly at promoter and enhancer regions, respectively.

(A) H3K4me1 peaks that were present in the *KMT2D*-wildtype cell line and either gained, lost or retained in both *KMT2D*-mutant cell lines were assessed for overlap with H3K27ac, H3K27me3 and H3K4me3 peaks in wildtype cells. H3K4me1 peaks that overlapped with at least one other histone mark in the wildtype cell line are shown. (B) H3K4me1 peaks gained in *KMT2D*-mutant cell lines often co-occur with the promoter mark H3K4me3 (overlapping with 59.2% of peaks). In contrast, H3K4me1 peaks lost in *KMT2D*-mutant cell lines often co-occur with enhancer mark H3K27ac (overlapping with 60.1% of peaks), while only 22.9% of peaks overlap with promoter mark H3K4me3. H3K4me1 peaks retained in *KMT2D*-mutant cell lines predominantly co-occur with H3K4me3 (overlapping with 59.2% of peaks). These results are compatible with the notion that gain of H3K4me1 in *KMT2D*-mutant cells tends to take place predominantly at promoter regions, while loss of H3K4me1 in *KMT2D*-mutant cells tends to take place predominantly at enhancer regions.

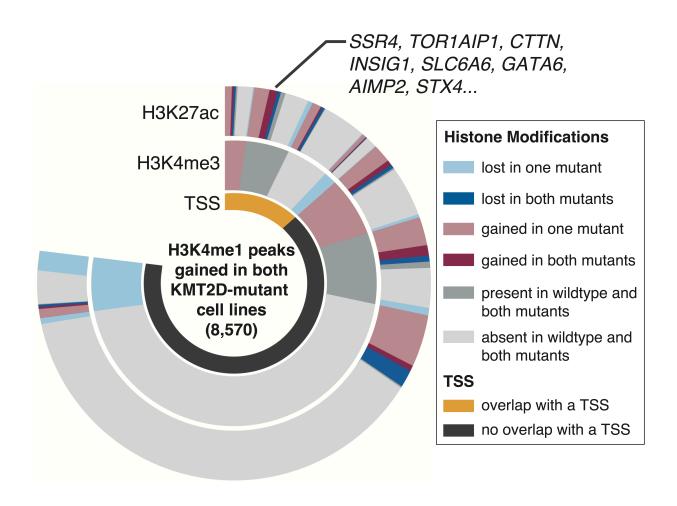


Figure 3.5 A subset of H3K4me1 peaks gained in *KMT2D*-mutant cell lines overlap with TSS regions.

H3K4me1 peaks gained in *KMT2D*-mutant cell lines were characterized in terms of their overlap with TSS regions (inner ring) and H3K4me3/H3K27ac peaks in *KMT2D*-mutant and wildtype cell lines. H3K4me1 peaks are represented radially. The inner track corresponds to whether each peak shared overlap with a TSS region (orange). The two outer tracks correspond to the changes in overlapping histone modifications H3K4me3 and H3K27ac in *KMT2D*-mutant cell lines with respect to wildtype. While a large proportion (4,212/8,570; 49.1%) of H3K4me1 peaks did not overlap with a TSS and showed no overlap with H3K4me3/H3K27ac in either *KMT2D*-mutant or wildtype cell lines (corresponding to the region of black inner ring and light-gray outer rings),

74 H3K4me1 peaks overlapped with a TSS, along with a H3K4me3 peak which was unchanged in *KMT2D*-mutant cell lines, and gained H3K27ac in both *KMT2D*-mutant cell lines. 8/74 (10.8%) of these genes were candidate targets of the MAPK8/JNK1 signaling pathway, and are shown at the top.

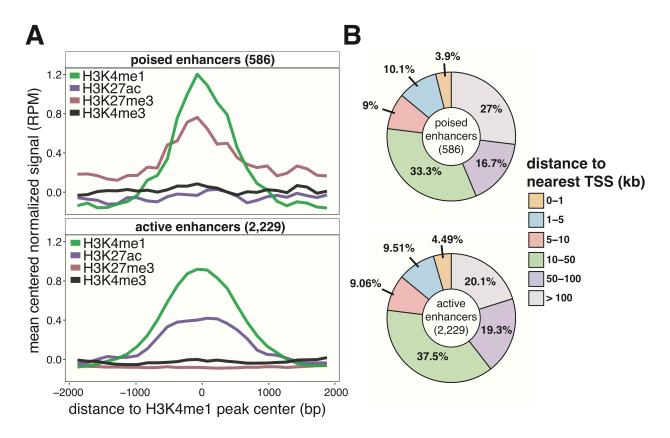


Figure 3.6 Poised and active enhancers detected in wildtype HEK293 cells are distal to promoter regions.

(A) Mean-centered normalized coverage (RPM) of ChIP-seq libraries H3K4me1/3 and H3K27ac/me3 in the *KMT2D*-wildtype cell line, centered on H3K4me1 peaks. Coverage profiles are shown for 586 poised enhancers (H3K4me1 and H3K27me3) and 2,229 active enhancers (H3K4me1 and H3K27ac). (B) Poised and active enhancers were mapped to their nearest TSS, with the majority (69.1% and 75.4%, respectively) located between 1-100 kb from their nearest TSS.

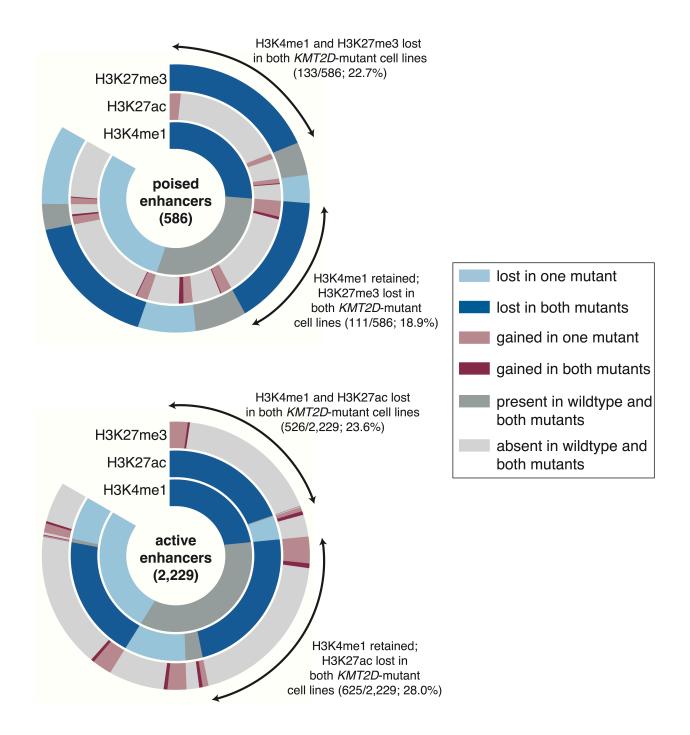


Figure 3.7 Mutation of *KMT2D* is associated with loss of histone modifications H3K4me1 and H3K27me3/ac at poised and active enhancers, respectively.

Poised and active enhancer regions identified in the *KMT2D*-wildtype cell line were assessed for changes in histone modifications in *KMT2D*-mutant cell lines. Enhancers are represented radially, with circular tracks representing overlap (> 150 bp) of H3K4me1, H3K27ac and H3K27me3 with each enhancer. Track colors represent histone mark changes in the two *KMT2D*-mutant cell lines relative to wildtype. Both poised and active enhancers showed loss of H3K4me1 peaks in both *KMT2D*-mutant cell lines (31.6% and 28.1%, respectively). *KMT2D* mutation was associated with loss of both H3K4me1 and H3K27me3 in 133/586 (22.7%) poised enhancers, while 526/2,229 (23.6%) active enhancers lost both H3K4me1 and H3K27ac. These results were compatible with the notion that *KMT2D* mutation is associated with alterations in the enhancer landscapes of *KMT2D*-mutant cells. Interestingly, loss of H3K27me3/ac occurred at a subset of poised and active enhancers that retained H3K4me1 in the *KMT2D*-mutant cell lines, indicating that presence of functional KMT2D, rather than H3K4me1, is required for establishment of H3K27me3/ac at poised and active enhancers (respectively).

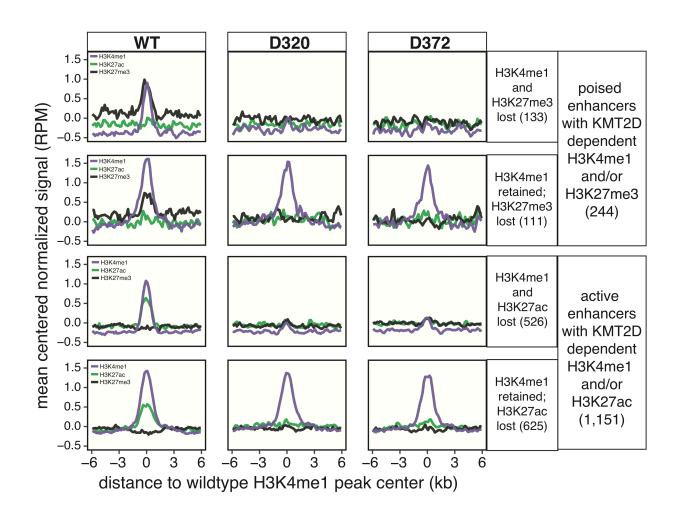


Figure 3.8 A subset of KMT2D-dependent poised and active enhancers lose H3K27me3/ac (respectively), despite retaining H3K4me1, in *KMT2D*-mutant cells.

Mean-centered normalized signal (RPM) is shown for ChIP-seq libraries H3K4me1 and H3K27ac/me3, centered at wildtype H3K4me1 peaks, for *KMT2D*-mutant (D320, D372) and wildtype cell lines. 111/244 (45.5%) poised enhancers showed loss of H3K27me3 peaks despite retaining the overlapping H3K4me1 peak. A similar result was observed for active enhancers, as 625/1,151 (54.3%) active enhancers lost H3K27ac despite retaining H3K4me1.

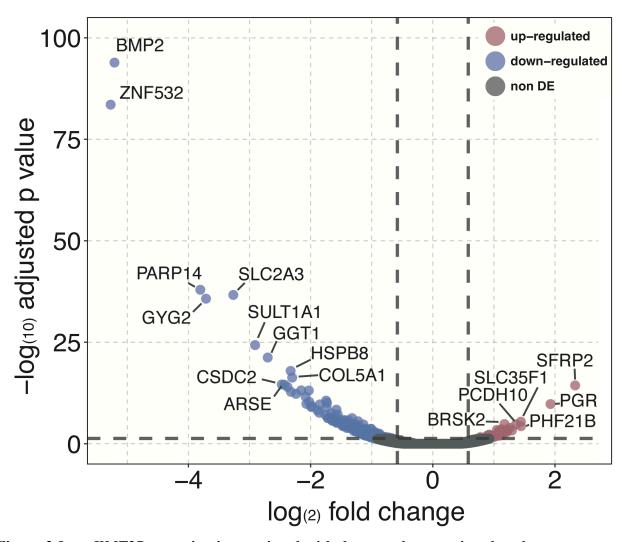


Figure 3.9 KMT2D mutation is associated with decreased transcript abundance.

Differential expression analysis was performed to compare mRNA levels between KMT2D-wildtype (n = 1) and mutant (n = 2) HEK293 cell lines. Log₂ fold change (x-axis) and negative log₁₀ adjusted p value (y-axis) is shown for each gene assayed, with down-regulated (adjusted p < 0.05 and fold change < -1.5) and up-regulated (adjusted p < 0.05 and fold change > 1.5) genes shown in blue and red, respectively. 220/265 (83.0%) genes were down-regulated while only 45/265 (17.0%) genes were up-regulated, indicating significantly decreased transcript abundance associated with KMT2D mutation.

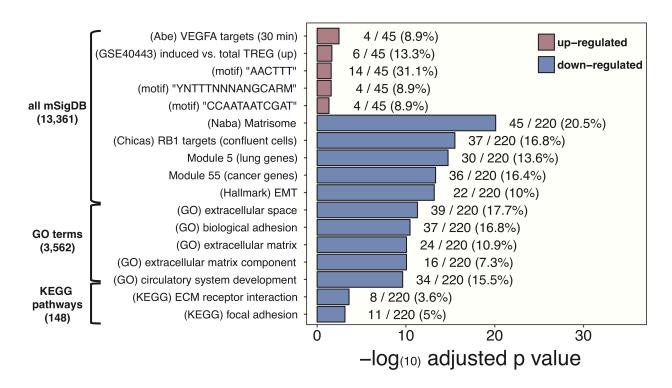
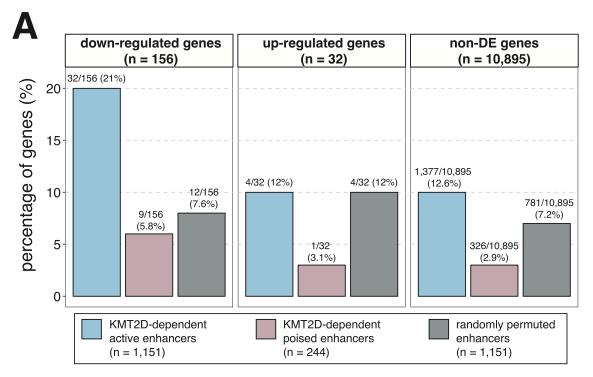


Figure 3.10 KMT2D-dependent genes are enriched for pathways related to cytoskeletal organization and cellular adhesion.

The 220 and 45 down (blue) and up (red) regulated genes (respectively) identified as DE in *KMT2D*-mutant HEK293 cell lines were both subjected to three separate enrichment analyses to test for enrichment of 13,361 mSigDB gene sets, 3,562 GO terms and 148 KEGG pathways. Negative log₁₀ adjusted p values are shown for the top most significantly enriched gene sets for each of the three groups of gene sets. Shown to the right of each bar is the number of up and down-regulated genes that belong in each gene set. Consistent with the notion of a role for KMT2D in the positive regulation of genes involved in extracellular matrix organization, top enriched gene sets for down-regulated genes included matrisome, EMT, adhesion and ECM-related pathways.



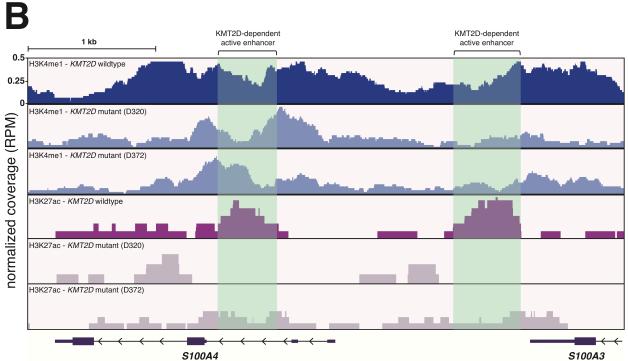


Figure 3.11 A subset of DE genes are within 100 kb of a KMT2D-dependent enhancer.

(A) Proportion of genes within 100 kb of KMT2D-dependent AEs (blue) PEs (red) and randomly permuted enhancers (gray). 21% of down-regulated genes were located within 100 kb of a KMT2D-dependent active enhancer, indicating that loss of a nearby enhancer could be responsible for the down-regulation of some genes in *KMT2D*-mutant cells. (B) Normalized ChIP-seq read coverage (RPM) of H3K4me1 and H3K27ac, showing reduction of H3K4me1 and H3K27ac peaks at two active enhancer regions in close proximity to the gene *S100A4*, which was found to be significantly (adj. p < 0.05) down-regulated in *KMT2D*-mutant HEK293 cell lines.

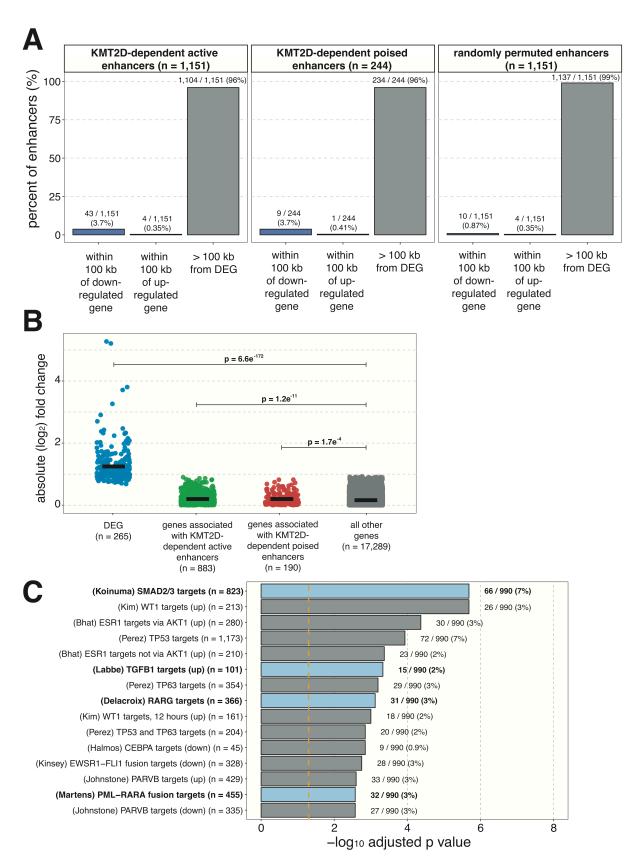
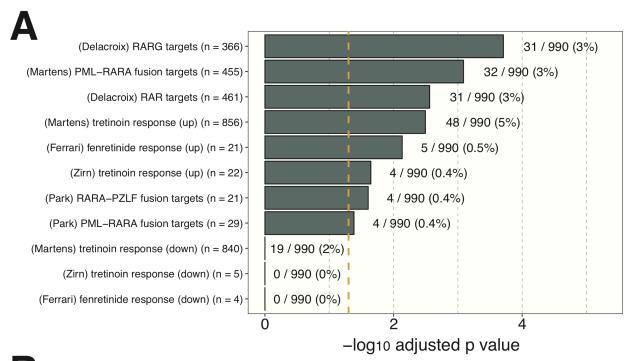


Figure 3.12 Genes associated with KMT2D-dependent active enhancers include targets of of the TGF-beta and retinoic acid signaling pathways.

(A) The majority of KMT2D-dependent active (96%) and poised (96%) enhancers were not located within 100 kb of a KMT2D-dependent DE gene. (B) Absolute fold changes (*KMT2D*-mutant versus wildtype) of genes DE (adj. p < 0.05; blue), associated with KMT2D-dependent active (green) and poised (red) enhancers, and all other genes (gray). Genes associated with KMT2D-dependent enhancers (identified using GREAT) were not DE in *KMT2D*-mutant cells. P values were determined using Wilcoxon tests followed by BH multiple hypothesis correction. (C) Genes associated with KMT2D-dependent enhancers were significantly enriched for targets of SMAD2/3 (adj. p = $2.1e^{-6}$), TGF-beta signaling (adj. p = $4.7e^{-4}$), retinoic acid receptor gamma (RARG; adj. p = $7.6e^{-4}$) and the PML-RARA fusion protein (adj. p = $2.7e^{-3}$). P values were determined using hypergeometric tests followed by BH multiple hypothesis correction. The



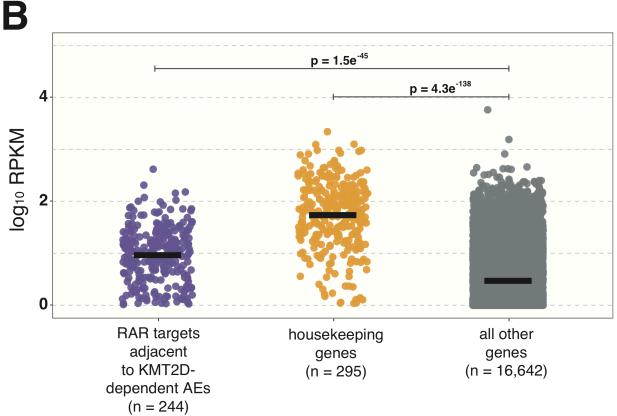


Figure 3.13 RA target genes associated with KMT2D-dependent active enhancers are poised for transcriptional activation.

(A) RA signaling gene sets significantly enriched (adj. p < 0.05) among genes associated with KMT2D-dependent active enhancers correspond to genes known to be positively regulated by RA treatment, while genes negatively regulated by RA treatment (the three bottom-most gene sets) are not significantly enriched (adj. p > 0.05). The dashed orange vertical line indicates the threshold of significance (adj. p = 0.05). (B) mRNA levels of RA target genes (purple), housekeeping genes (orange) and all other genes (gray) in *KMT2D*-wildtype cells. mRNA levels of RA target genes associated with KMT2D-dependent active enhancers are observed to be relatively low, indicating that they could be increased upon RA treatment.

Chapter 4: Discussion

4.1 KMT2D mutation is associated with alterations in the epigenetic regulatory landscape Frequent occurrence of KMT2D LOF mutations in cancer, together with the role of KMT2D as a histone methyltransferase, provide a rationale for the detailed investigation of epigenetic alterations associated with KMT2D loss. Previous studies have demonstrated an association between KMT2D deficiency and global loss of H3K4me1 in several distinct cell types^{33,68–70,72,73}. I was also able to demonstrate this association, using KMT2D-mutant HEK293A cell lines, through the analyses discussed in sections 3.1.1 and 3.1.2. Most notably, 15.9% of H3K4me1 peaks were lost in both KMT2D-mutant cell lines relative to wildtype. Interestingly, a proportion (11.8%) of H3K4me1 peaks were present in both KMT2D-mutant cell lines but not detected in wildtype cells, indicating that KMT2D deficiency is also associated with new establishment of H3K4me1 in some areas of the genome. This observation is consistent with results of a recent study, which demonstrated gain of H3K4me1 peaks in mouse ESCs that had both KMT2C and KMT2D knocked out⁷¹. Since the gain of H3K4me1 was observed in cells that lacked function of both canonical H3K4me1 methyltransferases (KMT2C/D), it remains unclear how de novo H3K4me1 could be established in these cells. While motif analysis revealed high levels of enrichment of STB2, HAL9, CTCF and RBP1-LIKE binding sites among H3K4me1 peaks gained, these proteins do not posses domains with known methyltransferase capability.

The observation that H3K4me1 peaks lost in *KMT2D*-mutant cell lines have a greater tendency to overlap with enhancer-related histone marks (H3K27ac/me3), rather than the promoter mark H3K4me3, aligns with previous studies demonstrating a preference for KMT2D to bind to distal

elements rather than promoters^{68,70,71}. Upon more detailed investigation of enhancer regions in *KMT2D*-mutant and wildtype cell lines, the abrogation of H3K4me1 and H3K27ac at 23.6% of active enhancer regions in *KMT2D*-mutant cells was consistent with results of a previous study (demonstrating loss of H3K4me1 and H3K27ac at 29% of active enhancers⁷⁰). Interestingly, the loss of H3K27ac (independent of H3K4me1) at active enhancer regions in *KMT2D*-mutant HEK293 cells was more pronounced compared to results described by Wang *et al.* (2016)⁷⁰. The results described in Section 3.1.4 indicate loss of H3K27ac in both *KMT2D*-mutant HEK293 cell lines at 74.6% of all active enhancers, whereas Wang and colleagues described the 71% of active enhancers that did not lose both H3K4me1 and H3K27ac to have "exhibited little change in H3K27ac in DKO cells". Loss of H3K27ac at active enhancers that retained H3K4me1 in *KMT2D*-mutant cell lines indicates that presence of H3K4me1 alone is not sufficient for H3K27 acetylation, and that perhaps functional KMT2D plays a role in the recruitment of the necessary H3K27 acetylase. This observation is supported by the study performed by Wang and colleagues, who reported KMT2D-dependent binding of H3K27 acetylase EP300 to enhancers⁷⁰.

The loss of H3K27ac observed in *KMT2D*-mutant HEK293 cells could perhaps be explained by the decreased protein levels of CREBBP in the same *KMT2D*-mutant HEK293 cell lines, as shown by Ryan Huff⁹². CREBBP proteins levels do not appear to have been assayed by Wang and colleagues, who instead focused their analyses on EP300. To provide insight into the possibility of one acetyltransferase being more highly expressed (relative to the other) in HEK293 cell lines and mouse ESCs, I compared mRNA levels of the two canonical acetylases (CREBBP and EP300) in the HEK293 cell line data used in this thesis and the mouse ESCs used by Wang and colleagues (GSE50534). RNA-seq data from the MLL3^{-/-} MLL4^{f/f} mouse ESCs

showed higher *EP300* expression compared to *CREBBP* (5,094 reads versus 1956 reads, respectively), while the RNA-seq data from wildtype HEK293A show the opposite, with *CREBBP* expression over four-fold higher than *EP300* (44,291 reads versus 9,966 reads, respectively), indicating that perhaps CREBBP is the dominant H3K27 acetylase in HEK293 while EP300 plays that role in mouse ESCs. If this were the case, the decrease in CREBBP protein could take place in both *KMT2D*-mutant mouse ESCs and HEK293 cell lines, but have a more pronounced effect on H3K27ac levels in HEK293 where CREBBP is the major H3K27 acetylase.

While the effect of KMT2D abrogation on active enhancers has been studied before ⁷⁰, I was able to extend upon these observations by exploring epigenetic changes occurring at poised enhancers in *KMT2D*-mutant cells. Similar to active enhancers, a significant proportion (23.6%) of poised enhancers lost both H3K4me1 and the histone mark located on H3K27 (H3K27me3 in the case of poised enhancers). Furthermore, 18.9% of poised enhancers lost H3K27me3 while retaining H3K4me1 in *KMT2D*-mutant cell lines, indicating that, similar to H3K27ac, presence of H3K4me1 alone is not sufficient for H3K27 tri-methylation. Whereas it is intuitive to consider the capability for KMT2D and CREBBP to interact at active enhancer regions, as they are both activators of transcription, it is perhaps more difficult to rationalize an interaction between KMT2D and H3K27 tri-methylase EZH2, as the two enzymes functionally antagonize one another in terms of their abilities to activate and repress transcription, respectively. Moreover, the KMT2D-containing COMPASS complex contains the H3K27 de-methylase UTX, which directly functions in removing H3K27me3¹³⁴. The widespread, H3K4me1-independent loss of

H3K27me3 at poised enhancers in *KMT2D*-mutant HEK293 cell lines therefore remains a paradox that provides rationale for further investigation into the KMT2D-EZH2 regulatory axis.

4.2 KMT2D mutation is associated with transcriptional alterations

Previous studies of KMT2D-dependent gene expression have indicated a role for KMT2D in maintaining expression of genes involved in a diverse set of pathways, and are consistent with the notion of KMT2D-dependent genes being context dependent ^{93,94}. To expand on the current knowledge of the KMT2D regulome, I identified sets of genes with expression patterns positively and negatively associated with *KMT2D* mutation in HEK293 cell lines.

The observation that the majority of KMT2D-dependent genes were down-regulated in *KMT2D*-mutant cell lines is consistent with the notion of KMT2D functioning as a transcriptional activator. Furthermore, the down-regulation of *LAMB3* in *KMT2D*-mutant HEK293 supported the validity of the *KMT2D*-mutant HEK293 system as a model for studying KMT2D function, as *LAMB3* has been previously found to be down-regulated in both *KMT2D*-mutant HeLa and HCT116 cell lines^{93,94} and therefore represents a context-independent target of KMT2D regulation.

While transcriptional alterations associated with *KMT2D* mutation have been investigated in several cellular contexts, the direct relevancy of KMT2D-dependent genes in tumorigenesis has only been demonstrated in the context of B-cell lymphoma^{33,73}. It therefore remains unclear how *KMT2D* mutations may contribute to malignancy in the many other forms of cancer *KMT2D* is found mutated in (Figure 1.2). *KMT2D*-mutant HEK293 cell lines showed down-regulation of

PARP14, a member of the PARP family of genes, many of which are known to function in DNA damage repair¹³⁵. Importantly, PARP14 depletion has been shown to be associated with deficiencies in DNA damage repair¹³⁶. Deficiency in DNA damage repair, mediated by decreased *PARP14* expression levels in *KMT2D*-mutant cells, could potentially represent a mechanism by which *KMT2D* mutation facilitates tumorigenesis, and is supported a previous study performed by Kantidakis *et al.* (2016)¹³⁷, who found alterations in *KMT2D* mutation to be associated with deficiencies in DNA damage repair in mouse embryonic fibroblasts (MEFs)¹³⁷. However, whether PARP family genes are dysregulated in *KMT2D*-mutant cancers, along with whether *KMT2D*-mutant HEK293 cell lines demonstrate deficiencies in DNA damage response, remains to be determined.

S100 family proteins are involved in a variety of distinct cellular pathways that are relevant in tumorigenesis¹³⁸, and *S100A2/3/4/5/14/16* have previously been found to be down-regulated in a *KMT2D*-mutant HCT116 cell line⁹⁴. I was able to reproduce the association between *S100* family gene expression and *KMT2D* mutation in HEK293 cell lines, which showed significant down-regulation of *S100A4/16*. *S100A16* encodes a calcium binding protein that is ubiquitously expressed and highly conserved in mammals¹³⁹. The mechanism by which *S100A16* down-regulation in *KMT2D*-mutant cells could contribute to malignancy is unclear, as *S100A16* has instead been found to be up-regulated in several cancer types¹³⁹. However, there exist other genes that show different patterns of dysregulation when mutated in different cancer types. One example of this can be seen with the gene *EZH2*, in which gain and loss of function alterations are recurrently found in myelodysplastic syndromes (MDS) and T-cell acute lymphoblastic leukemia, respectively¹⁴⁰. Down-regulation of *S100A4* in *KMT2D*-mutant cells is particularly

interesting, as S100A4 is known to enhance the apoptotic function of tumor suppressor TP53¹⁴¹. Thus, down-regulation of *S100A4* in *KMT2D*-mutant cells could facilitate malignancy by hindering TP53 activity, although this remains to be explored in *KMT2D*-mutant tumors and cell lines.

The over-representation of genes involved in cellular adhesion and cytoskeletal organization among genes dysregulated in KMT2D-mutant HEK293 cell lines is consistent with observations in several previous studies of KMT2D function^{68,93}. In addition to reproducing results of previous studies in the HEK293 cell line model system, I was able to extend upon these observations by identifying enrichment of two pathways that provide further insight into the potential role of KMT2D mutations in tumorigenesis. Angiogenesis involves the local formation of blood vessels to meet oxygen and nutrient demands of surrounding cells, and was included among the initial six hallmarks of cancer¹². Abe et al. (2001)¹²⁵ identified a set of 29 genes upregulated in HUVEC cells at 30 minutes after stimulation with vascular endothelial growth factor A (VEGFA)¹²⁵ – a protein well-known for it's functioning in angiogenesis¹⁴². This set of genes, positively regulated by VEGFA, was enriched among up-regulated genes in KMT2D-mutant HEK293 cell lines, with up-regulated members of the gene set including tribbles pseudokinase 1 (TRIBI) early growth response 1 and 2 (EGR1/2) and tachykinin precursor 1 (TACI). While the relationship between KMT2D and angiogenesis is not apparent in current literature, it has been shown that KMT2A depletion is associated with increased angiogenesis, through modulation of HOX family gene expression, in HUVEC cells¹⁴³. Given the ability of the KMT2D-containing complex to regulate HOX family gene expression¹³⁴, and the up-regulation of VEGFA targets in KMT2D-mutant cell lines (shown here), a role of KMT2D mutation in promoting angiogenesis

represents a potential mechanism by which *KMT2D* mutations contribute to tumorigenesis, and would perhaps benefit from further investigation.

Additional insight into how *KMT2D* mutation may contribute to malignancy was generated through the observation of significant over-representation of genes belonging to the cancer module (Module 55) of mSigDB among genes down-regulated in *KMT2D*-mutant cells. Module 55 (http://robotics.stanford.edu/~erans/cancer/modules/module_55) is a computationally generated mSigDB gene set composed of 834 genes found to consistently overlap with genes dysregulated across a wide spectrum of cancer types; 36 of which were down-regulated in *KMT2D*-mutant cells. These included several genes with known tumor suppressive functions such as *BCAM*¹⁴⁴, *TRIM29*¹⁴⁵, *DLC1*¹⁴⁶, and *EMP3*¹⁴⁷, thereby indicating that *KMT2D* mutation may facilitate oncogenesis through repression of other genes that normally function in tumor suppression, although this remains to be explored in the context of *KMT2D*-mutant tumor samples.

Detailed comparison of H3K4me1 distributions between *KMT2D*-mutant and wildtype cell lines revealed that 1,506 genes gained H3K4me1 at their TSS regions in both *KMT2D*-mutant cell lines (Section 3.1.3). 4 of these 1,506 genes (*EPHA8, FAM155B, TAC1* and *RIPPLY2*) were significantly (adj. p < 0.05) up-regulated in *KMT2D*-mutant cell lines, indicating that gain of H3K4me1 at TSS regions was associated with increased mRNA abundance for a subset of genes in *KMT2D*-mutant cells. Interestingly, 74 of these 1,506 genes showed presence of H3K4me3 at their TSS regions, and gained H3K27ac in addition to H3K4me1 in both *KMT2D*-mutant cell lines. Among these 74 genes was the significantly (adj. p < 0.05) up-regulated gene *FAM155B*,

whose function remains poorly characterized in the literature. GSEA of the 74 genes with TSS regions that gained both H3K4me1 and H3K27ac (and retained H3K4me3) showed enrichment of genes known to be down-regulated by MAPK8/JNK1¹²³, and included *SSR4*, *INSIG1*, *AIMP2*, *SLC6A6*, *GATA6*, *STX4*, *CTTN* and *TOR1AIP1*. However, it remains unclear how the addition of histone modifications positively associated with transcription (H3K4me1 and H3K27ac) at the TSS regions of these genes could contribute to tumorigenesis, as these 8 genes did not show differences in mRNA abundance in *KMT2D*-mutant cells compared to wildtype cells (Wilcoxon p = 0.26).

4.3 Integrative analysis of the KMT2D regulome reveals convergence between KMT2D-dependent enhancers and the TGF-beta and RA signaling networks

Identification of alterations in the transcriptional and epigenetic landscapes of *KMT2D*-mutant cells lead to the integrative analyses described in Chapter 3.3. Given that enhancers are known to regulate target genes from an average distance of $< 100 \text{ kb}^{48}$, it was expected that a proportion of KMT2D-dependent genes would be within this distance of a KMT2D-dependent enhancer. This was true for 32 down-regulated genes, which had annotated TSS regions within 100 kb of a KMT2D-dependent active enhancer, and indicated that a proportion of genes with significant (adj. p < 0.05) changes in expression in *KMT2D*-mutant cells may have resulted from KMT2D-dependent loss of nearby active enhancers. Interestingly, *S100A4* was among these genes, which not only supported previous findings regarding the role of KMT2D in regulating *S100A* family gene expression is associated with KMT2D-dependent loss of H3K4me1 and H3K27ac at nearby enhancer regions.

The observation that 79% of significantly (adj. p < 0.05) down-regulated genes were not within 100 kb of a KMT2D-dependent active enhancer indicated that KMT2D-dependent expression changes in these genes are independent of histone modification alterations in nearby enhancers. One possible explanation regarding how these genes may be dysregulated in *KMT2D*-mutant cells is that they are regulated by KMT2D-dependent enhancers at distances greater than 100 kb. Long-range (> 100 kb) enhancer-TSS interactions have been demonstrated in previous studies^{48,124}, including a recent study of KMT2D-mediated enhancer interactions⁷¹. A second explanation can be derived from another recent study regarding KMT2D function, in which KMT2D-dependent transcriptional alterations were associated with decreased Pol II binding at enhancers, in addition to decreased enhancer RNA (eRNA) synthesis¹⁴⁸. It is thus possible that the significantly down-regulated genes in *KMT2D*-mutant HEK293 cell lines could be adjacent to enhancers that experience reduced Pol II binding and/or eRNA expression, independent of histone modification alteration, although this was not able to be directly tested using the data available in my thesis project.

The fact that the majority (> 95%) of KMT2D-dependent poised and active enhancers were not within 100 kb of a DE gene indicated that epigenetic alterations in KMT2D-dependent enhancers were not sufficient for inducing significant (adj. p < 0.05) changes in expression of nearby genes. However, the consequences of KMT2D-dependent enhancer alteration remained an open question. The observation that genes associated with KMT2D-dependent enhancers were not dysregulated in KMT2D-mutant cells was consistent with results from a study performed by Wang *et al.* (2016)⁷⁰, who also showed that KMT2D-dependent enhancers were dispensable for

maintaining transcription of their associated genes⁷⁰. Instead, the authors found that genes associated with KMT2D-dependent enhancers were only dysregulated in *KMT2D*-mutant cells during differentiation, indicating the requirement of additional transcriptional co-regulators for KMT2D-dependent enhancer disruption to affect transcription. Using GSEA, I was able to show that genes associated with KMT2D-dependent active enhancers in HEK293 cell lines were enriched for members of the TGF-beta and RA signaling networks, highlighting the activation of these two pathways as candidate contexts in which to study the transcriptional consequences of KMT2D-dependent enhancer alteration.

The TGF-beta signaling network comprises a complex signal transduction cascade, mediated by TGF-beta receptor-ligand interactions, that ultimately regulates gene expression patterns relevant to several processes such as organism development¹⁴⁹ and programmed cell death¹⁵⁰.

Interestingly, deficiencies in TGF-beta signaling have been implicated in the development of cancers in which *KMT2D* is frequently mutated, such as diffuse large B-cell lymphomas^{151,152}. The observation that genes associated with KMT2D-dependent active enhancers are enriched for targets of TGF-beta transcription factors (SMAD2/3)¹⁵³ as well as genes up-regulated upon TGF-beta stimulation¹⁵⁴ indicates that KMT2D-dependent active enhancers play a role in TGF-beta-mediated activation of gene expression. Furthermore, the fact that KMT2D-dependent active enhancers lose activating histone modifications H3K4me1 and H3K27ac in *KMT2D*-mutant cells indicates that *KMT2D* mutation may lead to a decreased capability for TGF-beta signals to activate gene transcription. Finally, given the implication of TGF-beta signaling deficiencies in cancer^{151,152}, the impairment of TGF-beta-mediated activation of transcription in *KMT2D*-mutant cells may represent a mechanism by which mutation of *KMT2D* contributes to tumorigenesis. I

would hypothesize that the expression of TGF-beta response genes (in particular, those contained in the TGF-beta-related gene sets that were enriched among genes associated with KMT2D-dependent active enhancers) is increased in wildtype cells upon stimulation with TGF-beta ligands, while the increase in expression of these genes is attenuated in *KMT2D*-mutant cells. While experiments designed to address this hypothesis were outside the scope of my thesis work, investigation into the association between TGF-beta signaling and KMT2D remains a topic for future research.

Cellular response to RA involves direct binding of RA ligands to retinoic acid receptors (RARs) RARA/B/G, which subsequently bind retinoid X receptors (RXRA/B/G) to form heterodimeric proteins that bind and regulate transcription of target genes¹⁵⁵. The transcriptional activation of RAR/RXR target genes orchestrates a variety of developmental pathways, highlighting RA signaling as a key developmental component of mammalian cells¹⁵⁵. Disruption of RA signaling has been associated with several types of cancer, including acute promyelocytic leukemia (APL), in which the PML-RARA fusion event acts as a driver alteration that is thought to contribute to oncogenesis through defects in RA target gene activation¹⁵⁶. Deficiencies in RA signaling have been associated with KMT2D mutation in two previous studies, which showed a reduced ability for KMT2D-mutant cells to active RA response genes upon treatment of RA 92,94. The observation that genes associated with KMT2D-dependent active enhancers were enriched for RA response genes is consistent with the notion of the KMT2D and RA regulatory networks converging, and extends upon previous work by indicating loss of active enhancers as a means by which KMT2D mutation may disrupt RA signaling. Furthermore, disruptions in the RA signaling axis have been associated with lymphoma development in mice¹⁵⁷, thereby positioning malfunctions in RA

signaling as a possible mechanism by which *KMT2D* mutations contribute to oncogenesis, and highlighting investigation into RA signaling deficiencies in *KMT2D*-mutant tumors as an additional topic for future research.

4.4 Concluding remarks

Overall, the research presented in this thesis provides additional insight into the impact of KMT2D mutation on the epigenetic and transcriptomic landscapes of human cells. Through comprehensive profiling of KMT2D-dependent H3K4me1, I was able to reveal a subset of regions that gain H3K4me1 in cells lacking functional KMT2D, a result that has only been observed in one recent publication regarding KMT2D function, thereby motivating future investigation regarding the activity of other methyltransferases in KMT2D-mutant cells. Additionally, by performing the first study of KMT2D-dependent epigenetic alterations in poised enhancer regions, I generated the hypothesis that KMT2D is required in the maintenance of poised enhancer regions, providing rationale towards future studies regarding convergence between KMT2D and EZH2 regulatory axes. My integrative analysis of KMT2D-dependent enhancers and genes was consistent with the results of a recent study, which showed that epigenetic alteration of KMT2D-dependent enhancers alone is not sufficient for invoking transcriptional changes in nearby genes. Through identification of the enrichment of TGF-beta and RA signaling genes among genes associated with KMT2D-dependent active enhancers, I provided evidence consistent with the notion of KMT2D being involved in transcriptional activation of genes belonging to the TGF-beta and RA signaling networks. Given the relationship between both the TGF-beta and RA signaling axes and malignancy, my results highlighted the relationship between KMT2D-dependent active enhancers and response to TGF-beta/RA

signaling as a potential topic of future research regarding the consequences of *KMT2D* mutations and their relevance to cancer.

Bibliography

- 1. Finkel, T., Serrano, M. & Blasco, M. A. The common biology of cancer and ageing. *Nature* 448, 767–774 (2007).
- Canadian Cancer Society. Canadian Cancer Statistics 2016. (Canadian Cancer Society's Advisory Committee, 2016).
- 3. Sudhakar, A. History of Cancer, Ancient and Modern Treatment Methods. *J. Cancer Sci. Ther.* **1,** i–iv (2009).
- 4. Breasted, J. H. *The Edwin Smith Surgical Papyrus*. (The University of Chicago Oriental Institute Publications).
- 5. The History of Cancer. in (ed. American Cancer Society) (2014).
- 6. Southam, G. The Nature and Treatment of Cancer. Br. Med. J. (1858).
- 7. Zeidman, I. Metastasis: a review of recent advances. Cancer Res. 17, 157–162 (1957).
- 8. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
- 9. Saiki, R. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).
- 10. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- 11. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).
- 12. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- 13. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

- 14. Hebenstreit, D. Methods, Challenges and Potentials of Single Cell RNA-seq. *Biology* **1**, 658–667 (2012).
- 15. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177–182 (1987).
- 16. Nowell, P. C. & Hungerford, D. A. Chromosome studies in human leukemia. II. Chronic granulocytic leukemia. *J. Natl. Cancer Inst.* **27**, 1013–1035 (1961).
- 17. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- 18. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12,** 32–42 (2011).
- 19. Tuller, T. *et al.* An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**, 344–354 (2010).
- 20. H.J. Muller. Further studies on the nature and causes of gene mutations. *Int. Congr. Genet.* (1932).
- 21. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34,** 155–163 (2015).
- 22. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science* **253**, 49–53 (1991).
- Olivier, M., Hollstein, M. & Hainaut, P. TP53 Mutations in Human Cancers: Origins,
 Consequences, and Clinical Use. *Cold Spring Harb. Perspect. Biol.* 2, a001008–a001008 (2010).
- 24. Zilfou, J. T. & Lowe, S. W. Tumor Suppressive Functions of p53. *Cold Spring Harb*. *Perspect. Biol.* **1,** a001883–a001883 (2009).

- 25. Chen, Z. *et al.* Crucial role of p53-dependent cellular senescence in suppression of Ptendeficient tumorigenesis. *Nature* **436**, 725–730 (2005).
- 26. Morton, J. P. *et al.* Mutant p53 drives metastasis and overcomes growth arrest/senescence in pancreatic cancer. *Proc. Natl. Acad. Sci.* **107**, 246–251 (2010).
- 27. Kang, H. J., Chun, S.-M., Kim, K.-R., Sohn, I. & Sung, C. O. Clinical Relevance of Gain-Of-Function Mutations of p53 in High-Grade Serous Ovarian Carcinoma. *PLoS ONE* **8**, e72609 (2013).
- 28. T.S. Painter and H.J. Muller. Parallel cytology and genetics of induced translocations and deletions in Drosophila. *J Hered* **20**, 287–298 (1929).
- 29. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* **42**, 181–185 (2010).
- 30. Yap, D. B. *et al.* Somatic mutations at EZH2 Y641 act dominantly through a mechanism of selectively altered PRC2 catalytic activity, to increase H3K27 trimethylation. *Blood* **117**, 2451–2459 (2011).
- 31. Schwab, M. Oncogene amplification in solid tumors. Semin. Cancer Biol. 9, 319–325 (1999).
- 32. Good, B. H. & Desai, M. M. Deleterious passengers in adapting populations. *Genetics* **198**, 1183–1208 (2014).
- 33. Ortega-Molina, A. *et al.* The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nat. Med.* **21**, 1199–1208 (2015).
- 34. Wu, R.-C., Wang, T.-L. & Shih, I.-M. The emerging roles of ARID1A in tumor suppression. *Cancer Biol. Ther.* **15,** 655–664 (2014).

- 35. Aoki, K. & Taketo, M. M. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. *J. Cell Sci.* **120,** 3327–3335 (2007).
- 36. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.* (2012). doi:10.1038/nrm3330
- 37. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**, 415–428 (2002).
- 38. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
- 39. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).
- 40. Watt, F. & Molloy, P. L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev.* **2**, 1136–1143 (1988).
- 41. Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nat. Genet.* **42,** 1093–1100 (2010).
- 42. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).
- 43. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
- 44. O'Geen, H., Echipare, L. & Farnham, P. J. Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. in *Epigenetics Protocols* (ed. Tollefsbol, T. O.) **791,** 265–286 (Humana Press, 2011).
- 45. Au, S. L.-K., Wong, C. C.-L., Lee, J. M.-F., Wong, C.-M. & Ng, I. O.-L. EZH2-Mediated H3K27me3 Is Involved in Epigenetic Repression of Deleted in Liver Cancer 1 in Human Cancers. *PloS One* **8**, e68226 (2013).

- 46. Hahn, M. A. *et al.* Loss of the Polycomb Mark from Bivalent Promoters Leads to Activation of Cancer-Promoting Genes in Colorectal Tumors. *Cancer Res.* **74,** 3617–3629 (2014).
- 47. Voigt, P., Tee, W.-W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev.* 27, 1318–1338 (2013).
- 48. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci.* **111,** E2191–E2199 (2014).
- 49. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49,** 825–837 (2013).
- 50. Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.* **15,** 293–306 (2014).
- 51. Probst, A. V., Dunleavy, E. & Almouzni, G. Epigenetic inheritance during the cell cycle.

 Nat. Rev. Mol. Cell Biol. 10, 192–206 (2009).
- 52. Fahrner, J. A., Eguchi, S., Herman, J. G. & Baylin, S. B. Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. *Cancer Res.* **62**, 7213–7218 (2002).
- 53. Koczkowska, M. *et al.* Detection of somatic *BRCA1/2* mutations in ovarian cancer next-generation sequencing analysis of 100 cases. *Cancer Med.* **5**, 1640–1646 (2016).
- 54. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
- 55. Zhu, Q. *et al.* BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* **477**, 179–184 (2011).
- 56. Safran, M. et al. GeneCards Version 3: the human gene integrator. *Database J. Biol. Databases Curation* **2010**, baq020 (2010).

- 57. Schuettengruber, B., Martinez, A.-M., Iovino, N. & Cavalli, G. Trithorax group proteins: switching genes on and keeping them active. *Nat. Rev. Mol. Cell Biol.* **12,** 799–814 (2011).
- 58. Mas-y-Mas, S. *et al.* The Human Mixed Lineage Leukemia 5 (MLL5), a Sequentially and Structurally Divergent SET Domain-Containing Protein with No Intrinsic Catalytic Activity. *PLOS ONE* **11**, e0165139 (2016).
- 59. Stassen, M. J., Bailey, D., Nelson, S., Chinwalla, V. & Harte, P. J. The Drosophila trithorax proteins contain a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins. *Mech. Dev.* **52**, 209–223 (1995).
- 60. Shilatifard, A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu. Rev. Biochem.* **81,** 65–95 (2012).
- 61. Miller, T. *et al.* COMPASS: A complex of proteins associated with a trithorax-related SET domain protein. *Proc. Natl. Acad. Sci.* **98**, 12902–12907 (2001).
- 62. Smith, E., Lin, C. & Shilatifard, A. The super elongation complex (SEC) and MLL in development and disease. *Genes Dev.* **25**, 661–672 (2011).
- 63. Goo, Y.-H. *et al.* Activating signal cointegrator 2 belongs to a novel steady-state complex that contains a subset of trithorax group proteins. *Mol. Cell. Biol.* **23**, 140–149 (2003).
- 64. Kaikkonen, M. U. *et al.* Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Mol. Cell* **51**, 310–325 (2013).
- 65. Lee, S., Lee, J., Lee, S.-K. & Lee, J. W. Activating Signal Cointegrator-2 Is an Essential Adaptor to Recruit Histone H3 Lysine 4 Methyltransferases MLL3 and MLL4 to the Liver X Receptors. *Mol. Endocrinol.* **22**, 1312–1319 (2008).

- 66. Lee, S. *et al.* Crucial Roles for Interactions between MLL3/4 and INI1 in Nuclear Receptor Transactivation. *Mol. Endocrinol.* **23**, 610–619 (2009).
- 67. Kim, D.-H., Lee, J., Lee, B. & Lee, J. W. ASCOM Controls Farnesoid X Receptor

 Transactivation through Its Associated Histone H3 Lysine 4 Methyltransferase Activity. *Mol. Endocrinol.* **23**, 1556–1562 (2009).
- 68. Lee, J.-E. *et al.* H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *eLife* **2**, (2013).
- 69. Hu, D. *et al.* The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. *Mol. Cell. Biol.* **33**, 4745–4754 (2013).
- 70. Wang, C. *et al.* Enhancer priming by H3K4 methyltransferase MLL4 controls cell fate transition. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 11871–11876 (2016).
- 71. Yan, J. *et al.* Histone H3 Lysine 4 methyltransferases MLL3 and MLL4 Modulate Long-range Chromatin Interactions at Enhancers. (2017). doi:10.1101/110239
- 72. Guo, C. *et al.* KMT2D maintains neoplastic cell proliferation and global histone H3 lysine 4 monomethylation. *Oncotarget* **4,** 2144–2153 (2013).
- 73. Zhang, J. *et al.* Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nat. Med.* **21,** 1190–1198 (2015).
- 74. Wang, A. *et al.* Epigenetic Priming of Enhancers Predicts Developmental Competence of hESC-Derived Endodermal Lineage Intermediates. *Cell Stem Cell* **16**, 386–399 (2015).
- 75. Sakabe, N., Savic, D. & Nobrega, M. A. Transcriptional enhancers in development and disease. *Genome Biol.* **13**, 238 (2012).
- 76. Herz, H.-M., Hu, D. & Shilatifard, A. Enhancer Malfunction in Cancer. *Mol. Cell* **53**, 859–866 (2014).

- 77. Adam, M. P. & Hudgins, L. Kabuki syndrome: a review. *Clin. Genet.* **67**, 209–219 (2005).
- 78. Hannibal, M. C. *et al.* Spectrum of MLL2 (ALR) mutations in 110 cases of Kabuki syndrome. *Am. J. Med. Genet. A.* **155,** 1511–1516 (2011).
- 79. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42,** 790–793 (2010).
- 80. Bjornsson, H. T. *et al.* Histone deacetylase inhibition rescues structural and functional brain deficits in a mouse model of Kabuki syndrome. *Sci. Transl. Med.* **6,** 256ra135-256ra135 (2014).
- 81. Mallo, M. & Alonso, C. R. The regulation of Hox gene expression during animal development. *Development* **140**, 3951–3963 (2013).
- 82. Glaser, S. *et al.* Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. *Dev. Camb. Engl.* **133**, 1423–1432 (2006).
- 83. Bernard, O. A. & Berger, R. Molecular basis of 11q23 rearrangements in hematopoietic malignant proliferations. *Genes. Chromosomes Cancer* **13**, 75–85 (1995).
- 84. Slany, R. K. The molecular biology of mixed lineage leukemia. *Haematologica* **94,** 984–993 (2009).
- 85. Ferrando, A. A. Gene expression signatures in MLL-rearranged T-lineage and B-precursor acute leukemias: dominance of HOX dysregulation. *Blood* **102**, 262–268 (2003).
- 86. Rao, R. C. & Dou, Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. *Nat. Rev. Cancer* **15,** 334–346 (2015).
- 87. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298–303 (2011).

- 88. Beà, S. *et al.* Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18250–18255 (2013).
- 89. Parsons, D. W. *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435–439 (2011).
- 90. Spina, V. *et al.* The genetics of nodal marginal zone lymphoma. *Blood* **128**, 1362–1373 (2016).
- 91. Tan, J. et al. Genomic landscapes of breast fibroepithelial tumors. Nat. Genet. 47, 1341–1345 (2015).
- 92. Huff, R. D. Generation and characterization of a lysine (K)-specific methyltransferase 2D knockout human cell line. (2015). doi:10.14288/1.0166181
- 93. Issaeva, I. *et al.* Knockdown of ALR (MLL2) reveals ALR target genes and leads to alterations in cell adhesion and growth. *Mol. Cell. Biol.* **27**, 1889–1903 (2007).
- 94. Guo, C. *et al.* Global identification of MLL2-targeted loci reveals MLL2's role in diverse signaling pathways. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 17603–17608 (2012).
- 95. Pattanayak, V., Ramirez, C. L., Joung, J. K. & Liu, D. R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat. Methods* **8,** 765–770 (2011).
- 96. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
- 97. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space.

 Database J. Biol. Databases Curation **2011**, bar030 (2011).
- 98. R Development Core Team. R: a language and environment for statistical computing. *R Found. Stat. Comput.* **Version 2.0.1,** (2004).

- 99. Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).
- 100. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13,** R77 (2012).
- 101. Zeileis, A., Hornik, K. & Murrell, P. Escaping RGBland: Selecting colors for statistical graphics. *Comput. Stat. Data Anal.* **53**, 3259–3270 (2009).
- 102. Erich Neuwirth. RColorBrewer: ColorBrewer Palettes. (2014).
- 103. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
- 104. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
- 105. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
- 106. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
- 107. Misha Bilenky. FindER: A sesitive analytical tool to study epigenetic modifications and protein-DNA interactions from ChIP-seq data. *Canadian Epigenetics, Environment and Health Research Consortium Network* Available at: http://www.epigenomes.ca/tools-and-software/finder/index.html. (Accessed: 13th June 2017)
- 108. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26,** 841–842 (2010).

- 109. Pellacani, D. *et al.* Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Rep.* **17**, 2060–2074 (2016).
- 110. Neph, S., Reynolds, A. P., Kuehn, M. S. & Stamatoyannopoulos, J. A. Operating on Genomic Ranges Using BEDOPS. *Methods Mol. Biol. Clifton NJ* **1418**, 267–281 (2016).
- 111. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- 112. Butterfield, Y. S. *et al.* JAGuaR: Junction Alignments to Genome for RNA-Seq Reads. *PLoS ONE* **9**, e102398 (2014).
- 113. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).
- 114. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
- 115. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550 (2005).
- 116. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinforma. Oxf. Engl.*27, 1739–1740 (2011).
- 117. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
- 118. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

- 119. Bailey, T. *et al.* Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Comput. Biol.* **9**, e1003326 (2013).
- 120. Zanconato, F. *et al.* Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat. Cell Biol.* **17**, 1218–1227 (2015).
- 121. Biddie, S. C. *et al.* Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. *Mol. Cell* **43**, 145–155 (2011).
- 122. Choukrallah, M.-A., Song, S., Rolink, A. G., Burger, L. & Matthias, P. Enhancer repertoires are reshaped independently of early priming and heterochromatin dynamics during B cell differentiation. *Nat. Commun.* **6**, 8324 (2015).
- 123. Yoshimura, K. *et al.* Regression of abdominal aortic aneurysm by inhibition of c-Jun N-terminal kinase. *Nat. Med.* **11,** 1330–1338 (2005).
- 124. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues.

 Nature **507**, 455–461 (2014).
- 125. Abe, M. & Sato, Y. cDNA microarray analysis of the gene expression profile of VEGF-activated human umbilical vein endothelial cells. *Angiogenesis* **4**, 289–298 (2001).
- 126. Weiss, J. M. *et al.* Neuropilin 1 is expressed on thymus-derived natural regulatory T cells, but not mucosa-generated induced Foxp3+ T reg cells. *J. Exp. Med.* **209,** 1723–1742, S1 (2012).
- 127. Naba, A. *et al.* The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Mol. Cell. Proteomics MCP* **11**, M111.014647 (2012).
- 128. Chicas, A. *et al.* Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. *Cancer Cell* **17,** 376–387 (2010).

- 129. Yao, L., Berman, B. P. & Farnham, P. J. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit. Rev. Biochem. Mol. Biol.* **50**, 550–573 (2015).
- 130. Schnetz, M. P. *et al.* CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. *PLoS Genet.* **6**, e1001023 (2010).
- 131. Thakurela, S., Sahu, S. K., Garding, A. & Tiwari, V. K. Dynamics and function of distal regulatory elements during neurogenesis and neuroplasticity. *Genome Res.* **25,** 1309–1324 (2015).
- 132. Koenecke, N., Johnston, J., Gaertner, B., Natarajan, M. & Zeitlinger, J. Genome-wide identification of Drosophila dorso-ventral enhancers by differential histone acetylation analysis. *Genome Biol.* **17**, (2016).
- 133. Taher, L. & Ovcharenko, I. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinforma. Oxf. Engl.* **25**, 578–584 (2009).
- 134. Agger, K. *et al.* UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. *Nature* **449**, 731–734 (2007).
- 135. Amé, J.-C., Spenlehauer, C. & de Murcia, G. The PARP superfamily. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **26,** 882–893 (2004).
- 136. Nicolae, C. M. *et al.* A novel role for the mono-ADP-ribosyltransferase PARP14/ARTD8 in promoting homologous recombination and protecting against replication stress. *Nucleic Acids Res.* **43**, 3143–3153 (2015).
- 137. Kantidakis, T. *et al.* Mutation of cancer driver *MLL2* results in transcription stress and genome instability. *Genes Dev.* **30**, 408–420 (2016).

- 138. Chen, H., Xu, C., Jin, Q. 'e & Liu, Z. S100 protein family in human cancer. *Am. J. Cancer Res.* **4,** 89–115 (2014).
- 139. Marenholz, I. & Heizmann, C. W. S100A16, a ubiquitously expressed EF-hand protein which is up-regulated in tumors. *Biochem. Biophys. Res. Commun.* **313,** 237–244 (2004).
- 140. Kim, K. H. & Roberts, C. W. M. Targeting EZH2 in cancer. *Nat. Med.* 22, 128–134 (2016).
- 141. Grigorian, M. & Lukanidin, E. [Activator of metastasis in cancer cells, Mst1/S100A4 protein binds to tumor suppressor protein p53]. *Genetika* **39**, 900–908 (2003).
- 142. Hoeben, A. *et al.* Vascular endothelial growth factor and angiogenesis. *Pharmacol. Rev.* **56,** 549–580 (2004).
- 143. Diehl, F., Rössig, L., Zeiher, A. M., Dimmeler, S. & Urbich, C. The histone methyltransferase MLL is an upstream regulator of endothelial-cell sprout formation. *Blood* 109, 1472–1478 (2007).
- 144. Akiyama, H. *et al.* The FBI1/Akirin2 Target Gene, BCAM, Acts as a Suppressive Oncogene. *PLoS ONE* **8**, e78716 (2013).
- 145. Dükel, M. *et al.* The Breast Cancer Tumor Suppressor TRIM29 Is Expressed via ATM-dependent Signaling in Response to Hypoxia. *J. Biol. Chem.* **291**, 21541–21552 (2016).
- 146. Li, G. *et al.* Full activity of the deleted in liver cancer 1 (DLC1) tumor suppressor depends on an LD-like motif that binds talin and focal adhesion kinase (FAK). *Proc. Natl. Acad. Sci. U. S. A.* **108,** 17129–17134 (2011).
- 147. Fumoto, S. *et al.* EMP3 as a tumor suppressor gene for esophageal squamous cell carcinoma. *Cancer Lett.* **274,** 25–32 (2009).

- 148. Dorighi, K. M. *et al.* Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol. Cell* **66**, 568–576.e4 (2017).
- 149. Kitisin, K. *et al.* Tgf-Beta signaling in development. *Sci. STKE Signal Transduct. Knowl. Environ.* **2007**, cm1 (2007).
- 150. Schuster, N. & Krieglstein, K. Mechanisms of TGF-beta-mediated apoptosis. *Cell Tissue Res.* **307**, 1–14 (2002).
- 151. Chen, G. *et al.* Resistance to TGF-?1 correlates with aberrant expression of TGF-? receptor II in human B-cell lymphoma cell lines. *Blood* **109**, 5301–5307 (2007).
- 152. Bakkebø, M., Huse, K., Hilden, V. I., Smeland, E. B. & Oksvold, M. P. TGF-β-induced growth inhibition in B-cell lymphoma correlates with Smad1/5 signalling and constitutively active p38 MAPK. *BMC Immunol.* **11,** 57 (2010).
- 153. Koinuma, D. *et al.* Chromatin immunoprecipitation on microarray analysis of Smad2/3 binding sites reveals roles of ETS1 and TFAP2A in transforming growth factor beta signaling. *Mol. Cell. Biol.* **29**, 172–186 (2009).
- 154. Labbé, E. *et al.* Transcriptional cooperation between the transforming growth factor-beta and Wnt pathways in mammary and intestinal tumorigenesis. *Cancer Res.* **67**, 75–84 (2007).
- 155. Rhinn, M. & Dolle, P. Retinoic acid signalling during development. *Development* **139**, 843–858 (2012).
- 156. Ablain, J. & de Thé, H. Retinoic acid signaling in cancer: The parable of acute promyelocytic leukemia: Retinoic Acid Signaling in Cancer. *Int. J. Cancer* **135**, 2262–2272 (2014).

157. Manshouri, T. *et al.* Downregulation of RAR alpha in mice by antisense transgene leads to a compensatory increase in RAR beta and RAR gamma and development of lymphoma. *Blood* **89**, 2507–2515 (1997).

Appendices

Appendix A Differentially expressed genes identified in KMT2D-mutant HEK293 cell lines

gene	mean expression	log(2) FC	IfcSE	stat	p value	adj. p value
BMP2	1600	-5.2	0.25	-21	6.80E-99	1.30E-94
ZNF532	890	-5.3	0.26	-20	3.20E-88	2.90E-84
PARP14	370	-3.8	0.28	-14	1.90E-42	1.20E-38
SLC2A3	8200	-3.3	0.24	-13	4.80E-41	2.20E-37
GYG2	370	-3.7	0.28	-13	5.10E-40	1.90E-36
SULT1A1	840	-2.9	0.26	-11	1.60E-28	5.10E-25
GGT1	690	-2.7	0.26	-10	2.20E-25	5.80E-22
HSPB8	1300	-2.3	0.24	-9.6	5.00E-22	1.20E-18
COL5A1	11000	-2.3	0.25	-9.2	2.60E-20	5.30E-17
CSDC2	230	-2.5	0.28	-8.8	1.40E-18	2.60E-15
ARSE	170	-2.4	0.28	-8.8	1.70E-18	2.80E-15
SFRP2	1600	2.3	0.27	8.7	2.80E-18	4.30E-15
RP11-215E13.1	150	-2.4	0.28	-8.6	7.60E-18	1.10E-14
NES	730	-2.2	0.26	-8.4	5.30E-17	7.00E-14
EPPK1	2100	-2	0.24	-8.4	6.40E-17	7.90E-14
COL9A3	420	-2.3	0.28	-8.3	1.40E-16	1.70E-13
S100A16	140	-2.2	0.28	-8.1	4.90E-16	5.30E-13
DHRS2	1900	-2.1	0.26	-7.9	2.60E-15	2.60E-12
KRT8	75000	-1.7	0.23	-7.6	2.50E-14	2.40E-11
IL32	610	-2	0.27	-7.5	5.20E-14	4.80E-11
FN1	18000	-1.7	0.23	-7.5	9.30E-14	8.20E-11
SLC2A14	260	-2.1	0.28	-7.4	1.80E-13	1.50E-10
PGR	980	1.9	0.26	7.3	2.10E-13	1.70E-10
L1CAM	890	-1.8	0.25	-7.3	2.70E-13	2.10E-10
LINC00842	750	-2	0.28	-7.2	7.40E-13	5.50E-10
EPS8L2	210	-2	0.28	-7.1	1.30E-12	9.00E-10
ARHGAP23	1100	-1.7	0.25	-7.1	1.50E-12	1.00E-09
EDIL3	2300	-1.6	0.23	-6.9	6.20E-12	4.10E-09
COL3A1	500	-1.9	0.28	-6.8	9.40E-12	6.00E-09
H19	300	-1.9	0.28	-6.6	3.90E-11	2.40E-08
KRT19	5700	-1.7	0.25	-6.6	5.60E-11	3.30E-08

gene	mean	log(2) FC	IfcSE	stat	p value	adj.
	expression					p value
COL1A1	1400	-1.6	0.24	-6.4	1.20E-10	7.00E-08
CRLF1	1000	-1.8	0.27	-6.4	1.40E-10	8.00E-08
ZNF257	89	-1.6	0.26	-6.2	4.20E-10	2.30E-07
RP11-655M14.13	520	-1.7	0.28	-6.2	7.10E-10	3.80E-07
SH3TC2	130	-1.7	0.27	-6.1	9.00E-10	4.60E-07
SYNPO	470	-1.7	0.28	-6.1	9.90E-10	5.00E-07
RBBP7	9800	-1.3	0.22	-6.1	1.00E-09	5.00E-07
IL2RB	87	-1.6	0.26	-6.1	1.20E-09	5.90E-07
MATN2	530	-1.7	0.28	-6	2.60E-09	1.20E-06
ADAM19	1900	-1.5	0.24	-5.9	3.20E-09	1.40E-06
PRAME	1700	-1.4	0.23	-5.9	3.60E-09	1.60E-06
VAMP8	200	-1.6	0.28	-5.9	4.70E-09	2.00E-06
BST2	120	-1.6	0.27	-5.8	5.10E-09	2.10E-06
LAMB3	420	-1.6	0.28	-5.8	5.40E-09	2.20E-06
RAB11B-AS1	1000	-1.5	0.25	-5.8	6.40E-09	2.60E-06
SPARC	9500	-1.2	0.21	-5.8	7.30E-09	2.90E-06
TENM1	140	-1.6	0.28	-5.8	8.70E-09	3.40E-06
SLC35F1	2700	1.4	0.25	5.7	1.00E-08	3.80E-06
SFN	180	-1.6	0.28	-5.7	1.20E-08	4.40E-06
TINAGL1	220	-1.5	0.27	-5.6	1.80E-08	6.40E-06
ARSD	71	-1.4	0.25	-5.6	2.00E-08	7.00E-06
CSPG4	320	-1.6	0.28	-5.6	2.20E-08	7.60E-06
RP11-78F17.1	68	-1.4	0.25	-5.6	2.30E-08	8.00E-06
GSTM4	130	-1.5	0.26	-5.6	2.60E-08	8.80E-06
BRSK2	6700	1.2	0.21	5.5	4.10E-08	1.40E-05
ACVRL1	110	-1.4	0.26	-5.5	4.80E-08	1.60E-05
LOXL1	2600	-1.3	0.23	-5.5	5.00E-08	1.60E-05
PCDH10	3700	1.4	0.25	5.4	5.40E-08	1.70E-05
XIST	88000	-1.1	0.21	-5.4	5.70E-08	1.80E-05
OLR1	69	-1.3	0.24	-5.4	6.10E-08	1.90E-05
JPH2	110	-1.5	0.27	-5.4	6.60E-08	2.00E-05
ENPP2	380	-1.5	0.27	-5.4	8.30E-08	2.40E-05
IGF2	610	-1.5	0.27	-5.3	9.10E-08	2.60E-05
HSPB7	62	-1.3	0.24	-5.3	1.10E-07	3.00E-05
EFEMP2	560	-1.5	0.28	-5.3	1.30E-07	3.60E-05
KLK1	130	-1.4	0.26	-5.3	1.30E-07	3.70E-05
TGM2	380	-1.4	0.28	-5.2	1.60E-07	4.30E-05

gene	mean	log(2) FC	IfcSE	stat	p value	adj.
	expression					p value
TENC1	730	-1.3	0.26	-5.2	1.80E-07	4.80E-05
PHF21B	520	1.4	0.28	5.2	1.80E-07	4.90E-05
PSG4	210	-1.5	0.28	-5.2	2.10E-07	5.40E-05
CCL2	60	-1.2	0.24	-5.2	2.30E-07	5.80E-05
KIAA1161	2200	-1.3	0.24	-5.2	2.50E-07	6.30E-05
TAGLN	440	-1.4	0.27	-5.1	3.10E-07	7.70E-05
MEIS3	780	-1.4	0.27	-5.1	3.20E-07	7.90E-05
LPAR5	76	-1.3	0.26	-5.1	4.30E-07	1.00E-04
MLH3	950	1.3	0.25	5	4.90E-07	0.00012
KRT80	120	-1.3	0.26	-5	5.40E-07	0.00013
KRT8P45	490	-1.4	0.27	-5	5.60E-07	0.00013
RP1-93I3.1	220	-1.4	0.28	-5	5.80E-07	0.00013
DDR2	650	-1.3	0.27	-5	7.20E-07	0.00016
TCF15	55	-1.1	0.23	-4.9	7.90E-07	0.00018
RP11-328K4.1	80	-1.3	0.26	-4.9	8.40E-07	0.00019
FHL1	55000	-1	0.21	-4.9	9.60E-07	0.00021
PLS3	1.00E+05	-1	0.21	-4.9	9.80E-07	0.00021
SALL3	3100	-1.1	0.22	-4.9	9.70E-07	0.00021
EPHA8	830	1.2	0.25	4.9	1.20E-06	0.00026
TRIB1	12000	1.1	0.23	4.8	1.30E-06	0.00026
RAC2	77	-1.2	0.25	-4.8	1.40E-06	0.00029
KIAA1462	3000	-1.2	0.24	-4.8	1.50E-06	0.00031
MAGEA6	53	-1.1	0.23	-4.8	1.60E-06	0.00032
LOXL2	3900	-1.2	0.25	-4.8	1.60E-06	0.00032
MGLL	680	-1.2	0.26	-4.8	1.90E-06	0.00038
IL6R	770	-1.3	0.27	-4.7	2.10E-06	0.00041
KIF1A	13000	1	0.22	4.7	2.10E-06	0.00041
BCAM	2300	-1.1	0.23	-4.7	2.10E-06	0.00041
GAS7	1900	-1.1	0.23	-4.7	2.40E-06	0.00046
BTBD19	120	-1.2	0.26	-4.7	3.00E-06	0.00058
EGR1	1600	1.2	0.25	4.7	3.10E-06	0.00058
SOX21	400	1.3	0.28	4.7	3.10E-06	0.00058
QPRT	1000	-1.2	0.25	-4.7	3.20E-06	0.00058
EHBP1L1	560	-1.3	0.27	-4.6	4.00E-06	0.00072
S100A4	240	-1.3	0.28	-4.6	4.30E-06	0.00077
RNF122	1200	-1.1	0.25	-4.6	4.40E-06	0.00079
FAM155B	3700	1.1	0.24	4.6	4.50E-06	0.00079

gene	mean	log(2) FC	IfcSE	stat	p value	adj.
	expression					p value
PLXDC2	1300	1.1	0.24	4.6	5.30E-06	0.00093
NOS3	58	-1.1	0.24	-4.5	5.50E-06	0.00096
ZNF503-AS2	1500	1.2	0.27	4.5	5.70E-06	0.00099
LGALS1	610	-1.2	0.27	-4.5	6.50E-06	0.0011
FGF12	1400	1.1	0.24	4.5	6.60E-06	0.0011
MICAL2	3800	-1.1	0.25	-4.5	7.30E-06	0.0012
STRA6	55	-1.1	0.24	-4.5	7.20E-06	0.0012
AP1M2	470	-1.2	0.27	-4.5	7.30E-06	0.0012
SEPP1	930	-1.2	0.27	-4.4	8.70E-06	0.0014
ZNF492	230	-1.2	0.28	-4.4	8.70E-06	0.0014
TPM1	7600	-0.95	0.21	-4.4	8.80E-06	0.0014
CARD10	2400	-1.1	0.24	-4.4	9.40E-06	0.0015
ALDH1A2	3200	1.1	0.25	4.4	9.60E-06	0.0015
MATK	1200	1.2	0.27	4.4	9.60E-06	0.0015
LAMA5	12000	-1	0.24	-4.4	1.00E-05	0.0015
LOXL4	530	-1.2	0.27	-4.4	1.10E-05	0.0016
FSTL3	1900	-1	0.24	-4.4	1.20E-05	0.0018
FAM46B	900	-1.2	0.28	-4.4	1.20E-05	0.0019
RRAS	790	-1.2	0.27	-4.4	1.30E-05	0.0019
FLNC	14000	-0.93	0.21	-4.4	1.30E-05	0.0019
SH3KBP1	6100	-0.94	0.22	-4.4	1.30E-05	0.0019
CNN2	4200	-1	0.24	-4.4	1.30E-05	0.0019
CASP4	100	-1.2	0.27	-4.4	1.30E-05	0.002
KRT18	40000	-1	0.23	-4.3	1.40E-05	0.002
TRIM29	44	-0.96	0.22	-4.3	1.50E-05	0.0021
RP11-307C19.1	91	-1.1	0.25	-4.3	1.50E-05	0.0021
SH3TC1	56	-1	0.23	-4.3	1.60E-05	0.0023
LASP1	3700	-1	0.23	-4.3	1.70E-05	0.0024
ITGA3	1400	-1.1	0.25	-4.3	1.70E-05	0.0024
PRR5L	420	-1.2	0.27	-4.3	1.80E-05	0.0025
COL12A1	5800	-1.1	0.25	-4.3	1.90E-05	0.0026
MAP1A	1400	-1	0.24	-4.3	2.00E-05	0.0027
DLC1	2500	-1	0.24	-4.3	2.10E-05	0.0029
TRIM15	67	-1	0.24	-4.2	2.40E-05	0.0032
CTGF	27000	-0.87	0.21	-4.2	2.50E-05	0.0033
ADCYAP1	46	-0.93	0.22	-4.2	2.70E-05	0.0035
ZNF804A	2300	1.1	0.25	4.2	2.90E-05	0.0038

gene	mean	log(2) FC	IfcSE	stat	p value	adj.
	expression					p value
C1orf106	540	-1.1	0.26	-4.1	3.40E-05	0.0045
FAM19A5	41	-0.88	0.21	-4.1	3.70E-05	0.0048
PCYT1B	1800	-0.98	0.24	-4.1	4.10E-05	0.0052
TNC	2200	-1	0.25	-4.1	4.00E-05	0.0052
SYT11	3300	0.92	0.23	4.1	4.70E-05	0.0059
LIMS2	280	-1.1	0.28	-4.1	4.90E-05	0.0061
ALPK2	69	-1	0.25	-4.1	5.10E-05	0.0063
HCN4	250	-1.1	0.28	-4.1	5.10E-05	0.0063
RP11-196G18.22	1900	0.93	0.23	4	5.60E-05	0.0068
EGR2	330	1.1	0.28	4	5.70E-05	0.0069
SYP	1200	1.1	0.28	4	6.10E-05	0.0074
PLEC	23000	-0.95	0.24	-4	6.30E-05	0.0076
JDP2	430	-1.1	0.27	-4	6.40E-05	0.0077
PHLDB2	3600	-0.93	0.23	-4	6.80E-05	0.0081
OLFM1	1000	-1.1	0.28	-4	7.40E-05	0.0086
PLEKHB1	2000	-0.94	0.24	-4	7.30E-05	0.0086
ARMCX2	3600	0.95	0.24	4	7.40E-05	0.0087
FST	280	1.1	0.28	4	7.60E-05	0.0088
GPM6A	690	1.1	0.27	3.9	8.00E-05	0.0092
ELOVL3	520	1.1	0.28	3.9	8.00E-05	0.0092
UNC93B1	1800	-0.93	0.23	-3.9	8.20E-05	0.0093
TAC1	350	1.1	0.28	3.9	8.30E-05	0.0094
SARDH	290	-1.1	0.28	-3.9	8.50E-05	0.0095
PLEKHA4	180	-1.1	0.28	-3.9	8.60E-05	0.0096
BARHL2	1300	1	0.26	3.9	8.70E-05	0.0097
TACC2	2500	-1	0.25	-3.9	9.00E-05	0.0099
EMP3	1400	-0.96	0.25	-3.9	9.00E-05	0.0099
NRN1	2500	0.89	0.23	3.9	9.10E-05	0.0099
ADAMTS15	880	-1.1	0.27	-3.9	9.20E-05	0.01
SP100	1000	-0.97	0.25	-3.9	9.30E-05	0.01
TCIRG1	3000	-0.94	0.24	-3.9	9.70E-05	0.01
RP11-43F13.3	89	-1	0.26	-3.9	9.90E-05	0.011
RIMS2	310	1.1	0.28	3.9	9.90E-05	0.011
ETV7	230	-1.1	0.28	-3.9	0.00011	0.011
F10	120	-1.1	0.28	-3.9	0.00011	0.011
AXL	160	-1.1	0.28	-3.9	0.00011	0.012
ADCY7	590	-1	0.26	-3.9	0.00011	0.012

gene	mean	log(2) FC	IfcSE	stat	p value	adj.
	expression					p value
ADAMTS1	10000	-0.84	0.22	-3.9	0.00012	0.012
EPAS1	2900	-0.99	0.26	-3.9	0.00012	0.012
SOX21-AS1	490	1	0.27	3.8	0.00013	0.013
SVEP1	1300	-0.95	0.25	-3.8	0.00014	0.014
AL592183.1	1100	-0.93	0.25	-3.8	0.00015	0.015
TGFBI	86	-0.97	0.25	-3.8	0.00015	0.015
EHD2	69	-0.95	0.25	-3.8	0.00016	0.015
ALPPL2	95	-0.99	0.26	-3.8	0.00016	0.016
NYNRIN	1800	-0.91	0.24	-3.8	0.00016	0.016
INA	81	-0.96	0.25	-3.8	0.00017	0.016
VLDLR	1600	-0.9	0.24	-3.7	0.00018	0.018
PREX1	1600	-0.97	0.26	-3.7	0.00019	0.018
PITPNM1	3900	-0.91	0.24	-3.7	0.00019	0.019
LMNA	13000	-0.86	0.23	-3.7	2.00E-04	0.019
AL603965.1	40	-0.81	0.22	-3.7	2.00E-04	0.019
PLA2G4A	810	1	0.27	3.7	2.00E-04	0.019
GPR124	2500	-0.95	0.26	-3.7	0.00021	0.02
CDKL5	720	-0.94	0.26	-3.7	0.00022	0.021
SLC7A8	3200	-0.88	0.24	-3.7	0.00023	0.021
TRIML2	44	-0.79	0.21	-3.7	0.00023	0.022
PHKA2	3000	-0.83	0.23	-3.7	0.00023	0.022
A4GALT	220	-1	0.28	-3.7	0.00024	0.022
EEF1A2	19000	0.9	0.25	3.7	0.00024	0.022
CDH4	2400	-0.85	0.23	-3.7	0.00025	0.022
KRT8P10	230	-1	0.28	-3.7	0.00025	0.023
DLX3	720	-0.96	0.26	-3.7	0.00025	0.023
ETV5	6700	0.78	0.21	3.6	0.00027	0.024
PLXND1	9100	-0.9	0.25	-3.6	0.00028	0.025
RIPPLY2	94	0.96	0.26	3.6	0.00028	0.025
RPS6KA3	13000	-0.77	0.21	-3.6	0.00028	0.025
CYP2S1	510	-1	0.28	-3.6	0.00028	0.025
JUP	13000	-0.83	0.23	-3.6	0.00029	0.025
RGS7	560	0.97	0.27	3.6	3.00E-04	0.026
NEDD4L	6400	-0.77	0.21	-3.6	3.00E-04	0.026
KLHDC7A	77	-0.86	0.24	-3.6	3.00E-04	0.026
TMEM40	43	-0.79	0.22	-3.6	0.00031	0.027
ZNF835	46	-0.81	0.22	-3.6	0.00031	0.027

gene	mean	log(2) FC	IfcSE	stat	p value	adj.
	expression					p value
STXBP5L	1500	0.9	0.25	3.6	0.00031	0.027
AMBN	37	-0.73	0.2	-3.6	0.00032	0.027
SHANK1	220	-1	0.28	-3.6	0.00032	0.027
KNDC1	580	-0.96	0.27	-3.6	0.00033	0.027
HTRA3	840	-0.91	0.25	-3.6	0.00034	0.028
BMP5	40	-0.79	0.22	-3.6	0.00033	0.028
IRF5	590	0.95	0.26	3.6	0.00034	0.028
RP11-94L15.2	710	0.92	0.26	3.6	0.00034	0.028
RIN2	150	-0.98	0.27	-3.6	0.00033	0.028
NUAK2	610	-0.94	0.26	-3.6	0.00034	0.028
SEMA3B	180	-1	0.28	-3.6	0.00034	0.028
IL31RA	240	-0.99	0.28	-3.6	0.00034	0.028
SGK223	1400	-0.92	0.26	-3.6	0.00034	0.028
RP11-849I19.1	250	-1	0.28	-3.6	0.00035	0.028
RASGRF1	56	-0.86	0.24	-3.6	0.00035	0.028
RAPGEF3	490	-0.98	0.27	-3.6	0.00037	0.03
ST6GALNAC5	590	0.95	0.27	3.6	0.00038	0.03
ABCD1	760	-0.94	0.26	-3.6	0.00038	0.03
MGST1	5600	0.95	0.27	3.5	0.00039	0.031
IFITM2	4800	-0.78	0.22	-3.5	0.00039	0.031
PKP3	240	-1	0.28	-3.5	0.00039	0.031
MYEOV	120	-0.91	0.26	-3.5	4.00E-04	0.031
PLXNA4	520	-0.99	0.28	-3.5	0.00043	0.033
CNRIP1	1700	0.83	0.24	3.5	0.00043	0.033
RP11-38P22.2	700	0.93	0.27	3.5	0.00044	0.034
SPINT1	480	-0.99	0.28	-3.5	0.00044	0.034
АТР9В	2500	-0.81	0.23	-3.5	0.00045	0.034
FBLIM1	2700	-0.9	0.26	-3.5	0.00045	0.034
CARD11	1500	-0.68	0.2	-3.5	0.00048	0.036
CSF2RA	42	-0.75	0.22	-3.5	0.00049	0.037
PLAT	1400	-0.86	0.25	-3.5	0.00049	0.037
NCAM1	4100	-0.79	0.23	-3.5	0.00049	0.037
CTSB	5800	-0.76	0.22	-3.5	0.00051	0.038
OLFML3	84	-0.9	0.26	-3.5	0.00051	0.038
GCNT2	280	-0.97	0.28	-3.5	0.00055	0.04
LTBP3	9500	-0.77	0.22	-3.4	0.00057	0.042
CCNYL2	280	-0.96	0.28	-3.4	0.00057	0.042

gene	mean expression	log(2) FC	IfcSE	stat	p value	adj. p value
TBC1D30	2400	0.87	0.25	3.4	0.00057	0.042
ELFN1	560	-0.96	0.28	-3.4	0.00059	0.043
BAI1	390	-0.95	0.28	-3.4	6.00E-04	0.043
WDR66	460	-0.93	0.27	-3.4	6.00E-04	0.043
IGSF11	610	-0.94	0.27	-3.4	0.00061	0.044
MYH9	82000	-0.71	0.21	-3.4	0.00062	0.044
COL18A1	4600	-0.8	0.23	-3.4	0.00062	0.044
WFS1	11000	-0.76	0.22	-3.4	0.00065	0.046
HEY1	2700	-0.77	0.23	-3.4	0.00066	0.047
KIAA1644	2000	-0.85	0.25	-3.4	0.00067	0.047
SELENBP1	950	-0.94	0.28	-3.4	0.00068	0.048
ZNF469	68	-0.86	0.25	-3.4	0.00068	0.048

Appendix B mSigDB gene sets significantly enriched among genes up-regulated in

KMT2D-mutant cell lines

gene set	p value	adj.
		p value
ABE_VEGFA_TARGETS_30MIN	2.60E-07	0.0035
GSE40443_INDUCED_VS_TOTAL_TREG_UP	3.30E-06	0.022
AACTTT_UNKNOWN	7.80E-06	0.026
YNTTTNNNANGCARM_UNKNOWN	6.40E-06	0.026
V\$CDP_01	1.70E-05	0.046

Appendix C Top 200 mSigDB gene sets most significantly enriched among genes down-regulated in *KMT2D*-mutant cell lines

gene set	p value	adj.
		p value
NABA_MATRISOME	5.90E-25	7.80E-21
CHICAS_RB1_TARGETS_CONFLUENT	4.60E-20	3.10E-16
MODULE_5	4.00E-19	1.80E-15
MODULE_55	1.40E-17	4.60E-14
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	2.50E-17	6.50E-14
MODULE_122	2.90E-17	6.50E-14
CHEN_METABOLIC_SYNDROM_NETWORK	4.00E-17	7.70E-14
WU_CELL_MIGRATION	4.60E-17	7.70E-14
MODULE_88	6.40E-17	9.50E-14
MODULE_47	2.20E-16	2.90E-13
LIM_MAMMARY_STEM_CELL_UP	4.40E-16	5.30E-13
GO_EXTRACELLULAR_SPACE	1.40E-15	1.50E-12
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	1.30E-15	1.50E-12
MODULE_1	2.80E-15	2.60E-12
MODULE_38	8.30E-15	7.40E-12
GO_BIOLOGICAL_ADHESION	2.00E-14	1.50E-11
FORTSCHEGGER_PHF8_TARGETS_DN	2.00E-14	1.50E-11
GO_EXTRACELLULAR_MATRIX	9.70E-14	6.80E-11
GO_EXTRACELLULAR_MATRIX_COMPONENT	9.40E-14	6.80E-11
LINDGREN_BLADDER_CANCER_CLUSTER_2B	1.00E-13	7.00E-11
NABA_CORE_MATRISOME	1.50E-13	9.30E-11
SENESE_HDAC2_TARGETS_DN	2.40E-13	1.40E-10
SENESE_HDAC1_AND_HDAC2_TARGETS_DN	2.90E-13	1.60E-10
ISSAEVA_MLL2_TARGETS	2.80E-13	1.60E-10
GO_CIRCULATORY_SYSTEM_DEVELOPMENT	3.90E-13	1.90E-10
GO_CARDIOVASCULAR_SYSTEM_DEVELOPMENT	3.90E-13	1.90E-10
ROZANOV_MMP14_TARGETS_UP	3.70E-13	1.90E-10
MODULE_2	1.10E-12	5.40E-10
NABA_MATRISOME_ASSOCIATED	1.40E-12	6.70E-10
HENDRICKS_SMARCA4_TARGETS_UP	1.60E-12	7.10E-10
GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX	1.80E-12	7.60E-10
PID_INTEGRIN1_PATHWAY	2.60E-12	1.10E-09
GO_REGULATION_OF_MULTICELLULAR_ORGANISMAL_DEVELOPMEN	3.80E-12	1.50E-09

gene set	p value	adj.
		p value
Т		
GO_REGULATION_OF_CELLULAR_COMPONENT_MOVEMENT	4.70E-12	1.90E-09
HAN_SATB1_TARGETS_UP	8.20E-12	3.20E-09
GO_BASEMENT_MEMBRANE	8.60E-12	3.20E-09
GGGAGGRR_V\$MAZ_Q6	9.90E-12	3.60E-09
VERHAAK_GLIOBLASTOMA_MESENCHYMAL	1.10E-11	4.00E-09
GO_REGULATION_OF_CELL_DIFFERENTIATION	1.20E-11	4.20E-09
WONG_ADULT_TISSUE_STEM_MODULE	1.70E-11	5.60E-09
MODULE_84	2.60E-11	8.50E-09
JOHNSTONE_PARVB_TARGETS_3_UP	2.90E-11	9.20E-09
LEF1_UP.V1_UP	3.70E-11	1.10E-08
ESC_V6.5_UP_EARLY.V1_DN	3.70E-11	1.10E-08
WANG_SMARCE1_TARGETS_UP	4.70E-11	1.40E-08
PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_UP	5.50E-11	1.60E-08
KIM_GLIS2_TARGETS_UP	5.90E-11	1.70E-08
GO_REGULATION_OF_ANATOMICAL_STRUCTURE_MORPHOGENESIS	6.90E-11	1.90E-08
MODULE_12	1.50E-10	4.20E-08
LIU_PROSTATE_CANCER_DN	1.90E-10	4.90E-08
PASINI_SUZ12_TARGETS_DN	1.90E-10	4.90E-08
GO_POSITIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	2.20E-10	5.60E-08
LEI_MYB_TARGETS	2.20E-10	5.60E-08
GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	2.40E-10	6.10E-08
GO_TISSUE_DEVELOPMENT	2.50E-10	6.10E-08
MODULE_24	2.70E-10	6.40E-08
MODULE_52	2.90E-10	6.80E-08
GO_VASCULATURE_DEVELOPMENT	3.10E-10	7.20E-08
SENESE_HDAC1_TARGETS_DN	4.20E-10	9.60E-08
GO_POSITIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_PRO	4.60E-10	1.00E-07
CESS	F 20F 10	1 105 07
GO_ANCHORING_JUNCTION	5.20E-10	1.10E-07
DELYS_THYROID_CANCER_UP	5.20E-10	1.10E-07
MODULE_234	6.20E-10	1.30E-07
GO_LOCOMOTION GO_ROSITIVE_PEGLILATION_OF_CELL_DIFFERENTIATION	6.60E-10	1.40E-07
GO_POSITIVE_REGULATION_OF_CELL_DIFFERENTIATION	6.80E-10	1.40E-07
SWEET_KRAS_TARGETS_UP	7.00E-10	1.40E-07
SMID_BREAST_CANCER_LUMINAL_B_DN	7.60E-10	1.50E-07
DURAND_STROMA_S_UP	9.00E-10	1.80E-07

gene set	p value	adj.
		p value
PLASARI_TGFB1_TARGETS_10HR_UP	9.40E-10	1.80E-07
SENESE_HDAC3_TARGETS_DN	1.20E-09	2.20E-07
GO_ANATOMICAL_STRUCTURE_FORMATION_INVOLVED_IN_MORPH	1.20E-09	2.20E-07
OGENESIS		
NABA_ECM_GLYCOPROTEINS	1.40E-09	2.60E-07
RODWELL_AGING_KIDNEY_UP	1.60E-09	2.90E-07
HUANG_FOXA2_TARGETS_DN	1.60E-09	2.90E-07
GO_REGULATION_OF_CELL_MORPHOGENESIS	1.90E-09	3.30E-07
MODULE_6	2.00E-09	3.60E-07
KOYAMA_SEMA3B_TARGETS_UP	2.50E-09	4.30E-07
V\$AP1_Q6	2.70E-09	4.70E-07
MODULE_128	2.80E-09	4.80E-07
GO_CELL_SUBSTRATE_JUNCTION	3.00E-09	5.00E-07
MODULE_170	3.30E-09	5.50E-07
MODULE_79	3.90E-09	6.40E-07
TGANTCA_V\$AP1_C	4.10E-09	6.60E-07
BOQUEST_STEM_CELL_CULTURED_VS_FRESH_UP	4.80E-09	7.60E-07
GO_RESPONSE_TO_EXTERNAL_STIMULUS	5.00E-09	7.80E-07
MODULE_118	5.50E-09	8.50E-07
MODULE_44	6.70E-09	1.00E-06
GO_STRUCTURAL_MOLECULE_ACTIVITY	7.10E-09	1.10E-06
MODULE_60	7.10E-09	1.10E-06
GO_INTERMEDIATE_FILAMENT_CYTOSKELETON	7.40E-09	1.10E-06
WGGAATGY_V\$TEF1_Q6	1.60E-08	2.30E-06
GO_CELL_SURFACE	2.00E-08	2.80E-06
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	2.00E-08	2.80E-06
MARTENS_TRETINOIN_RESPONSE_UP	2.00E-08	2.80E-06
GO_LOCALIZATION_OF_CELL	2.20E-08	3.00E-06
GO_CELL_MOTILITY	2.20E-08	3.00E-06
GO_BLOOD_VESSEL_MORPHOGENESIS	2.20E-08	3.00E-06
RICKMAN_TUMOR_DIFFERENTIATED_WELL_VS_POORLY_DN	2.30E-08	3.10E-06
GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_ORGANIZA	2.40E-08	3.10E-06
TION		
ONDER_CDH1_TARGETS_2_DN	2.30E-08	3.10E-06
DANG_REGULATED_BY_MYC_DN	2.40E-08	3.20E-06
GNF2_PTX3	2.90E-08	3.80E-06
GO_ANGIOGENESIS	3.00E-08	3.90E-06

gene set	p value	adj.
		p value
PETROVA_ENDOTHELIUM_LYMPHATIC_VS_BLOOD_DN	3.80E-08	4.80E-06
RB_DN.V1_DN	4.80E-08	6.10E-06
GU_PDEF_TARGETS_UP	5.40E-08	6.80E-06
GO_CELL_DEVELOPMENT	5.70E-08	7.10E-06
GO_CELL_MORPHOGENESIS_INVOLVED_IN_DIFFERENTIATION	5.80E-08	7.10E-06
chrxp22	6.50E-08	8.00E-06
BOQUEST_STEM_CELL_UP	6.90E-08	8.40E-06
GO_INTERMEDIATE_FILAMENT	7.30E-08	8.70E-06
CROONQUIST_STROMAL_STIMULATION_UP	7.30E-08	8.70E-06
CAGGTG_V\$E12_Q6	8.10E-08	9.60E-06
LIU_SMARCA4_TARGETS	8.70E-08	1.00E-05
GILDEA_METASTASIS	8.80E-08	1.00E-05
KHETCHOUMIAN_TRIM24_TARGETS_UP	9.90E-08	1.10E-05
RODWELL_AGING_KIDNEY_NO_BLOOD_UP	1.10E-07	1.20E-05
REN_ALVEOLAR_RHABDOMYOSARCOMA_DN	1.10E-07	1.20E-05
GO_REGULATION_OF_CELL_PROLIFERATION	1.20E-07	1.30E-05
GO_RESPONSE_TO_WOUNDING	1.20E-07	1.30E-05
WIEDERSCHAIN_TARGETS_OF_BMI1_AND_PCGF2	1.20E-07	1.30E-05
GO_REGULATION_OF_CELL_ADHESION	1.30E-07	1.40E-05
HAN_SATB1_TARGETS_DN	1.30E-07	1.40E-05
GO_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	1.40E-07	1.50E-05
PEREZ_TP53_TARGETS	1.50E-07	1.60E-05
DODD_NASOPHARYNGEAL_CARCINOMA_UP	1.80E-07	1.90E-05
GO_MOVEMENT_OF_CELL_OR_SUBCELLULAR_COMPONENT	1.80E-07	1.90E-05
GSE2405_0H_VS_1.5H_A_PHAGOCYTOPHILUM_STIM_NEUTROPHIL_UP	1.90E-07	2.00E-05
RYTTCCTG_V\$ETS2_B	2.20E-07	2.30E-05
GO_CELLULAR_RESPONSE_TO_ORGANIC_SUBSTANCE	2.20E-07	2.30E-05
GO_MUSCLE_STRUCTURE_DEVELOPMENT	2.70E-07	2.70E-05
GO_REGULATION_OF_NEURON_PROJECTION_DEVELOPMENT	2.70E-07	2.70E-05
GO_EXTRACELLULAR_MATRIX_BINDING	2.80E-07	2.80E-05
GO_REGULATION_OF_CELL_PROJECTION_ORGANIZATION	2.90E-07	2.90E-05
GO_POSITIVE_REGULATION_OF_LOCOMOTION	2.90E-07	2.90E-05
CHIARADONNA_NEOPLASTIC_TRANSFORMATION_CDC25_UP	3.10E-07	3.00E-05
CROMER_TUMORIGENESIS_UP	3.40E-07	3.30E-05
GO_REPRODUCTIVE_SYSTEM_DEVELOPMENT	3.60E-07	3.50E-05
TGGAAA_V\$NFAT_Q4_01	3.70E-07	3.50E-05

December December	gene set	p value	adj.
JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP			-
GO_REGULATION_OF_CELL_GROWTH		3.70E-07	3.50E-05
NABA_ECM_REGULATORS 4.30E-07 4.00E-05	JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP	4.00E-07	3.80E-05
GOTZMANN_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP 4.60E-07 4.20E-05 LINDGREN_BLADDER_CANCER_HIGH_RECURRENCE 4.80E-07 4.40E-05 PANGAS_TUMOR_SUPPRESSION_BY_SMAD1_AND_SMAD5_UP 4.90E-07 4.50E-05 BERENIENO_TRANSFORMED_BY_RHOA_REVERSIBLY_DN 4.90E-07 4.50E-05 GO_REGULATION_OF_CELL_DEVELOPMENT 5.50E-07 5.40E-05 MODULE_220 6.10E-07 5.60E-05 DELACROIX_RARG_BOUND_MEF 6.30E-07 5.60E-05 GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION 6.30E-07 5.90E-05 CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.20E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.30E-07 7.70E-05 V\$STATSB_01	GO_REGULATION_OF_CELL_GROWTH	4.20E-07	3.90E-05
LINDGREN_BLADDER_CANCER_HIGH_RECURRENCE	NABA_ECM_REGULATORS	4.30E-07	4.00E-05
PANGAS_TUMOR_SUPPRESSION_BY_SMAD1_AND_SMAD5_UP 4.90E-07 4.50E-05 BERENJENO_TRANSFORMED_BY_RHOA_REVERSIBLY_DN 4.90E-07 4.50E-05 GO_REGULATION_OF_CELL_DEVELOPMENT 5.50E-07 4.90E-05 MODULE_220 6.10E-07 5.40E-05 DELACROIX_RARG_BOUND_MEF 6.30E-07 5.60E-05 GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION 6.70E-07 5.90E-05 CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANSTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.80E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 7.00E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05	GOTZMANN_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP	4.60E-07	4.20E-05
BERENJENO_TRANSFORMED_BY_RHOA_REVERSIBLY_DN 4.90E-07 4.50E-05 GO_REGULATION_OF_CELL_DEVELOPMENT 5.50E-07 4.90E-05 MODULE_220 6.10E-07 5.40E-05 DELACROIX_RARG_BOUND_MEF 6.30E-07 5.60E-05 GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION 6.70E-07 5.90E-05 CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.50E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 6.80E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 8.00E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 GESE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP 8.50E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STATSB_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04	LINDGREN_BLADDER_CANCER_HIGH_RECURRENCE	4.80E-07	4.40E-05
GO_REGULATION_OF_CELL_DEVELOPMENT 5.50E-07 4.90E-05 MODULE_220 6.10E-07 5.40E-05 DELACROIX_RARG_BOUND_MEF 6.30E-07 5.60E-05 GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION 6.30E-07 5.60E-05 CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.80E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 7.00E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.20E-05 TION 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STATSB_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 UP SMID_BRE	PANGAS_TUMOR_SUPPRESSION_BY_SMAD1_AND_SMAD5_UP	4.90E-07	4.50E-05
MODULE_220 6.10E-07 5.40E-05 DELACROIX_RARG_BOUND_MEF 6.30E-07 5.60E-05 GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION 6.30E-07 5.60E-05 CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.80E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 7.00E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 8.70E-07 7.20E-05 MARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 VID_BREAST_CANCER_BASAL_UP <	BERENJENO_TRANSFORMED_BY_RHOA_REVERSIBLY_DN	4.90E-07	4.50E-05
DELACROIX_RARG_BOUND_MEF 6.30E-07 5.60E-05 GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION 6.30E-07 5.60E-05 CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 6.80E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 7.00E-05 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04	GO_REGULATION_OF_CELL_DEVELOPMENT	5.50E-07	4.90E-05
GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION 6.30E-07 5.60E-05 CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.50E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 7.00E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 8.70E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_US_DUCTAL_NORMAL_US_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CARCINOMA_VS_DUCKS_CAR	MODULE_220	6.10E-07	5.40E-05
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN 6.70E-07 5.90E-05 ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.50E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 6.80E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 8.20E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 UP 5MID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04	DELACROIX_RARG_BOUND_MEF	6.30E-07	5.60E-05
ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE 6.80E-07 5.90E-05 GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.50E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 7.00E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 7.00E-05 TION 8.20E-07 7.00E-05 GSE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 UP 5.00E-05 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 <t< th=""><th>GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION</th><th>6.30E-07</th><th>5.60E-05</th></t<>	GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION	6.30E-07	5.60E-05
GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS 7.00E-07 6.00E-05 GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.50E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 6.80E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.50E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGE	CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN	6.70E-07	5.90E-05
GO_CELL_JUNCTION_ASSEMBLY 7.50E-07 6.50E-05 WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 6.80E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 7.20E-05 7.20E-05 GSE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.50E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 0.00011	ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE	6.80E-07	5.90E-05
WILCOX_RESPONSE_TO_PROGESTERONE_DN 8.00E-07 6.80E-05 GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC TION 8.20E-07 7.00E-05 GSE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS	7.00E-07	6.00E-05
GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC 8.20E-07 7.00E-05 TION 7.00E-05 GSE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00	GO_CELL_JUNCTION_ASSEMBLY	7.50E-07	6.50E-05
TION SSE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	WILCOX_RESPONSE_TO_PROGESTERONE_DN	8.00E-07	6.80E-05
GSE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP 8.50E-07 7.20E-05 HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUC	8.20E-07	7.00E-05
HARRIS_HYPOXIA 8.70E-07 7.20E-05 CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	TION		
CROMER_METASTASIS_DN 8.70E-07 7.20E-05 MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	GSE22935_UNSTIM_VS_48H_MBOVIS_BCG_STIM_MACROPHAGE_UP	8.50E-07	7.20E-05
MODULE_342 9.10E-07 7.50E-05 PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	HARRIS_HYPOXIA	8.70E-07	7.20E-05
PETROVA_PROX1_TARGETS_DN 9.30E-07 7.70E-05 V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	CROMER_METASTASIS_DN	8.70E-07	7.20E-05
V\$STAT5B_01 9.70E-07 8.00E-05 HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	MODULE_342	9.10E-07	7.50E-05
HINATA_NFKB_TARGETS_KERATINOCYTE_UP 1.20E-06 1.00E-04 PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	PETROVA_PROX1_TARGETS_DN	9.30E-07	7.70E-05
PID_AVB3_INTEGRIN_PATHWAY 1.20E-06 1.00E-04 TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	V\$STAT5B_01	9.70E-07	8.00E-05
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	HINATA_NFKB_TARGETS_KERATINOCYTE_UP	1.20E-06	1.00E-04
UP 1.30E-06 1.00E-04 SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011	PID_AVB3_INTEGRIN_PATHWAY	1.20E-06	1.00E-04
SMID_BREAST_CANCER_BASAL_UP 1.30E-06 1.00E-04 KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011		1.30E-06	1.00E-04
KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN 1.30E-06 1.00E-04 MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011		1.30F-06	1.00F-04
MODULE_176 1.30E-06 1.00E-04 GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011			
GO_REGULATION_OF_VASCULATURE_DEVELOPMENT 1.30E-06 1.00E-04 SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011			
SERVITJA_ISLET_HNF1A_TARGETS_UP 1.40E-06 0.00011			
	MODULE 75		
PID_INTEGRIN3_PATHWAY 1.50E-06 0.00012	_		
HALLMARK_ESTROGEN_RESPONSE_EARLY 1.60E-06 0.00012			

gene set	p value	adj. p value
KEGG_ECM_RECEPTOR_INTERACTION	1.70E-06	0.00013
GO_REGULATION_OF_CELL_MORPHOGENESIS_INVOLVED_IN_DIFFER ENTIATION	1.80E-06	0.00013
KOINUMA_TARGETS_OF_SMAD2_OR_SMAD3	1.80E-06	0.00014
KRAS.600_UP.V1_DN	1.80E-06	0.00014
GO_EPITHELIUM_DEVELOPMENT	1.90E-06	0.00014
CAHOY_ASTROGLIAL	1.90E-06	0.00014
GO_COLLAGEN_BINDING	2.00E-06	0.00014
ONDER_CDH1_TARGETS_2_UP	2.10E-06	0.00016
DUTERTRE_ESTRADIOL_RESPONSE_24HR_DN	2.20E-06	0.00016
GSE6259_CD4_TCELL_VS_CD8_TCELL_UP	2.20E-06	0.00016
GO_COLLAGEN_FIBRIL_ORGANIZATION	2.20E-06	0.00016
MODULE_438	2.20E-06	0.00016
GO_CELL_JUNCTION	2.30E-06	0.00016
GRAESSMANN_RESPONSE_TO_MC_AND_SERUM_DEPRIVATION_UP	2.30E-06	0.00016
GO_CELL_JUNCTION_ORGANIZATION	2.30E-06	0.00016
MIYAGAWA_TARGETS_OF_EWSR1_ETS_FUSIONS_DN	2.40E-06	0.00016
KRAS.300_UP.V1_DN	2.40E-06	0.00017
ARGGGTTAA_UNKNOWN	2.50E-06	0.00017
POOLA_INVASIVE_BREAST_CANCER_DN	2.70E-06	0.00019
HELLER_HDAC_TARGETS_SILENCED_BY_METHYLATION_UP	2.90E-06	2.00E-04
HELLER_SILENCED_BY_METHYLATION_UP	2.90E-06	2.00E-04
GO_REGULATION_OF_NEURON_DIFFERENTIATION	3.10E-06	0.00021
GO_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUCTION	3.20E-06	0.00022
CHEBOTAEV_GR_TARGETS_DN	3.20E-06	0.00022
GO_POSITIVE_REGULATION_OF_CELL_DEVELOPMENT	3.40E-06	0.00022

Appendix D GO terms significantly enriched among genes down-regulated in *KMT2D*-mutant cell lines

gene set	p value	adj.
		p value
GO_EXTRACELLULAR_SPACE	1.40E-15	5.00E-12
GO_BIOLOGICAL_ADHESION	2.00E-14	3.50E-11
GO_EXTRACELLULAR_MATRIX	9.70E-14	8.60E-11
GO_EXTRACELLULAR_MATRIX_COMPONENT	9.40E-14	8.60E-11
GO_CIRCULATORY_SYSTEM_DEVELOPMENT	3.90E-13	2.30E-10
GO_CARDIOVASCULAR_SYSTEM_DEVELOPMENT	3.90E-13	2.30E-10
GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX	1.80E-12	9.00E-10
GO_REGULATION_OF_MULTICELLULAR_ORGANISMAL_DEVELOPME	3.80E-12	1.70E-09
NT		
GO_REGULATION_OF_CELLULAR_COMPONENT_MOVEMENT	4.70E-12	1.90E-09
GO_BASEMENT_MEMBRANE	8.60E-12	3.10E-09
GO_REGULATION_OF_CELL_DIFFERENTIATION	1.20E-11	4.00E-09
GO_REGULATION_OF_ANATOMICAL_STRUCTURE_MORPHOGENESI	6.90E-11	2.00E-08
S		
GO_POSITIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	2.20E-10	5.90E-08
GO_TISSUE_DEVELOPMENT	2.50E-10	6.00E-08
GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	2.40E-10	6.00E-08
GO_VASCULATURE_DEVELOPMENT	3.10E-10	6.90E-08
GO_POSITIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_PR OCESS	4.60E-10	9.60E-08
GO_ANCHORING_JUNCTION	5.20E-10	1.00E-07
GO_LOCOMOTION	6.60E-10	1.20E-07
GO_POSITIVE_REGULATION_OF_CELL_DIFFERENTIATION	6.80E-10	1.20E-07
GO_ANATOMICAL_STRUCTURE_FORMATION_INVOLVED_IN_MORP HOGENESIS	1.20E-09	2.00E-07
GO_REGULATION_OF_CELL_MORPHOGENESIS	1.90E-09	3.00E-07
GO_CELL_SUBSTRATE_JUNCTION	3.00E-09	4.60E-07
GO_RESPONSE_TO_EXTERNAL_STIMULUS	5.00E-09	7.40E-07
GO_STRUCTURAL_MOLECULE_ACTIVITY	7.10E-09	1.00E-06
GO_INTERMEDIATE_FILAMENT_CYTOSKELETON	7.40E-09	1.00E-06
GO_CELL_SURFACE	2.00E-08	2.60E-06
GO_LOCALIZATION_OF_CELL	2.20E-08	2.60E-06
GO_CELL_MOTILITY	2.20E-08	2.60E-06
GO_BLOOD_VESSEL_MORPHOGENESIS	2.20E-08	2.60E-06

gene set	p value	adj.
		p value
GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_ORGANIZ ATION	2.40E-08	2.70E-06
GO_ANGIOGENESIS	3.00E-08	3.30E-06
GO_CELL_DEVELOPMENT	5.70E-08	6.00E-06
GO_CELL_MORPHOGENESIS_INVOLVED_IN_DIFFERENTIATION	5.80E-08	6.00E-06
GO_INTERMEDIATE_FILAMENT	7.30E-08	7.40E-06
GO_REGULATION_OF_CELL_PROLIFERATION	1.20E-07	1.10E-05
GO_RESPONSE_TO_WOUNDING	1.20E-07	1.10E-05
GO_REGULATION_OF_CELL_ADHESION	1.30E-07	1.20E-05
GO_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	1.40E-07	1.30E-05
GO_MOVEMENT_OF_CELL_OR_SUBCELLULAR_COMPONENT	1.80E-07	1.60E-05
GO_CELLULAR_RESPONSE_TO_ORGANIC_SUBSTANCE	2.20E-07	2.00E-05
GO_MUSCLE_STRUCTURE_DEVELOPMENT	2.70E-07	2.20E-05
GO_REGULATION_OF_NEURON_PROJECTION_DEVELOPMENT	2.70E-07	2.20E-05
GO_EXTRACELLULAR_MATRIX_BINDING	2.80E-07	2.20E-05
GO_REGULATION_OF_CELL_PROJECTION_ORGANIZATION	2.90E-07	2.30E-05
GO_POSITIVE_REGULATION_OF_LOCOMOTION	2.90E-07	2.30E-05
GO_REPRODUCTIVE_SYSTEM_DEVELOPMENT	3.60E-07	2.80E-05
GO_REGULATION_OF_CELL_GROWTH	4.20E-07	3.10E-05
GO_REGULATION_OF_CELL_DEVELOPMENT	5.50E-07	4.00E-05
GO_REGULATION_OF_CELL_SUBSTRATE_ADHESION	6.30E-07	4.50E-05
GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS	7.00E-07	4.90E-05
GO_CELL_JUNCTION_ASSEMBLY	7.50E-07	5.20E-05
GO_POSITIVE_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSD UCTION	8.20E-07	5.50E-05
GO_REGULATION_OF_VASCULATURE_DEVELOPMENT	1.30E-06	8.70E-05
GO_REGULATION_OF_CELL_MORPHOGENESIS_INVOLVED_IN_DIFFE RENTIATION	1.80E-06	0.00011
GO_EPITHELIUM_DEVELOPMENT	1.90E-06	0.00012
GO_COLLAGEN_BINDING	2.00E-06	0.00012
GO_COLLAGEN_FIBRIL_ORGANIZATION	2.20E-06	0.00014
GO_CELL_JUNCTION	2.30E-06	0.00014
GO_CELL_JUNCTION_ORGANIZATION	2.30E-06	0.00014
GO_REGULATION_OF_NEURON_DIFFERENTIATION	3.10E-06	0.00018
GO_REGULATION_OF_INTRACELLULAR_SIGNAL_TRANSDUCTION	3.20E-06	0.00018
GO_POSITIVE_REGULATION_OF_CELL_DEVELOPMENT	3.40E-06	0.00019
GO_POSITIVE_REGULATION_OF_CELL_COMMUNICATION	3.40E-06	0.00019

gene set	p value	adj.
		p value
GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT	3.50E-06	0.00019
GO_REGULATION_OF_GROWTH	3.50E-06	0.00019
GO_POSITIVE_REGULATION_OF_CELL_PROJECTION_ORGANIZATION	3.60E-06	0.00019
GO_CELLULAR_COMPONENT_MORPHOGENESIS	3.90E-06	2.00E-04
GO_MULTICELLULAR_ORGANISMAL_MACROMOLECULE_METABOLI C_PROCESS	4.30E-06	0.00022
GO_GROWTH_FACTOR_BINDING	4.40E-06	0.00023
GO_REGULATION_OF_CELL_DEATH	5.30E-06	0.00026
GO_NEGATIVE_REGULATION_OF_CELL_DEATH	5.40E-06	0.00027
GO_NEUROGENESIS	6.20E-06	0.00029
GO_ORGAN_MORPHOGENESIS	6.10E-06	0.00029
GO_NEGATIVE_REGULATION_OF_LOCOMOTION	6.10E-06	0.00029
GO_RESPONSE_TO_ACID_CHEMICAL	6.40E-06	0.00029
GO_CELL_ADHESION_MOLECULE_BINDING	6.30E-06	0.00029
GO_INTEGRIN_BINDING	6.40E-06	0.00029
GO_POSITIVE_REGULATION_OF_CELL_PROLIFERATION	7.60E-06	0.00034
GO_CYTOKINE_BINDING	7.60E-06	0.00034
GO_NEGATIVE_REGULATION_OF_CELL_SUBSTRATE_ADHESION	8.30E-06	0.00036
GO_RECEPTOR_ACTIVITY	8.50E-06	0.00037
GO_REPRODUCTION	9.40E-06	4.00E-04
GO_CARGO_RECEPTOR_ACTIVITY	9.60E-06	0.00041
GO_RESPONSE_TO_OXYGEN_CONTAINING_COMPOUND	1.00E-05	0.00042
GO_REGULATION_OF_EXTENT_OF_CELL_GROWTH	1.30E-05	0.00052
GO_MULTICELLULAR_ORGANISM_METABOLIC_PROCESS	1.30E-05	0.00052
GO_REGULATION_OF_ACTIN_FILAMENT_BUNDLE_ASSEMBLY	1.30E-05	0.00052
GO_WOUND_HEALING	1.40E-05	0.00055
GO_TUBE_DEVELOPMENT	1.40E-05	0.00057
GO_PROTEIN_COMPLEX_BINDING	1.50E-05	0.00057
GO_ARTERY_DEVELOPMENT	1.60E-05	0.00061
GO_RESPONSE_TO_CYTOKINE	1.80E-05	0.00067
GO_LEUKOCYTE_MIGRATION	1.80E-05	7.00E-04
GO_MUSCLE_CELL_DIFFERENTIATION	1.90E-05	0.00072
GO_HEART_DEVELOPMENT	2.40E-05	0.00088
GO_RESPONSE_TO_MECHANICAL_STIMULUS	2.60E-05	0.00094
GO_IMMUNE_SYSTEM_PROCESS	2.60E-05	0.00095
GO_RESPONSE_TO_ABIOTIC_STIMULUS	2.70E-05	0.00098
GO_POSITIVE_REGULATION_OF_PHOSPHORUS_METABOLIC_PROCE	3.00E-05	0.001

gene set	p value	adj. p value
SS		
GO_POSITIVE_REGULATION_OF_PHOSPHATE_METABOLIC_PROCESS	3.00E-05	0.001
GO_DEVELOPMENTAL_PROCESS_INVOLVED_IN_REPRODUCTION	3.20E-05	0.0011
GO_ACTIN_BINDING	3.50E-05	0.0012
GO_CYTOSKELETON_ORGANIZATION	3.50E-05	0.0012
GO_POSITIVE_REGULATION_OF_NEURON_PROJECTION_DEVELOPM ENT	4.00E-05	0.0014
GO_CELL_SUBSTRATE_JUNCTION_ASSEMBLY	4.70E-05	0.0016
GO_NEGATIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_P ROCESS	5.00E-05	0.0017
GO_REGULATION_OF_EPITHELIAL_TO_MESENCHYMAL_TRANSITION	5.20E-05	0.0017
GO_MACROMOLECULAR_COMPLEX_BINDING	5.60E-05	0.0018
GO_POSITIVE_REGULATION_OF_NEURON_DIFFERENTIATION	5.60E-05	0.0018
GO_POSITIVE_REGULATION_OF_CELL_MORPHOGENESIS_INVOLVED _IN_DIFFERENTIATION	6.00E-05	0.0019
GO_CELLULAR_RESPONSE_TO_ENDOGENOUS_STIMULUS	6.00E-05	0.0019
GO_POSITIVE_REGULATION_OF_PROTEIN_KINASE_B_SIGNALING	6.30E-05	0.002
GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT	6.40E-05	0.002
GO_REGULATION_OF_ACTIN_FILAMENT_BASED_PROCESS	6.50E-05	0.002
GO_REGULATION_OF_CELL_SIZE	7.00E-05	0.0021
GO_ODONTOGENESIS	7.30E-05	0.0022
GO_CELLULAR_RESPONSE_TO_CYTOKINE_STIMULUS	7.50E-05	0.0023
GO_COLLAGEN_TRIMER	7.60E-05	0.0023
GO_EMBRYO_DEVELOPMENT	8.10E-05	0.0024
GO_NEGATIVE_REGULATION_OF_CELL_PROLIFERATION	8.10E-05	0.0024
GO_POSITIVE_REGULATION_OF_CELL_ADHESION	8.10E-05	0.0024
GO_MUSCLE_ORGAN_DEVELOPMENT	8.20E-05	0.0024
GO_TISSUE_MORPHOGENESIS	8.30E-05	0.0024
GO_FEMALE_SEX_DIFFERENTIATION	9.80E-05	0.0028
GO_REGULATION_OF_STEM_CELL_DIFFERENTIATION	1.00E-04	0.003
GO_RESPONSE_TO_ETHANOL	0.00011	0.0031
GO_REGULATION_OF_PROTEIN_KINASE_B_SIGNALING	0.00011	0.0031
GO_POSITIVE_REGULATION_OF_STEM_CELL_DIFFERENTIATION	0.00012	0.0033
GO_ENDOPLASMIC_RETICULUM_LUMEN	0.00012	0.0033
GO_TISSUE_REMODELING	0.00012	0.0033
GO_EPITHELIAL_CELL_DIFFERENTIATION	0.00013	0.0034
GO_NEGATIVE_REGULATION_OF_CELL_ADHESION	0.00013	0.0034

gene set	p value	adj.
		p value
GO_RESPIRATORY_SYSTEM_DEVELOPMENT	0.00013	0.0034
GO_CELL_MATRIX_ADHESION	0.00013	0.0034
GO_EXTERNAL_SIDE_OF_PLASMA_MEMBRANE	0.00013	0.0034
GO_INTEGRIN_MEDIATED_SIGNALING_PATHWAY	0.00013	0.0034
GO_RECEPTOR_COMPLEX	0.00016	0.004
GO_REGULATION_OF_CHEMOTAXIS	0.00015	0.004
GO_ENDOCYTOSIS	0.00016	0.0042
GO_POSITIVE_REGULATION_OF_MOLECULAR_FUNCTION	0.00018	0.0044
GO_ACTIN_FILAMENT_BUNDLE	0.00019	0.0047
GO_SKIN_DEVELOPMENT	0.00019	0.0047
GO_HEAD_DEVELOPMENT	2.00E-04	0.005
GO_CELL_SUBSTRATE_ADHESION	0.00021	0.0052
GO_INTRINSIC_COMPONENT_OF_PLASMA_MEMBRANE	0.00023	0.0055
GO_POSITIVE_REGULATION_OF_PROTEIN_MODIFICATION_PROCES S	0.00023	0.0055
GO_REGULATION_OF_PHOSPHORUS_METABOLIC_PROCESS	0.00023	0.0055
GO_REGULATION_OF_MULTICELLULAR_ORGANISMAL_METABOLIC_ PROCESS	0.00024	0.0055
GO_BLOOD_VESSEL_REMODELING	0.00024	0.0055
GO_LAMININ_BINDING	0.00024	0.0055
GO_RESPONSE_TO_LIPID	0.00024	0.0056
GO_GLYCOPROTEIN_BINDING	0.00027	0.0062
GO_STRIATED_MUSCLE_CELL_DIFFERENTIATION	0.00027	0.0062
GO_SCAVENGER_RECEPTOR_ACTIVITY	0.00027	0.0063
GO_RECEPTOR_BINDING	0.00028	0.0064
GO_ENDOTHELIAL_CELL_MIGRATION	0.00028	0.0064
GO_NEGATIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	0.00029	0.0066
GO_REGULATION_OF_RESPONSE_TO_EXTERNAL_STIMULUS	3.00E-04	0.0066
GO_CELLULAR_RESPONSE_TO_OXYGEN_CONTAINING_COMPOUND	3.00E-04	0.0066
GO_RESPONSE_TO_ALCOHOL	3.00E-04	0.0066
GO_RESPONSE_TO_ENDOGENOUS_STIMULUS	3.00E-04	0.0066
GO_ACTOMYOSIN	0.00031	0.0067
GO_REGULATION_OF_BODY_FLUID_LEVELS	0.00032	0.0069
GO_HEART_MORPHOGENESIS	0.00032	0.0069
GO_REGULATION_OF_AXONOGENESIS	0.00033	0.007
GO_CORTICAL_ACTIN_CYTOSKELETON	0.00034	0.0072
GO_EPITHELIAL_TO_MESENCHYMAL_TRANSITION	0.00034	0.0072

gene set	p value	adj.
		p value
GO_HEART_VALVE_DEVELOPMENT	0.00036	0.0076
GO_CYTOKINE_ACTIVITY	4.00E-04	0.0084
GO_PLACENTA_DEVELOPMENT	4.00E-04	0.0084
GO_SKELETAL_SYSTEM_DEVELOPMENT	0.00041	0.0084
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	0.00041	0.0084
GO_GLIAL_CELL_MIGRATION	0.00042	0.0085
GO_TAXIS	0.00045	0.009
GO_REGULATION_OF_CELL_SHAPE	0.00045	0.009
GO_MESENCHYMAL_CELL_DIFFERENTIATION	0.00045	0.009
GO_SINGLE_ORGANISM_CELL_ADHESION	0.00046	0.0092
GO_RAS_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY	0.00047	0.0094
GO_POSITIVE_REGULATION_OF_EPITHELIAL_TO_MESENCHYMAL_T RANSITION	0.00047	0.0094
GO_DEFENSE_RESPONSE	0.00051	0.01
GO_POSITIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS	0.00052	0.01
GO_RESPONSE_TO_TRANSFORMING_GROWTH_FACTOR_BETA	0.00058	0.011
GO_POSITIVE_REGULATION_OF_AXON_EXTENSION	6.00E-04	0.012
GO_MESENCHYME_DEVELOPMENT	0.00061	0.012
GO_OSSIFICATION	0.00068	0.013
GO_RESPONSE_TO_GROWTH_FACTOR	0.00069	0.013
GO_RESPONSE_TO_BIOTIC_STIMULUS	0.00071	0.013
GO_POSITIVE_REGULATION_OF_CELL_GROWTH	0.00074	0.014
GO_CELL_ACTIVATION	0.00076	0.014
GO_CELL_PROJECTION_ORGANIZATION	0.00077	0.014
GO_RESPONSE_TO_CORTICOSTEROID	0.00077	0.014
GO_SEX_DIFFERENTIATION	0.00081	0.015
GO_REGULATION_OF_HYDROLASE_ACTIVITY	0.00083	0.015
GO_INFLAMMATORY_RESPONSE	0.00089	0.016
GO_HEPARIN_BINDING	0.00089	0.016
GO_ENDOCRINE_SYSTEM_DEVELOPMENT	0.00089	0.016
GO_ODONTOGENESIS_OF_DENTIN_CONTAINING_TOOTH	9.00E-04	0.016
GO_ENDOPEPTIDASE_ACTIVITY	0.00092	0.016
GO_AMEBOIDAL_TYPE_CELL_MIGRATION	0.00093	0.016
GO_REGULATION_OF_RHO_PROTEIN_SIGNAL_TRANSDUCTION	0.00094	0.017
GO_SCAFFOLD_PROTEIN_BINDING	0.00094	0.017
GO_UROGENITAL_SYSTEM_DEVELOPMENT	0.00097	0.017
GO_REGULATION_OF_POTASSIUM_ION_TRANSPORT	0.00097	0.017

gene set	p value	adj.
		p value
GO_ACTOMYOSIN_STRUCTURE_ORGANIZATION	0.00097	0.017
GO_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS	0.001	0.017
GO_CELLULAR_RESPONSE_TO_ACID_CHEMICAL	0.001	0.017
GO_SITE_OF_POLARIZED_GROWTH	0.001	0.017
GO_POSITIVE_REGULATION_OF_LYASE_ACTIVITY	0.001	0.018
GO_MULTICELLULAR_ORGANISMAL_HOMEOSTASIS	0.0011	0.018
GO_POSITIVE_REGULATION_OF_CYCLIC_NUCLEOTIDE_METABOLIC_	0.0011	0.019
PROCESS	2 2211	2.212
GO_ENDOTHELIAL_CELL_DIFFERENTIATION	0.0011	0.019
GO_SECRETORY_GRANULE_LUMEN	0.0012	0.019
GO_WATER_HOMEOSTASIS	0.0012	0.019
GO_POSITIVE_REGULATION_OF_CYCLASE_ACTIVITY	0.0012	0.019
GO_POSITIVE_REGULATION_OF_ACTIN_FILAMENT_BUNDLE_ASSEM	0.0012	0.019
BLY	0.0040	0.010
GO_AORTA_DEVELOPMENT	0.0012	0.019
GO_CELL_CORTEX_PART	0.0012	0.019
GO_CELLULAR_RESPONSE_TO_EXTERNAL_STIMULUS	0.0012	0.019
GO_TISSUE_MIGRATION	0.0013	0.021
GO_MUSCLE_CELL_DEVELOPMENT	0.0013	0.021
GO_RESPONSE_TO_ORGANIC_CYCLIC_COMPOUND	0.0014	0.022
GO_PLASMA_MEMBRANE_ORGANIZATION	0.0014	0.022
GO_CORTICAL_CYTOSKELETON	0.0014	0.022
GO_SECRETORY_GRANULE	0.0014	0.022
GO_NEGATIVE_REGULATION_OF_RESPONSE_TO_STIMULUS	0.0014	0.022
GO_SIGNAL_TRANSDUCER_ACTIVITY	0.0015	0.023
GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS	0.0015	0.024
GO_EMBRYONIC_PLACENTA_DEVELOPMENT	0.0016	0.024
GO_IN_UTERO_EMBRYONIC_DEVELOPMENT	0.0016	0.024
GO_PERINUCLEAR_REGION_OF_CYTOPLASM	0.0016	0.025
GO_ACTIN_CYTOSKELETON	0.0016	0.025
GO_MULTI_ORGANISM_REPRODUCTIVE_PROCESS	0.0017	0.025
GO_MEMBRANE_REGION	0.0017	0.026
GO_IMMUNE_RESPONSE	0.0018	0.026
GO_CALCIUM_ION_BINDING	0.0017	0.026
GO_CELL_PROLIFERATION	0.0018	0.026
GO_REGULATION_OF_ANATOMICAL_STRUCTURE_SIZE	0.0017	0.026
GO_RESPONSE_TO_VITAMIN	0.0018	0.026

gene set	p value	adj.
		p value
GO_ACTIN_FILAMENT_BASED_PROCESS	0.0018	0.027
GO_TRABECULA_FORMATION	0.0019	0.028
GO_COMPLEX_OF_COLLAGEN_TRIMERS	0.0019	0.028
GO_MEMORY	0.002	0.029
GO_RESPONSE_TO_BMP	0.002	0.029
GO_CELLULAR_RESPONSE_TO_BMP_STIMULUS	0.002	0.029
GO_POSITIVE_REGULATION_OF_PEPTIDYL_TYROSINE_PHOSPHORYL	0.002	0.029
ATION		
GO_REGULATION_OF_DENDRITE_DEVELOPMENT	0.002	0.029
GO_PLASMA_MEMBRANE_REGION	0.002	0.029
GO_REGULATION_OF_ERK1_AND_ERK2_CASCADE	0.0021	0.03
GO_SARCOMERE_ORGANIZATION	0.0022	0.031
GO_REGULATION_OF_CELLULAR_COMPONENT_SIZE	0.0022	0.031
GO_POSITIVE_REGULATION_OF_GROWTH	0.0023	0.032
GO_CELL_CELL_JUNCTION	0.0023	0.032
GO_MULTI_MULTICELLULAR_ORGANISM_PROCESS	0.0023	0.032
GO_MODIFIED_AMINO_ACID_BINDING	0.0023	0.032
GO_MATERNAL_PROCESS_INVOLVED_IN_FEMALE_PREGNANCY	0.0023	0.032
GO_POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY	0.0023	0.032
GO_CYTOPLASMIC_REGION	0.0023	0.032
GO_REGULATION_OF_DEVELOPMENTAL_GROWTH	0.0024	0.033
GO_RECEPTOR_MEDIATED_ENDOCYTOSIS	0.0024	0.033
GO_CONTRACTILE_FIBER	0.0024	0.033
GO_STEM_CELL_DIFFERENTIATION	0.0024	0.033
GO_NEGATIVE_REGULATION_OF_POTASSIUM_ION_TRANSPORT	0.0024	0.033
GO_CARDIAC_EPITHELIAL_TO_MESENCHYMAL_TRANSITION	0.0024	0.033
GO_POSITIVE_REGULATION_OF_HYDROLASE_ACTIVITY	0.0025	0.034
GO_CELLULAR_RESPONSE_TO_BIOTIC_STIMULUS	0.0025	0.034
GO_TUBE_MORPHOGENESIS	0.0026	0.034
GO_MORPHOGENESIS_OF_AN_EPITHELIUM	0.0026	0.035
GO_OVULATION_CYCLE	0.0026	0.035
GO_ENDOTHELIUM_DEVELOPMENT	0.0026	0.035
GO_BIOMINERAL_TISSUE_DEVELOPMENT	0.0027	0.035
GO_CELL_CORTEX	0.0027	0.036
GO_RESPONSE_TO_INTERFERON_GAMMA	0.0028	0.036
GO_RESPONSE_TO_PROSTAGLANDIN	0.0028	0.036
GO_NEGATIVE_REGULATION_OF_CELL_MATRIX_ADHESION	0.0028	0.036

gene set	p value	adj. p value
GO_REGULATION_OF_EXTRACELLULAR_MATRIX_ORGANIZATION	0.0028	0.036
GO POSITIVE REGULATION OF CARTILAGE DEVELOPMENT	0.0028	0.036
GO RESPONSE TO HORMONE	0.0028	0.036
GO ENDOCYTIC VESICLE	0.0028	0.036
GO CYTOKINE RECEPTOR ACTIVITY	0.0029	0.037
GO POSITIVE REGULATION OF OSTEOBLAST DIFFERENTIATION	0.0029	0.037
GO RESPONSE TO RETINOIC ACID	0.0029	0.037
GO LEUKOCYTE DIFFERENTIATION	0.003	0.038
GO SULFUR COMPOUND BINDING	0.003	0.038
GO NEURON PROJECTION DEVELOPMENT	0.0031	0.038
GO_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY	0.0031	0.038
GO_POSITIVE_REGULATION_OF_TYROSINE_PHOSPHORYLATION_OF	0.0031	0.039
_STAT3_PROTEIN		
GO_INTRACELLULAR_SIGNAL_TRANSDUCTION	0.0032	0.039
GO_POSITIVE_REGULATION_OF_PURINE_NUCLEOTIDE_METABOLIC	0.0032	0.04
_PROCESS		
GO_POSITIVE_REGULATION_OF_NUCLEOTIDE_METABOLIC_PROCES	0.0032	0.04
S		
GO_REGULATION_OF_DEFENSE_RESPONSE	0.0034	0.041
GO_GLYCOSAMINOGLYCAN_BINDING	0.0034	0.041
GO_VESICLE_LUMEN	0.0034	0.041
GO_REGULATION_OF_LYASE_ACTIVITY	0.0034	0.041
GO_SALIVARY_GLAND_DEVELOPMENT	0.0035	0.042
GO_RESPONSE_TO_PEPTIDE	0.0036	0.043
GO_CELLULAR_RESPONSE_TO_ALCOHOL	0.0036	0.043
GO_POSITIVE_REGULATION_OF_CAMP_METABOLIC_PROCESS	0.0036	0.043
GO_CELL_LEADING_EDGE	0.0037	0.044
GO_REGULATION_OF_HEART_CONTRACTION	0.0037	0.044
GO_REGULATION_OF_RESPONSE_TO_WOUNDING	0.0038	0.045
GO_CELL_MORPHOGENESIS_INVOLVED_IN_NEURON_DIFFERENTIAT	0.0038	0.045
GO_REGULATION_OF_CARTILAGE_DEVELOPMENT	0.0038	0.045
GO_SIGNALING_RECEPTOR_ACTIVITY	0.0039	0.046
GO_NEGATIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS	0.004	0.046
GO_RESPONSE_TO_OXYGEN_LEVELS	0.004	0.046
GO_RESPONSE_TO_AMINO_ACID	0.004	0.046
GO_MUSCLE_ORGAN_MORPHOGENESIS	0.0041	0.047

gene set	p value	adj.
		p value
GO_REGULATION_OF_BLOOD_CIRCULATION	0.0041	0.048
GO_REGULATION_OF_CELL_ACTIVATION	0.0042	0.048
GO_PLASMA_MEMBRANE_RECEPTOR_COMPLEX	0.0043	0.049
GO_STRUCTURAL_CONSTITUENT_OF_MUSCLE	0.0043	0.05
GO_CENTRAL_NERVOUS_SYSTEM_DEVELOPMENT	0.0044	0.05

Appendix E KEGG pathways significantly enriched among genes down-regulated in

KMT2D-mutant cell lines

		adj.
gene set	p value	p value
KEGG_ECM_RECEPTOR_INTERACTION	1.70E-06	0.00026
KEGG_FOCAL_ADHESION	9.90E-06	0.00073

Appendix F Curated gene sets significantly enriched among genes associated with KMT2D-dependent active enhancers

gene set	p value	adj.
		p value
KOINUMA_TARGETS_OF_SMAD2_OR_SMAD3	3.50E-09	2.10E-06
KIM_WT1_TARGETS_UP	4.60E-09	2.10E-06
BHAT_ESR1_TARGETS_VIA_AKT1_UP	1.40E-07	4.30E-05
PEREZ_TP53_TARGETS	5.20E-07	0.00012
BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	2.40E-06	0.00043
LABBE_TGFB1_TARGETS_UP	3.10E-06	0.00047
PEREZ_TP63_TARGETS	4.90E-06	0.00063
DELACROIX_RARG_BOUND_MEF	6.80E-06	0.00076
KIM_WT1_TARGETS_12HR_UP	1.00E-05	0.00099
PEREZ_TP53_AND_TP63_TARGETS	1.60E-05	0.0014
HALMOS_CEBPA_TARGETS_DN	1.70E-05	0.0014
KINSEY_TARGETS_OF_EWSR1_FLII_FUSION_DN	2.40E-05	0.0018
JOHNSTONE_PARVB_TARGETS_3_UP	3.60E-05	0.0025
MARTENS_BOUND_BY_PML_RARA_FUSION	5.60E-05	0.0027
JOHNSTONE_PARVB_TARGETS_2_DN	4.40E-05	0.0027
MIYAGAWA_TARGETS_OF_EWSR1_ETS_FUSIONS_UP	5.30E-05	0.0027
SENESE_HDAC1_AND_HDAC2_TARGETS_DN	4.60E-05	0.0027
JIANG_VHL_TARGETS	5.60E-05	0.0027
BHAT_ESR1_TARGETS_NOT_VIA_AKT1_DN	5.20E-05	0.0027
SENESE_HDAC1_TARGETS_UP	6.30E-05	0.0028
LABBE_TARGETS_OF_TGFB1_AND_WNT3A_UP	9.00E-05	0.0038
FORTSCHEGGER_PHF8_TARGETS_UP	0.00011	0.0044
SENESE_HDAC3_TARGETS_DN	0.00013	0.0048
SENESE_HDAC3_TARGETS_UP	0.00013	0.0048
GUO_HEX_TARGETS_UP	0.00018	0.0064
CHICAS_RB1_TARGETS_CONFLUENT	0.00027	0.0089
SCIAN_INVERSED_TARGETS_OF_TP53_AND_TP73_DN	0.00026	0.0089
DELACROIX_RAR_BOUND_ES	0.00028	0.0091
SANSOM_APC_TARGETS	0.00031	0.0095
STEIN_ESRRA_TARGETS	0.00036	0.011
GOZGIT_ESR1_TARGETS_DN	0.00049	0.014
NUYTTEN_NIPP1_TARGETS_DN	0.00062	0.016
WANG_SMARCE1_TARGETS_UP	0.00059	0.016

gene set	p value	adj.
		p value
WANG_CLIM2_TARGETS_UP	0.00062	0.016
PILON_KLF1_TARGETS_DN	0.00069	0.017
MACLACHLAN_BRCA1_TARGETS_UP	7.00E-04	0.017
WANG_SMARCE1_TARGETS_DN	0.00073	0.018
PASINI_SUZ12_TARGETS_DN	0.00078	0.018
WIERENGA_STAT5A_TARGETS_DN	0.00093	0.021
DAVICIONI_TARGETS_OF_PAX_FOXO1_FUSIONS_UP	0.00099	0.022
BENPORATH_SOX2_TARGETS	0.0011	0.024
OXFORD_RALA_OR_RALB_TARGETS_DN	0.0013	0.027
LOPEZ_MBD_TARGETS	0.0013	0.028
PLASARI_TGFB1_TARGETS_1HR_UP	0.0015	0.03
KATSANOU_ELAVL1_TARGETS_DN	0.0015	0.03
NUYTTEN_EZH2_TARGETS_UP	0.0017	0.033
DOUGLAS_BMI1_TARGETS_UP	0.0018	0.034
BASAKI_YBX1_TARGETS_UP	0.0019	0.034
UDAYAKUMAR_MED1_TARGETS_DN	0.0019	0.034
ABDELMOHSEN_ELAVL4_TARGETS	0.0019	0.034
KIM_WT1_TARGETS_12HR_DN	0.0021	0.038
FORTSCHEGGER_PHF8_TARGETS_DN	0.0022	0.038
ROZANOV_MMP14_TARGETS_UP	0.0024	0.04
TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_HSC_DN	0.0024	0.04
OSADA_ASCL1_TARGETS_UP	0.0026	0.043
MCCABE_BOUND_BY_HOXC6	0.0029	0.046
TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_16D_DN	0.003	0.046
BENPORATH_SUZ12_TARGETS	0.003	0.047
HUANG_GATA2_TARGETS_DN	0.0031	0.047