NORTHWESTERN UNIVERSITY

The COMPASS family of histone H3 lysine 4 methyltransferases in transcriptional regulation, stem cell pluripotency, and development

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Driskill Graduate Training Program in Life Sciences

By

Christie C. Sze

EVANSTON, ILLINOIS

June 2020

Abstract

The evolutionarily conserved COMPASS family of methyltransferases implements histone H3 lysine 4 (H3K4) methylation, an epigenetic mark associated with transcriptional activation. Given the high mutational prevalence of COMPASS subunits across cancers and neurodevelopmental disorders, understanding COMPASS function would lend important insights into disease pathogenesis to facilitate development of effective therapies. The H3K4 methylase Set1A is one of six COMPASS members identified in mammals, and possesses an enzymatic SET domain responsible for genome-wide H3K4 di- and tri-methylation (H3K4me2 and H3K4me3 respectively). Previous studies demonstrated that loss of full-length Set1A resulted in embryonic lethality and embryonic stem cell (ESC) viability defects; however, the dependency of ESC pluripotency on H3K4 methylation by Set1A was not investigated. As shown in this dissertation, the SET domain of Set1A was dispensable for ESC viability and self-renewal, although necessary for proper differentiation. In addition, deleting the Set1A SET domain did not perturb bulk H3K4me3, implicating possible compensatory roles played by other COMPASS methyltransferases. By investigating a series of ESC lines harboring compounding mutations of the COMPASS enzymes, an elaborate relationship was unveiled: despite the differential regulation of H3K4me3 deposition and peak breadth by Set1A-B vs. Mll2, Mll2 could still help preserve global H3K4me3 level and breadth in the absence of Set1A-B. These findings illustrate the biological flexibility of such enzymes in transcriptional regulation to ensure cell viability. Finally, ongoing efforts to elucidate the critical role of Set1A in ESC pluripotency include leveraging a CRISPR/Cas9 dropout screen to identify targets that functionally interact with Set1A$^{\Delta SET}$. These efforts point to an uncharacterized interplay between COMPASS and another family of chromatin modifiers in regulating ESC viability. Taken together, this dissertation

provides novel insights into the epigenetic complexities in regulating stem cell pluripotency and development, which will ultimately facilitate effective treatment development.

To my parents, my siblings, my friends, and my husband,
for their infinite patience, love, and support throughout graduate school.

Acknowledgements

They say it takes a village to achieve a feat—indeed, successfully completing my predoctoral studies would not have been possible without the immeasurable support and invaluable guidance from my professors, colleagues, friends, and family. I would like to spend the next several pages to express my gratitude to these important individuals.

- **Dr. Ali Shilatifard**: Thank you for granting me the fortunate opportunity to develop my scientific thinking under your extraordinary tutelage. I am forever grateful for all the training, mentorship, support, and feedback you have provided, from devising impactful research projects to writing successful NIH grants. Your passion and enthusiasm for science is inspiring, your perpetual encouragement for me to keep challenging myself and never to give up, together have pushed me to grow both as a scientist and as a person. I am also thankful for the rigorous yet collaborative lab environment you fostered, where I have always been able to engage with other brilliant colleagues over the past several years.
- **Drs. Ann Harris, Jaehyuk Choi, Marc Mendillo, and Shelley Berger**: I am incredibly lucky to have such supportive committee members. Thank you for the constant feedback, during and outside of committee meetings, to improve the scientific rigor of my work and to ensure I stay on track. I especially would like to thank Ann, for being a wonderful chair.
- **Dr. Kaixiang Cao**: I honestly would not have been able to get this far in my scientific career if it were not for the amount of dedication, patience, and time Kai has put into teaching me the day-to-day skills since I joined the Shilatifard lab, which include tricky ESC culturing and line generation, basic molecular cloning, paper discussions, and manuscript writing. Thank you for always reminding me to keep going even during rough times.

- **Dr. Bercin Cenik**: I did not think we would get along so incredibly well since you joined the lab that you are practically my graduate school twin. Thank you for being a constant source of emotional support (and desk- and bay-mate), and I am really glad that we finally have the opportunity to have a fun scientific collaboration.
- **Stacy Marshall**: When I first joined the lab, you were among the first few who genuinely welcomed me while you were lab manager. Since then, our friendship has blossomed, and you have always been there for me. I miss our occasional apple cider breaks and will definitely miss our weekly climbing sessions.
- **All other previous and current members of the Shilatifard lab, with specific mentions**: **Laura Shilatifard, Beverly Kirk, Dr. Lu Wang, Dr. Noah Birch, Dr. Michal Ugarenko, Caila Ryan, Emily Rendleman, Dr. Marc Morgan, Dr. Delphine Douillet, Dr. Edwin Smith, Dr. Fei X. Chen, Patrick Ozark, Sid Das, Bin Zheng, Didi Zha, and Avani Shah**: Every one of you has been inspirational and supportive during the last several years, allowing me to flourish both scientifically and emotionally as a graduate student. I will miss all of our social gatherings, one-on-one chats, and breakroom lunches; of course, I will miss all of you.
- **Department of Biochemistry and Molecular Genetics/Simpson Querrey Center for Epigenetics**: I have had the honor to engage with many brilliant colleagues here at Northwestern University, especially those in the BMG Department/SQE. It has been an absolute delight to be able to discuss research ideas, current events, and life goals with you.
- **Driskill Graduate Program and Northwestern University, especially Drs. Steve Anderson and Pamela Carpentier**: I am very grateful to you for all the support and advice. Thank you for always being available to answer my many questions and for endlessly providing reassurance throughout my time as a graduate student.

- **Drs. Krithika Ramachandran and Rajita Vatapalli**: My two graduate school partners-in-crime, who have both begun this journey with me since the very beginning. When we started classes, I often panicked on how to study for tests - these two were always happily and patiently navigating me out of my struggles until classes became easy and fun. Even after classes, they are always there throughout my professional and personal milestones. I would not have been able to finish this Ph.D. without them.
- **My roommate Karen DeRocher**: Honestly the best roommate I could ever ask for; thank you for always being there for me (especially after Garrett moved out early). All the hours sharing grad school successes and joys, while commiserating being in long-distance relationships, cheering each other on as we are both inching towards the finish line, plus other wonderful memories. I will miss you very much.
- **Additional graduate school friends, to name a few**: **Kevin Park, Dr. Nimrod Miller, Tianjiao Sun, Hayk Yegoryan, Dr. Shawn Chen, Dr. Ashwin Narayanan, and CSHL friends**: Thank you for all the wonderful memories and sharing your delights and pains as we push forward with our professional and personal decisions and developments.
- **Wellesley Sisters, especially Stefanie Chan and Dr. Aabha Sharma**: Thank you both for being there for me during my journey through grad school. I am extremely grateful for the support, love, and compassion from my Wellesley Sisters, especially those who have gone/are going through a similar undertaking after college.
- **Former colleagues from Clarion, especially Dr. Dennis Chang, Shlomo M. Harnas, Dr. Uciane Scarlett, and Dr. Vineet Prabhu**: I am truly blessed to have gotten to know you during my Clarion days, and that we have continued to maintain our connections since. You

have evolved into becoming my life coaches and were all early influences of my grad school experience. Thank you for listening to me and for all the valuable life advice.

- **Undergraduate mentors Drs. Yuichiro Suzuki, Laura Johnston, and Michelle LaBonte**: Thank you for providing the fundamental scientific training and support during my undergrad years that inspired me to pursue grad school. Yui, thank you for your continuing encouragement through my grad school years.
- **Friends from high school (i.e. DNA Sisters) and college, especially Dr. Tiffany Cheng, Ting Poon, J.D., Dr. Diana Cai, and Dr. Jennifer Gillman**: Thank you for being there for me, especially during the recent tough times. I am forever grateful for your help and advice.
- **My in-laws (Mom, Dad, Tiffany, and Wendy)**: Thank you for welcoming me as part of your family and for your earnest support, both before and after I married Garrett.
- **Vivian and Edwin**: My siblings have always been my biggest cheerleaders, both before and during grad school. Even though I am the oldest, Vivian and Edwin have always encouraged me to be the best I can and be optimistic. I most certainly would not have been able to get through grad school if it weren't for their unwavering encouragement and love.
- **Mom and Dad**: Thank you for all of your sacrifices that have made me the person I am today. Your unconditional love, understanding, and support have always kept me motivated, and I hope to continue to make you proud.
- **Garrett**: Thank you for being my bedrock, confidante, and lifelong best friend. You have always embraced me wholeheartedly, (including my paranoia and OCD), and I am truly lucky to have you in my life. Your infinite love, selflessness, and wisdom has helped me become a better scientist and person.

Lastly, I would like to add that at the time of writing this dissertation (Spring 2020), our country and the rest of the world are in the midst of fighting the COVID-19 pandemic. Many states, including Illinois, have instated stay-at-home orders to help slow the spread of the contagious virus. As a result, my working from home to write a Ph.D. dissertation has been a peculiar blessing of this state-wide lockdown. In all seriousness, I especially would like to extend my utmost gratitude and respect to all the front-line fighters and essential staff for your compassion and bravery in keeping our country safe and running. Thank you.

Table of Contents

List of figures

Abbreviations

H3K4 – Histone H3 lysine 4

ESC – Embryonic stem cell

RNAP II / Pol II – RNA polymerase II

KMT – Lysine methyltransferase

KDM – Lysine demethylase

SAM – S-adenosyl-L-methionine

COMPASS – Complex of Proteins Associated with Set1

Trx – Trithorax

ChIP – Chromatin immunoprecipitation

TSS – Transcription start site

PHD – Plant homeodomain

HAT – Histone acetyltransferase

HDAC – Histone deacetylase

Trr – Trithorax-related

RRM – RNA recognition motif

FYR – FY-rich

HMG – High mobility group

KO – Knockout

FLOS – Functional location on Set1A

PRE – Polycomb response element

MEF – Mouse embryonic fibroblast

WT – Wild-type

gDNA – Genomic DNA

cDNA – Complementary DNA

RPM – Reads per million

FPKM – Fragments per kilobase of exon per million

EB – Embryoid body

GO – Gene ontology

NTC – Non-template control

RA – Retinoic acid

RPKM – Reads per kilobase of transcript per million mapped reads

CPM – Counts per million

NGS – Next generation sequencing

OE – Overexpression

shRNA – short hairpin RNA

NT – nontargeting

MOI – Multiplicity of infection

DDS – Differential dependency score

AP – Alkaline phosphatase

HSC – Hematopoietic stem cell

NSC – Neural stem cell

LSC – Leukemic stem cell

AID – Auxin inducible degradation

FBS – Fetal bovine serum

1 Introduction

Parts of this chapter, including figures, have been reproduced, with or without modifications, from my published review: Christie C. Sze and Ali Shilatifard. Cold Spring Harb Perspect Med. 2016, 6(11):a026427.

1.1 Epigenetics underlying gene expression regulation

Nucleosomes compose the eukaryotic chromatin core and consist of DNA wrapped around an octamer of four histone proteins (H2A, H2B, H3, and H4) (*1*). The amino- and carboxy-terminal tails of histones that protrude from the core nucleosome particle are subject to a diverse array of post-translational modifications, including methylation, acetylation, phosphorylation, and ubiquitination, that regulate various cellular processes including transcription, DNA repair and cell cycle progression (*2, 3*). These modifications serve as recognition sites for transcription factors and RNA polymerase II (RNAP II) to ensure proper gene expression and directly alter local chromatin structure; otherwise, aberrant deposition of these histone marks would result in defective development and disease.

Histone methylation, which occurs on basic residues, i.e. lysines (K), arginines (R), and histidines (H), has emerged as an essential modification involved with transcriptional regulation and development (*2*). Whether a methyl mark is associated with transcriptional activation or repression is specific to the residue, type of histone, and degree of methylation (*2, 3*). Multiple histone lysine residues that can be mono-methylated (me1), di-methylated (me2), and tri-methylated (me3) have been characterized to date: K4, K9, K27, K36, and K79 on histone H3, and K20 on histone H4 (*2, 3*). Among these residues, H3K4, H3K36, and H3K79 are generally associated with active transcription, while H3K9, H3K27, and H4K20 are primarily connected to transcriptional repression (*2, 3*). Each methylation state at the individual lysines is established

and dynamically modulated by two groups of highly conserved enzymes: histone lysine methyltransferases (KMTs) and demethylases (KDMs), which add and remove methyl marks at specific sites respectively (*2, 3*). Most KMTs identified thus far contain a SET domain, which is the enzymatic region that uses S-adenosyl-L-methionine (SAM) as the methyl group donor and catalytically adds the methyl groups to histones (*4*). My dissertation centers around a specific family of KMTs, namely the COMPASS (COMplex of Proteins ASsociated with Set1) family of H3K4 methyltransferases, whose relevance in development and disease has been increasingly studied since the fortuitous discovery of a spontaneous mutation of its founding member Trithorax (Trx) in *Drosophila* that resulted in homeotic transformations (*5*). An overview of the COMPASS family and their biological impact is discussed further below in Section 1.3.

The development of highly specific antibodies against various histone marks for genome-wide chromatin immunoprecipitation (ChIP) studies has led to the identification of chromatin signatures that imply local chromatin biology and transcriptional state. For example, H3K4me3 is a mark typically associated with transcriptional activation, marking promoters or transcription start sites (TSS) (*6-8*). H3K4me3 can be recognized by certain structural domains, e.g. plant homeodomain (PHD) fingers and chromodomains, found on reader proteins, thereby dictating the localization of effector proteins such as chromatin remodelers and other histone-modifying enzymes to impact local chromatin dynamics (*9*). Other methylation marks enriched at particular regions have also been determined: H3K4me1 at enhancers (which are noncoding regulatory sequences that govern spatiotemporal patterns of gene expression by heightening the rate of transcription of target genes (*10-12*)), H3K36 and H3K79 methylations mainly at gene bodies (*13, 14*), and H3K9me3 as a constitutive heterochromatic feature (*15*). Furthermore, a unique

group of promoters are bivalently decorated with the active H3K4me3 and repressive H3K27me3 marks (*16*). These bivalent genes are further discussed in Section 1.2.

Another key post-translational modification is histone acetylation, which are implemented by histone acetyltransferases (HATs) and removed by histone deacetylases (HDACs). HATs catalyze by transferring an acetyl group from acetyl-CoA to the ε-amino acid side chain of certain lysine residues (*17*). Lysine acetylation neutralizes the positive charge on the histone tail, which consequently loosens intra- and inter-nucleosomal interactions to promote chromatin accessibility for transcriptional activation (*18, 19*). Similar to methylation, lysine acetylation can also serve as docking sites for other reader proteins containing certain motifs (e.g. bromodomains), and can be found at both TSS and non-TSS elements. Akin to the marking of bivalent genes, the presence of acetylated H3K27 (H3K27ac) and H3K4me1 can distinguish active from inactive (or poised) enhancers respectively (*12, 20-22*).

It is worth mentioning that chromatin is usually marked by different combinations of post-translational modifications, indicating that there is a collaborative effort among various chromatin modifiers and remodelers to ultimately attain a coordinated transcriptional and cellular response. A notable example of this histone crosstalk is that mono-ubiquitination of histone H2B is required for proper H3K4 methylation (*23-25*). Crosstalk can also occur at a non-histone level, where the non-enzymatic function of a chromatin modifier can influence the recruitment and activity of another chromatin regulator in context-dependent situations (*26, 27*). These findings collectively indicate that instead of a “code”, the interplay among the chromatin regulators and their relevant histone marks comprise a sophisticated chromatin language that guides transcriptional regulation (*28*).

1.2 Harnessing the full power of ESCs requires understanding its chromatin landscape

Derived from the inner cell mass of mouse blastocysts, embryonic stem cells (ESCs) are pluripotent stem cells that retain the ability to self-renew and have the capacity to differentiate into all cell lineages from the three germ layers (endoderm, mesoderm, and ectoderm) (*29*). Tremendous interest in ESCs stems from its broad applicability. In basic research, ESCs are used 1) to study the epigenetic landscape in stem cells, 2) for differentiation in culture to mimic *in vivo* development, especially as an alternative to overcome the hurdles of embryonic lethality and to study cell fate decisions, and 3) for its unique utility in mouse model generation owing to its capability of germline transmission (*30-33*). In the clinic, ESCs can be used as a valuable platform for disease modeling and drug discovery and provide the promise of regenerative medicine in the form of cell replacement therapies. Furthermore, the discovery of somatic cell reprogramming by ectopic expression of the four Yamanaka factors (Oct4, Sox2, Klf4, and c-Myc) has equipped the stem cell field with more powerful research tools and clinical potential (*34*). Therefore, understanding the regulatory mechanisms governing pluripotency is critical to ensure careful application of pluripotent stem cells for therapeutic purposes.

The classical method for growing ESCs has been to co-culture ESCs on top of a layer of mitotically-inactivated feeder cells, which provide trophic factors to ensure ESCs in an undifferentiated state, in the presence of serum-containing medium. Addition of the cytokine leukemia inhibitory factor (LIF), which inhibits Stat3 signaling, to the medium also helps to maintain self-renewal and prevent spontaneous differentiation. However, maintaining ESCs in serum-conditions results in heterogeneous cell morphology and mosaic expression of pluripotency factors (*35*). As a result, serum-free culture conditions were developed: specifically,

in addition to LIF, the use of two small molecule kinase inhibitors against Mapk and Gsk3 signaling (known as "2i" inhibitors) were identified to serve as therapeutic approaches to prevent unintended differentiation. These culture conditions enable ESCs to attain ground state pluripotency, which more closely imitates the inner cell mass environment, with ESCs achieving a relatively more homogenous expression of pluripotency factors (*36*).

ESC pluripotency is mediated by the intricate circuitry of transcription factors and chromatin regulators that must be tightly controlled to hinder premature differentiation (*37, 38*). Studies have shown that ESCs maintain an "open" chromatin state to foster a transcriptionally permissive environment, which would allow rapid activation of transcriptional programs upon receiving differentiation cues (*39*). As pluripotent stem cells undergo differentiation towards lineage commitment, chromatin becomes increasingly condensed, or more heterochromatic in nature (*39, 40*). Notably, these structural changes in chromatin are seen *in vivo*, as pluripotent cells in the inner cell mass have less compact chromatin than the lineage-committed cells (*41*). Extensive characterization of the ESC epigenome revealed the following: ESCs are generally more abundant in chromatin remodeling complexes and active histone marks, e.g., H3K4me3 and acetylation of H3 and H4 histones, with low H3K27me3 and DNA methylation levels (*31, 42*). Meanwhile, lineage-committed cells have increased levels of repressive marks (e.g., H3K27me3, H3K9me3, and CpG DNA methylation) genome-wide (*43-45*). Furthermore, the concept of "bivalency" was developed to explain how developmental genes in ESCs are silenced but poised for activation upon triggering of differentiation (*16*). These lineage-specific genes have bivalent promoters, which are decorated by both H3K4me3 and H3K27me3, and are quickly expressed during differentiation through loss of H3K27me3 (*16, 46*). These findings

collectively demonstrate the tight regulation of transcriptional programs underlying ESC pluripotency, and that thorough understanding of the regulatory proteins involved is fundamentally important. Figure 1.1 below summarizes the key transcriptional and epigenetic changes as pluripotent stem cells ultimately progress to terminal differentiation.

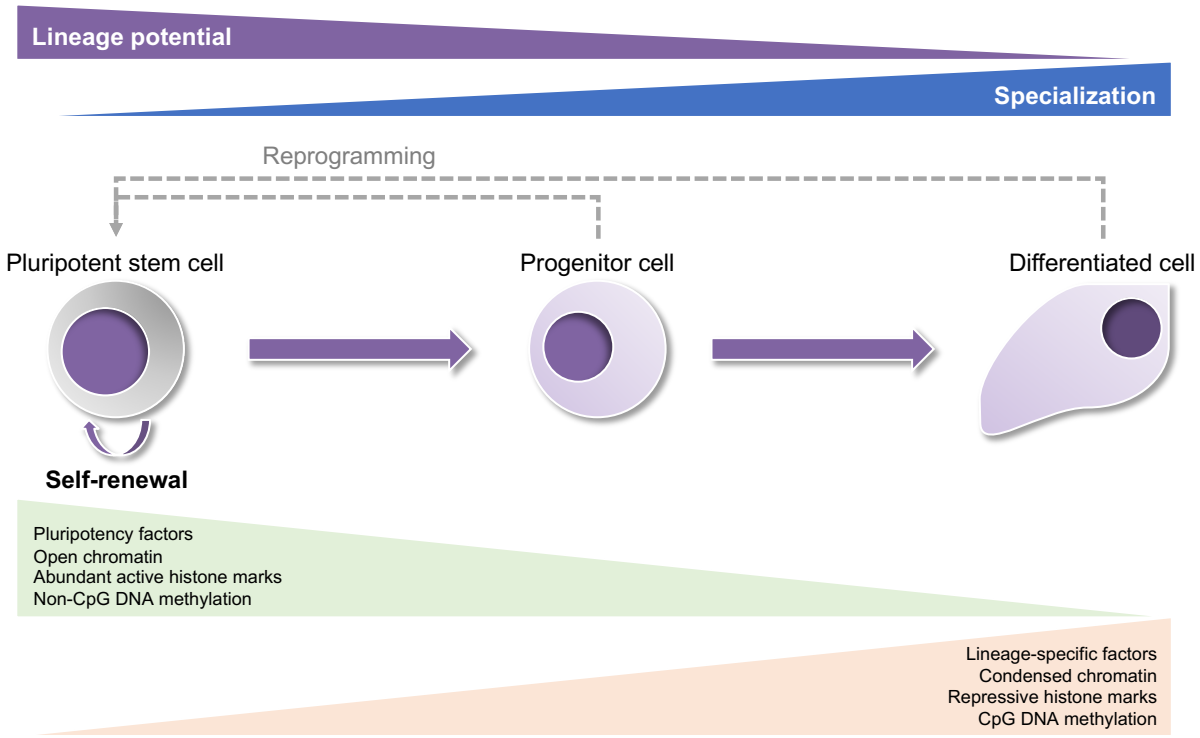

Figure 1.1. High-level overview of transcriptional and epigenetic landscape of pluripotent vs. differentiated cells.

Summary schematic illustrating the key transcriptional and epigenetic modifications underlying pluripotent stem cells and their progression towards terminal differentiation. Pluripotent stem cells, which encompass ESCs, generally have a transcriptionally permissive environment, and lineage-specific factors are in a silenced, but primed state. As pluripotent stem cells progress towards a more restrictive potency state, the chromatin structure becomes less hyperdynamic and more condensed, with an increase in expression of lineage-specific factors. Figure is adapted from CL Fisher and AG Fisher, 2011 (*47*).

1.3 The COMPASS family of histone H3 lysine 4 methyltransferases

The evolutionarily conserved COMPASS family of methyltransferases deposit methylation at H3K4 (H3K4me), a mark as discussed in Section 1.1 is typically associated with transcriptionally active chromatin. Interest in this family accelerated when MLL (mixed lineage leukemia) was discovered as an oncogenic fusion protein resulting from seemingly random translocations in patients with hematological malignancies (*48-51*), spearheading early efforts that focused on isolating MLL-containing complexes to understand the biochemical properties, functions, and regulation of MLL under normal conditions. An ancestral homolog of MLL, Set1, was identified in the budding yeast *Saccharomyces cerevisiae* and found to exist in a macromolecular complex that was named COMPASS (*52, 53*). Subsequent studies revealed a diverse family of COMPASS enzymes in metazoans. While yeast only has a single Set1/COMPASS capable of catalyzing mono-, di-, and tri-methylation on H3K4 (H3K4me1, H3K4me2, and H3K4me3, respectively) (*54*), *Drosophila* has three H3K4 methyltransferases: dSet1, Trithorax (Trx), and Trithorax-related (Trr) (*55, 56*). Mammals have two paralogs corresponding to each *Drosophila* member: Set1A and Set1B, orthologous to dSet1; Mll1 and Mll2, orthologous to Trx; and Mll3 and Mll4, orthologous to Trr (*57, 58*). The *Drosophila* and mammalian methylases also reside in COMPASS-like complexes, shown through ensuing studies to contain: 1) core components critical for enzymatic activity (Figure 1.2, dark blue), and 2) specific subunits believed to confer functional uniqueness to each complex (Figure 1.2, green) (*56, 58, 59*).

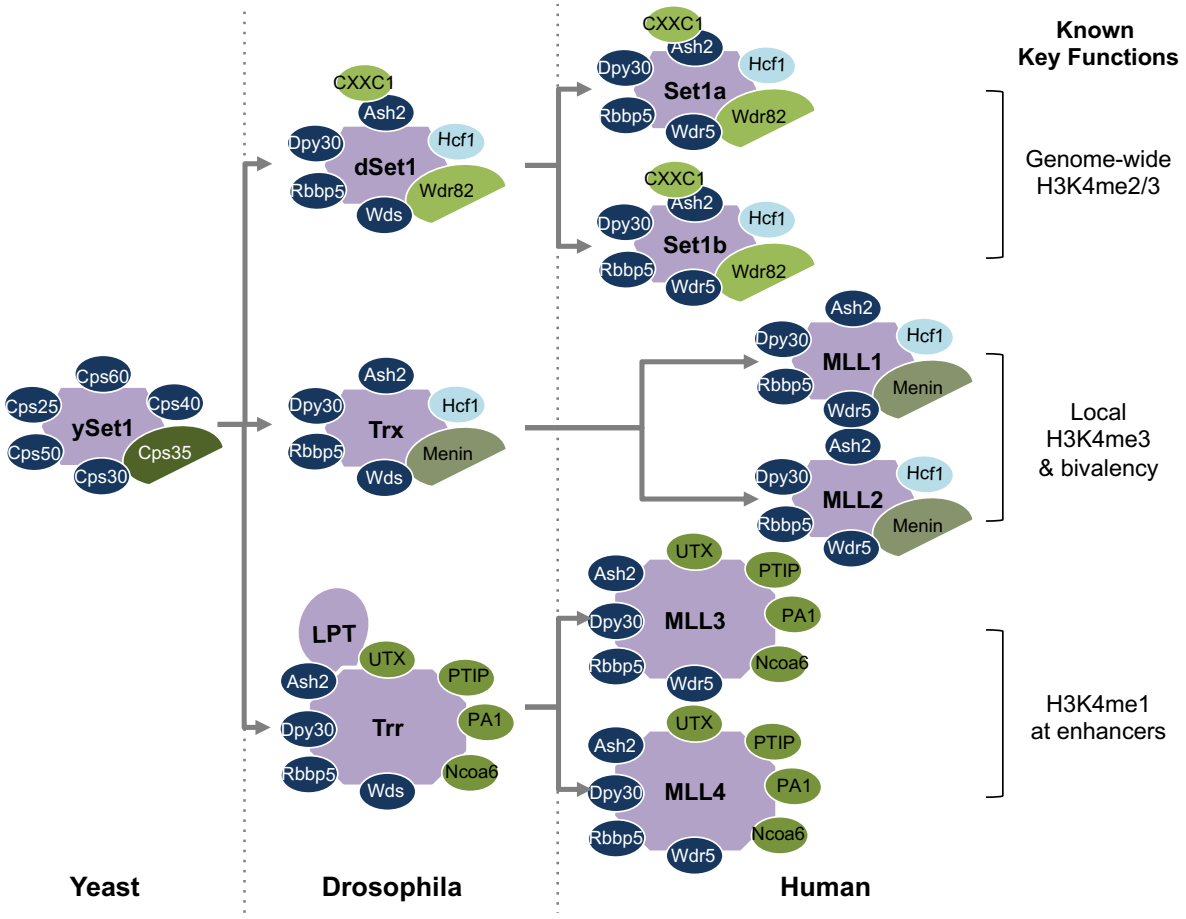

Figure 1.2. The COMPASS family of H3K4 methyltransferases in yeast, flies, and humans.

In yeast, there is only one Set1 methyltransferase capable of methylating histone H3K4. Flies have three COMPASS family members: dSet1, trithorax (Trx), and trithorax-related (Trr). Mammals have two paralogs for each of the three fly members for a total of six COMPASS members. Core subunits found in all COMPASS complexes are highlighted in dark blue, while subunits specific to the complex are marked in green. Hcf1 (light blue) is reportedly specific to be in Set1 and Trx branches but not in the Trr complex (*60*). Key methylation functions known to date for each branch of COMPASS are noted.

The methylase subunits of COMPASS all possess a catalytic 130 amino acid-long C-terminal motif called the SET domain, named after the *Drosophila* proteins Su(var)3-9, Enhancer of zeste [E(z)], and Trithorax (Trx) (*61, 62*). The SET domain enables COMPASS and other proteins containing this enzymatic motif to use S-adenosyl-L-methionine (SAM) as the methyl group donor and catalytically methylate lysine substrates (*4, 57*). In contrast, regions N-terminal to the SET domain differ across family members (Figure 1.3) (*63*). Mammalian Set1A and Set1B each have an N-terminal RNA recognition motif (RRM) and an N-SET domain directly preceding and juxtaposing the SET domain, while mammalian Mll1-4 contain varying numbers of PHD fingers, FY-rich (FYR) domains, and DNA-binding motifs such as AT-hooks, high mobility group (HMG) boxes, and CXXC domains (Figure 1.3) (*63, 64*).

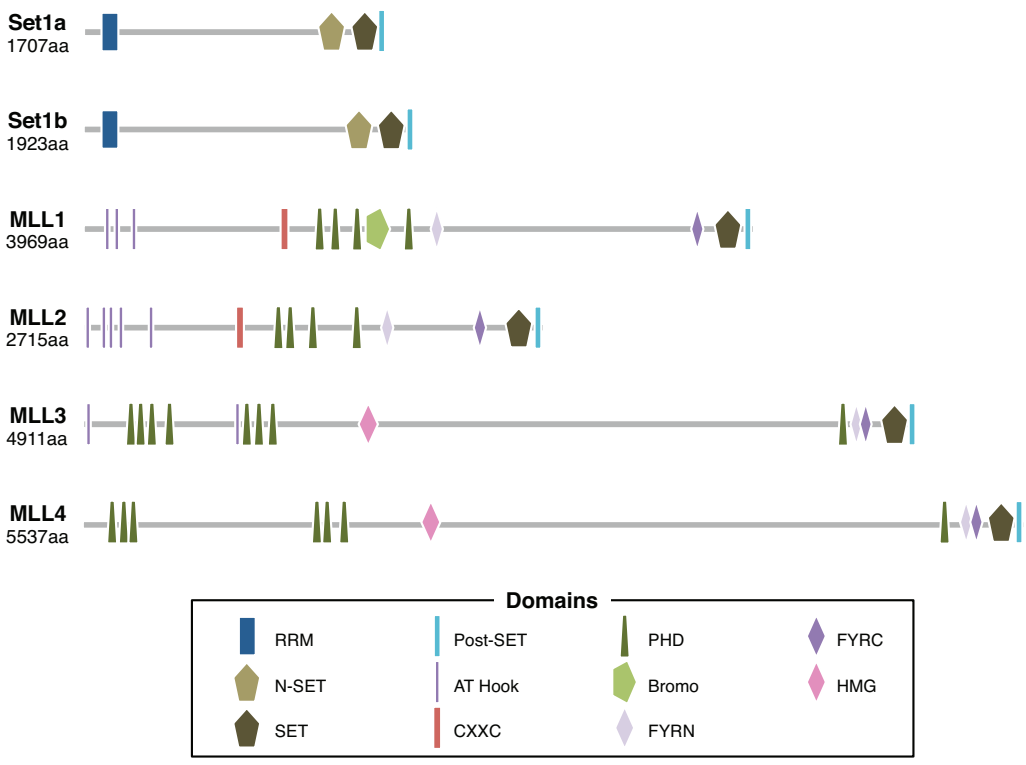

Figure 1.3. Known domain organization of COMPASS family members in humans.

Annotation of each domain structure follows SMART (http://smart.embl-heidelberg.de) (*65, 66*) utilizing protein sequences obtained from NCBI as accessed on January 8, 2016. Names for illustrated domains are specified in the box labeled "Domains." All COMPASS family members possess the highly conserved SET and post-SET domains at the C-terminus. Meanwhile, the N-terminus vastly varies across the subfamilies. Set1A and Set1B have N-terminal RNA recognition motifs (RRM) and an N-SET domain juxtaposing the SET domain. MLL-related proteins have several plant homeodomain (PHD) fingers and other domains associated with chromatin binding (e.g., AT-hooks, high mobility group (HMG) boxes, and CXXC domains). Mll1-4 methylases also have FY-rich (FYR) motifs, where Mll1 and Mll2 have FYRN and FYRC regions distant from each other, while Mll3 and Mll4 have such regions adjacent to each other. The diversity in domains contributes to the binding and functional properties of the COMPASS complexes.

Such domain architectural variability denotes functional and recruitment diversification of the COMPASS family. Indeed, accumulating evidence points to a model where H3K4 methylation responsibilities are divided among the COMPASS members in higher metazoans to ensure proper transcriptional modulation. Studies have shown that dSet1/Set1A/Set1B are responsible for genome-wide H3K4me2/me3 (*56, 64, 67, 68*), while Trx/Mll1/Mll2 catalyze H3K4me3 at specific loci, including the *Hox* gene (*69*) and bivalent promoters (marked by concurrent tri-methylation of H3K4 and H3K27 and poised to express developmental genes) in ESCs (*70*). As mentioned earlier, Trx was initially discovered to regulate *Hox* gene expression in *Drosophila*, specifically required for maintaining *Hox* gene activation (*5*). It is through shared protein homology with Trx that Trr was cloned (*71*). Trr and its mammalian homologs Mll3/Mll4 have been accredited as key H3K4 mono-methyltransferases at enhancers (*72-74*).

Unlike the division of their H3K4 methylation responsibilities, the recruitment of the COMPASS complexes to chromatin is not as well understood. Several possible mechanisms have been proposed regarding how the COMPASS complexes are selectively recruited to

specific loci. One frequently studied mechanism is sequence-specific targeting through the CXXC motif, which is found within the Cxxc1 protein, a key subunit in the Set1/COMPASS complexes (Figure 1.2), as well as the Mll1 and Mll2 enzymes (Figure 1.3). Although Cxxc1, Mll1, and Mll2 all recognize and bind to unmodified CpG-containing DNA via their CXXC domain, studies have shown that these CXXC domains have distinct DNA-binding specificities, therefore contributing to the differential targeting of Set1A-B/COMPASS and Mll1-2/COMPASS complexes to their respective genomic loci (*75-78*). Another possible recruitment mechanism worth noting is COMPASS interaction with chromatin, specifically by associating with certain histone modifications. As illustrated in Figure 1.3, the Mll1-4 methyltransferases harbor varying numbers of PHD fingers, which are structurally conserved modules known to recognize methylated or unmethylated residues on histones H3 or H4 (*79*). It has been speculated that the COMPASS enzymes could be a part of a positive feedback loop by potentially recognizing and binding to methylated H3K4 to further propagate H3K4 methylation. To date, only Mll1 has been shown to be recruited through its PHD finger to the *HoxA9* gene by recognizing the H3K4me3 mark (*80*). It has also been reported that Mll4's tandem PHD fingers could bind to un- or di-methylated arginine on histone H4 to dictate the localization and function of Mll4 (*81*). Furthermore, Wdr82, a key component of the Set1/COMPASS complexes, has been found to associate with chromatin in a histone H2B ubiquitination-dependent manner (*67, 82*). Lastly, targeted recruitment of various COMPASS complexes to chromatin could be explained by their affinity with site-specific transcription factors, co-activators, and scaffolding proteins. Notable examples include the following: recruitment of Set1/COMPASS could be mediated through its association with RNAP II (*83-87*); interactions with the pluripotency factor

Oct4 could target COMPASS to chromatin (*88, 89*); and factors such as Menin and Ledgf have been reported to mediate Mll/COMPASS recruitment (*90-93*).

As elucidated above, numerous studies are providing key insights on how each COMPASS member functions and is localized in unique cellular contexts. However, precise mechanisms underlying the role of COMPASS in transcriptional and developmental regulation continue to remain elusive. My graduate research primarily focuses on the Set1A/Set1B branch of the COMPASS family, whose roles in development and disease are summarized in the subsequent sections. Brief highlights of the roles of other COMPASS subfamilies (i.e., Mll1-4) in development and disease will also be presented in the ensuing sections.

1.3.1 COMPASS in transcriptional regulation, stem cell pluripotency, and development

Set1A/Set1B—Set1A, and its subfamilial relative Set1B, were identified based on sequence homology to yeast Set1 (*64*). The function of Set1A and Set1B in mammalian development was first examined by using gene trapping and conditional knockout (KO) in mice to unearth their vastly differing roles in early embryonic development. Deletion of either Set1A or Set1B resulted in embryonic lethality: *Set1A*-KO embryos died around E7.5, while genetically removing *Set1B* allowed the embryos to survive until E11.5 (*94*). Phenotypic analyses revealed that Set1A functions shortly after inner cell mass formation but pre-gastrulation, while Set1B is required post-gastrulation (*94*). Meanwhile, a follow up study found that maternal Set1B, but not Set1A, is necessary for progression through the first cell division prior to zygotic genome activation (*95*), delineating context-specific functions of COMPASS during early development.

Work in ESCs showed that Set1A, but not Set1B, is essential for ESC proliferation and self-renewal, for *Set1A*-KO ESCs could not be derived from blastocysts (*94*). Conditional Set1A deletion in ESCs increased G1, but decreased S phase, while conditional Set1B loss had no cell cycle effects (*94*). Set1A loss globally reduced H3K4me1/me2/me3 in ESCs and embryos, but Set1B deletion did not affect bulk H3K4 methylation (*94*). Similar findings regarding Set1A's role in ESCs and embryos were reported in an independent study (*89*). Set1A reportedly interacts with key pluripotent factor Oct4 and is recruited to target promoters of pluripotency genes for transcriptional activation via H3K4 methylation (*89*). Furthermore, the involvement of Set1A in cellular differentiation has been explored in several cellular contexts: Set1A is involved in hematopoietic lineage differentiation in culture and B-cell development *in vivo* (*96, 97*), and a study from our laboratory established a role for Set1A as a transcriptional activator of *Hox* gene expression during ESC differentiation (*98*).

Part of my graduate research, which will be elaborated in greater detail in the following chapter, demonstrated that the SET domain of Set1A is surprisingly dispensable for ESC self-renewal, although H3K4 methylation by Set1A is essential for proper differentiation (*99*). The study unveiled that the requirement of Set1A in ESC viability in the self-renewing state is independent of the catalytic domain. Another study in leukemia cells also showed that full-length Set1A is required for viability, although cell survival is not dependent on the enzymatic SET domain. Instead, the authors reported a newly identified N-terminal region named FLOS (Functional Location On Set1A), located downstream of the RRM domain, that is indispensable for viability and regulates the DNA damage response (*100*). Similarly, our laboratory has recently discovered a SET-independent function of Set1B: with its unique cytoplasmic-

interacting protein Bod1, Set1B is essential for suppressing fatty acid metabolism to promote cancer cell survival (*101*). Altogether, these studies underscore the biological roles of COMPASS beyond the enzymatic activity.

Mll1/Mll2—As stated previously, Mll1 and Mll2 are responsible for H3K4 methylation in a locus-specific manner. Mll1 deposits H3K4me2 and H3K4me3 at Polycomb response elements (PREs) (*102*) and gene promoters of the *Hoxa* and *Hoxc* clusters (*69*). The *Hox* gene clusters are evolutionarily conserved, important for determining segmental identity and body patterning. Deleting *Mll1* results in embryonic lethality at E10.5, owing to disrupted *Hox* gene expression causing mutant embryos to exhibit hematopoietic abnormalities (*103*). Interestingly, $Mll1^{\Delta SET}$ mutants are viable with defective skeletal development (*104*), once again manifesting the catalytic-dependent vs. independent aspects of COMPASS biological function.

Mll2 tri-methylates at bivalent promoters and a subset of non-TSS elements (*70, 75, 105*). In addition, Mll2 has been shown to be involved in neuronal transdifferentiation, and that Mll2 perturbation induces expressions of genes linked to dystonia (*106*). Unlike Mll1, Mll2 functions earlier in development, such that *Mll2*-null is embryonic lethal at E7.5 (*107*). Conditional mutagenesis determined Mll2 as crucial for germ cell development in mice of either sex (*108, 109*). Further investigation indicated a temporal requirement for Mll2 in development: Mll2 allegedly acts as the major H3K4 methyltransferase shortly after fertilization until the blastocyst stage (~E3.5), when Set1A takes over until Mll2, as well as Set1B, is required again post-gastrulation (*94, 108, 109*). These temporal switches of H3K4 depositors illustrate their cooperativity in regulating gene networks important for pluripotency and proper development.

Mll3/Mll4—Mll3/Mll4 have been accredited as key H3K4 mono-methyltransferases at enhancers, primarily implementing H3K4 mono-methylation at intergenic and intragenic regions (*72-74*). ChIP-seq studies revealed that Mll3/Mll4 bind to enhancer regions as well as TSSs, but depletion of Mll3/Mll4 resulted in genome-wide reduction of H3K4me1 primarily at enhancers (*72, 73*). Furthermore, loss of Mll3/Mll4 in mouse embryonic fibroblasts (MEFs) led to decreased H3K27ac with paralleled increase of H3K27me3 at putative enhancers (*72, 73*). Since H3K27ac and H3K27me marks are part of different enhancer chromatin signatures, these findings further implicated Mll3/Mll4 in enhancer regulation (*110*). Mll4, with partial redundancy with Mll3, was confirmed as a key mono-methylase at adipogenic and myogenic enhancers during differentiation, and other studies have implicated their role in regulating enhancer/promoter communication in macrophage activation, cardiac development, and B cell lymphomagenesis (*74, 111-113*). Unlike the divergent *in vivo* phenotypes seen with *Mll1*-KO vs. $Mll1^{\Delta SET}$, both *Mll4*-null and $Mll4^{\Delta SET}$ induced embryonic lethality at E10.5, potentially owing to the destabilization of the Mll4 protein upon SET domain removal *in vivo* (*114*). Interestingly, our laboratory demonstrated that mono-methylation at H3K4 by Trr, the fly ortholog to Mll3/Mll4, is dispensable for *Drosophila* viability (*115*). Furthermore, unlike *Mll4*-KO, *Mll3*-KO and $Mll4^{\Delta SET}$ ESCs do not exhibit naïve pluripotency withdrawal (*26, 98*).

1.3.2 COMPASS in disease

The studies discussed above illustrate the central roles that the COMPASS family and H3K4 methylation play in dynamic gene expression programs, stem cell pluripotency, and development. As a result, it is not surprising that multiple genome sequencing studies have

recently reported that various subunits of COMPASS are frequently mutated across cancers and neurodevelopmental disorders (*116-119*). In fact, as summarized in the subsequent sub-sections, the broad mutational landscape of COMPASS can lead to various disease pathologies, and ongoing research efforts in the field are beginning to provide key insights into the underlying mechanisms of COMPASS mis-regulation in pathogenesis.

Set1A/Set1B—Set1A is overexpressed in metastatic breast cancer, and its knockdown decreased metastases in nude mice (*120*). Set1A reportedly cooperates with a lncRNA to promote hepato-carcinogenesis (*121*), and regulates the cell cycle via miRNAs (*122*). Furthermore, several groups recently reported that Set1A coordinates DNA damage signaling and repair to ensure genomic integrity, although the mechanisms by which Set1A orchestrates this response varied across the studies (*100, 123-125*). Exome sequencing studies have also identified rare variants in *Set1A* in patients with schizophrenia, but none in the control population, implicating the potential role of Set1A in brain function (*118, 126, 127*). Relative to Set1A, Set1B is less well studied, though we and others have shown Set1B to have a role in cancer (*101, 128-130*). In fact, as mentioned earlier in Section 1.3.1, our laboratory recently discovered that Set1B, together with its cytoplasmic-interactor Bod1, regulates tumor metabolism in a non-catalytic manner (*101*). Two recent reports have also identified de novo *Set1B* variants in patients with epilepsy and autism (*119, 131*).

Mll1/Mll2—As introduced earlier in the overview section on COMPASS, Mll1 was first discovered in oncogenic fusion proteins resulting from chromosomal translocations associated with childhood leukemia (*58*). In addition, de novo loss-of-function heterozygous variants of *Mll1* are responsible for the Wiedemann-Steiner syndrome (*132*). Patients with this disorder have

poor muscle tone, intellectual disabilities, facial abnormalities, and short stature. Meanwhile, loss-of-function mutations of *Mll2* have been reported in patients with early-onset dystonia, a motor-related progressive disorder (*133, 134*).

Mll3/Mll4—Cancer genome sequencing studies have identified a myriad of somatic *Mll3*/*Mll4* mutations across different malignancies (*135-149*). In fact, extensive analyses of sequencing data revealed *Mll3*/*Mll4* mutations to be among the most frequent in human cancer (*116, 117*). Growing evidence suggests that Mll3/Mll4 are tumor suppressors (*113, 150-154*); closer observation of catalogued mutations of *Mll3* and *Mll4* data revealed a higher density of mutations in the N-terminal region of Mll3 containing clusters of PHD fingers, while mutations of *Mll4* are relatively more scattered throughout the protein (*59*). Our laboratory has recently discovered that the mutational hotspots found in *Mll3* perturbed the interaction between Mll3 and the tumor suppressor Bap1 to promote proliferation and survival of breast cancer cells (*154*). Interestingly, a few studies have pinpointed Mll4 in an oncogenic role indicating context-dependent Mll4 function in cancer (*155, 156*). In addition, loss-of-function mutations of *Mll4* results in the Kabuki syndrome, a neurodevelopmental disorder characterized by intellectual disability, skeletal abnormalities, and growth delays (*157, 158*). Meanwhile, loss-of-function heterozygous mutations of *Mll3* have been associated with Kleefstra syndrome, another neurodevelopmental disorder characterized by language and motor delays, autistic behavior, and intellectual disabilities (*159*). To date, functional studies investigating *Mll3* mutations in patient cells have yet to be conducted.

1.4 Dissertation objectives and overview

The COMPASS family of H3K4 methyltransferases, including the associating complex subunits, have been identified to be frequently mutated across various cancers and neurodevelopmental diseases. Therefore, understanding COMPASS function will lend important mechanistic insights into their role underlying disease pathogenesis, which is necessary for the development of effective therapeutics. The overall objective of my dissertation study is to elucidate the functional role of the COMPASS family of H3K4 methyltransferases in regulating ESC pluripotency, focusing primarily on Set1A/COMPASS. As discussed earlier in this chapter, Set1A is the only COMPASS methyltransferase whose full genetic knockout perturbs ESC viability, showcasing the importance of this protein in ESC pluripotency.

My dissertation is comprised of three main stories, detailed in Chapters 2-4. Chapter 2 discusses the discovery that the catalytic SET domain of Set1A was surprisingly dispensable for ESC viability and self-renewal, although necessary for proper ESC differentiation. The extended viability of deleting the SET domain of Set1A *in vivo* is also presented, along with the discovery that the SET domain of Set1A is seemingly crucial for proper mesoderm development. Furthermore, removing the SET domain of Set1A in ESCs did not perturb global H3K4 methylation, which suggested the plausibility that other COMPASS family members might play compensatory roles in sustaining global H3K4 methylation. This inspired my research detailed in Chapter 3, where I discuss generating a series of ESC lines harboring compounding mutations of the COMPASS family of methyltransferases to analyze the consequential effects of such alterations on H3K4 methylation. The systematic unraveling of the functional specialization and redundancy among the COMPASS members in regulating H3K4 methylation level and breadth

revealed their critical role in regulating transcriptional output and cellular identity. These results collectively illustrate the incredible extent of biological flexibility of such enzymes in regulating transcription in a context-dependent manner in stem cell maintenance. Following on these findings, current ongoing efforts aiming to elucidate the critical role of Set1A in ESC pluripotency are detailed in Chapter 4, where a genome-wide negative selection screen was employed to identify targets that functionally interact with a catalytically dead Set1A in regulating ESC viability. Altogether, this dissertation illuminates the basic functional significance of Set1A and unweaves the epigenetic complexities in regulating stem cell pluripotency and development. These findings could consequently be extrapolated to assist our understanding of the disease liability of Set1A/COMPASS and ultimately facilitate the development of effective targeted therapeutics.

2 Histone H3K4 methylation-dependent and independent functions of Set1A/COMPASS in embryonic stem cell pluripotency

Majority of the work from this chapter, including figures, has been reproduced, with or without modifications, from my publication: Christie C. Sze, Kaixiang Cao, Clayton K. Collings, Stacy A. Marshall, Emily J. Rendleman, Patrick A. Ozark, Fei Xavier Chen, Marc A. Morgan, Lu Wang, and Ali Shilatifard. Genes Dev. 2017, 31(17):1732-1737.

2.1 Introduction

The highly conserved COMPASS (COMplex of Proteins ASsociated with Set1) family of methylases implement methylation at lysine 4 of histone H3 (H3K4) (*58, 59*), a mark associated with transcriptionally active chromatin. The H3K4 methyltransferase Set1A is one of six COMPASS members identified in mammals, and has consistently been shown to deposit bulk H3K4 di- and trimethylation (H3K4me2/me3 respectively) across the genome (*56, 64, 67, 68, 89, 94, 98*). Set1A is critical for early mouse embryonic development, as *Set1A*-knockout (KO) results in early embryonic lethality at E7.5 (*94*). Defects exhibited in *Set1A*-KO embryos suggest that Set1A functions shortly after inner cell mass formation but before gastrulation (*94*). Unlike any of its other COMPASS familial relatives (*94, 98, 105, 160*), Set1A is essential for embryonic stem cell (ESC) viability: ESCs could not be derived from *Set1A*-KO blastocysts (*89, 94*) and homozygous *Set1A*-KO ESCs could not be generated via CRISPR/Cas9 (*98*). Depletion of *Set1A* hinders ESC proliferation and triggers their apoptosis (*89, 94, 98*), affirming the importance of Set1A in ESC survival and identity. The involvement of Set1A in cellular differentiation has been explored in several cellular contexts. Set1A reportedly mediates hematopoietic lineage differentiation in culture (*96*) and B-cell development in vivo (*97*). Our laboratory has also recently established a role for Set1A as a transcriptional activator of *Hox* gene expression during ESC differentiation (*98*). To date, we still have very limited knowledge of the role of Set1A in

regulating ESC pluripotency; specifically, whether the function of Set1A in ESC maintenance is dependent on its methylase activity remains elusive.

2.2 Results & discussion

To investigate the role of the enzymatic activity compared to the rest of the Set1A protein in ESC pluripotency, we began by determining if deletion of its C-terminal catalytic SET domain adversely affects ESC self-renewal and viability (Figure 2.1-2.2). Using CRISPR/Cas9, we deleted a ~1.7kb endogenous genomic sequence coding for the SET domain of Set1A (Figure 2.1B-C; Appendix 7.1). We identified two homozygous mutant clones containing the SET domain deletion of Set1A (*Set1A*$^{\Delta SET}$) via PCR genotyping, and further verified by Sanger sequencing and RNA sequencing (Figure 2.1B-C; Figure 2.2A). Sanger sequencing of both the genomic DNA and messenger RNA revealed that these mutant clones harbor the SET domain deletion that would result in the introduction of an early STOP codon, leading to protein truncation that also removes the post-SET domain (Figure 2.1A, C). Truncated Set1A protein stability is comparable to that of wildtype (WT) protein detected in the parental ESCs (Figure 2.2B). Remarkably, *Set1A*$^{\Delta SET}$ ESCs retain their self-renewal characteristics: mutant cells display a strikingly similar morphology to the WT cells, as evidenced by alkaline-phosphatase staining (Figure 2.2C). Expression of the principal pluripotency factors *Nanog*, *Oct4*, and *Sox2* is not altered in *Set1A*$^{\Delta SET}$ ESCs (Figure 2.1D; Figure 2.2D), and further global analyses demonstrated that the overall transcriptome is indeed similar (Figure 2.2E). Although deletion of *Set1A* was shown to perturb ESC pluripotency, our data show that the SET domain of Set1A is not required

for ESC self-renewal, and the self-renewal function associated with Set1A requires a region of the protein upstream of the catalytic domain.

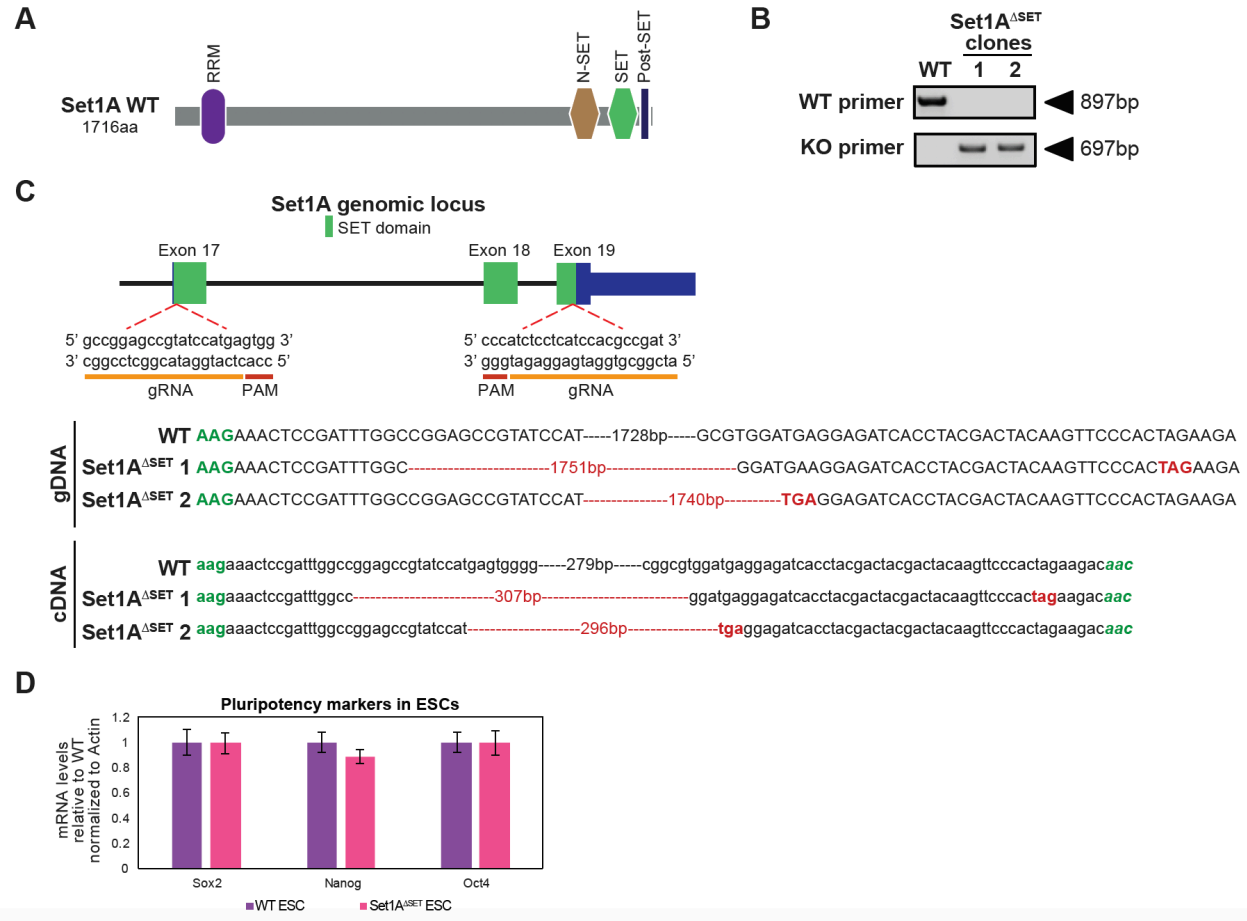

Figure 2.1. CRISPR/Cas9-generated *Set1A*$^{\Delta SET}$ ESCs.

(A) Diagram of the known domain organization of mouse Set1A protein. (B) PCR genotyping results of WT vs. *Set1A*$^{\Delta SET}$ ESCs. Arrowheads indicate base pair (bp) size of PCR products. (C) Top: Schematic of the *Set1A* genomic locus and the two CRISPR/Cas9 cut sites targeting the SET domain (green). gRNA sequences are in orange, and PAM sequences are in red. Bottom: Sanger sequencing of genomic DNA (gDNA) and complementary DNA (cDNA) revealed the precise sequences deleted (indicated in bp) in *Set1A*$^{\Delta SET}$ ESCs. Early STOP codon introduced (highlighted in red) as a result of SET domain deletion. Start and stop codons of SET domain are indicated in green. (D) qRT-PCR analysis of expression levels of pluripotency factors *Sox2*, *Nanog*, and *Oct4* in ESCs.

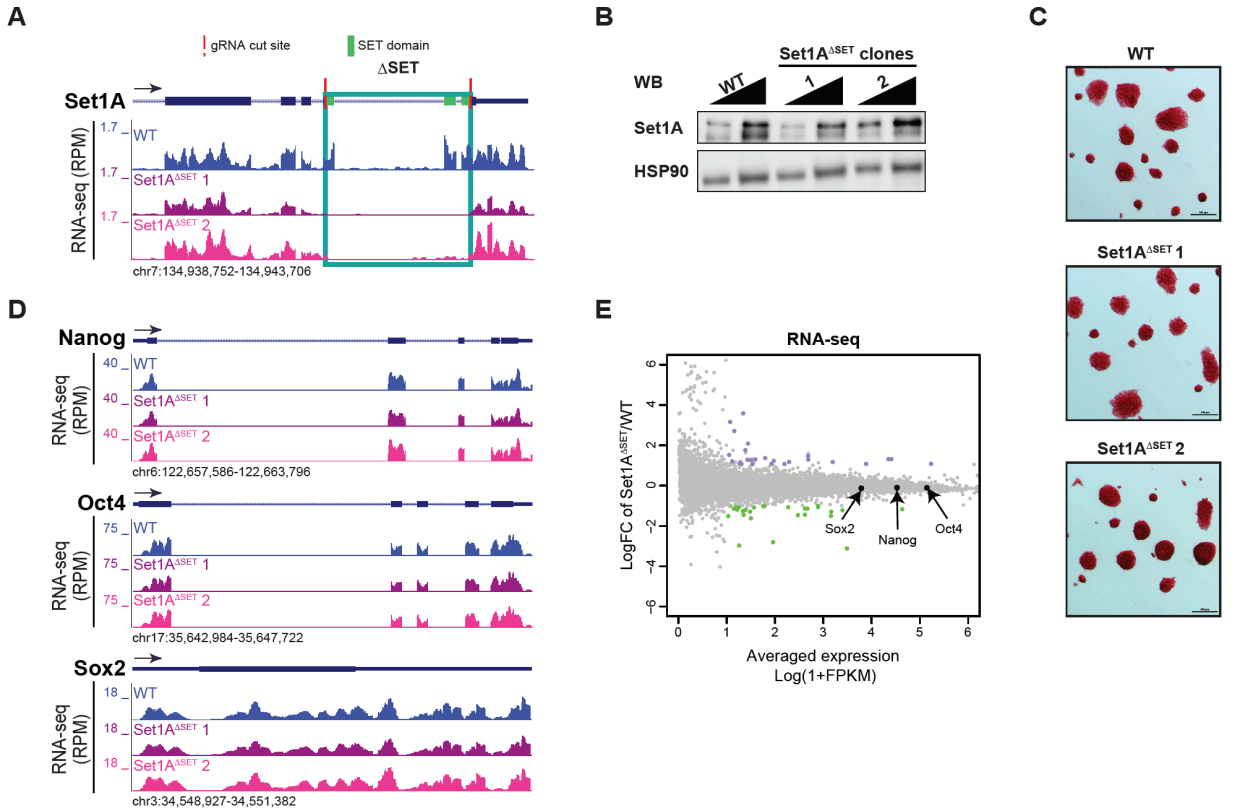

Figure 2.2. SET domain deletion of Set1A (*Set1A$^{\Delta SET}$*) does not perturb ESC self-renewal.

(A) RNA-seq results confirm CRISPR/Cas9-mediated deletion (green box) of the genomic sequence coding for SET domain of Set1A for two homozygous clones. RPM: reads per million. (B) Western blot of Set1A levels in WT and *Set1A$^{\Delta SET}$* ESCs, with HSP90 as the loading control. Samples were loaded at a 1:2. (C) Representative images of alkaline-phosphatase staining of WT and *Set1A$^{\Delta SET}$* ESC colonies. Black scale bar = 100µm. (D) RNA-seq tracks of pluripotency factors *Nanog*, *Oct4*, and *Sox2* between WT ESCs and mutant clones. (E) MA plot comparing the global transcriptome between parental and *Set1A$^{\Delta SET}$* ESCs. Expressions of pluripotency factors *Nanog*, *Oct4*, and *Sox2* are indicated. FPKM: fragments per kilobase of exon per million.

We assessed H3K4 methylation in the *Set1A*$^{\Delta SET}$ ESCs compared to that of parental WT cells. As observed by western blot, bulk H3K4me1/me2/me3 levels are comparable between WT and mutant ESCs (Figure 2.3A). To determine the changes in the levels and the pattern of H3K4me3 on chromosomes, ChIP-seq analyses of global H3K4 methylation in WT and *Set1A*$^{\Delta SET}$ clones were performed, and these data further support our findings (Figure 2.3B). To identify genomic sites with H3K4me3 changes specific to Set1A, we performed ChIP-seq of Set1A. A total of 10,240 Set1A binding regions were characterized, with promoter localization being the predominant annotation as calculated by HOMER (Figure 2.4A). Examination of genome-wide Set1A occupancy revealed similar Set1A binding in WT and *Set1A*$^{\Delta SET}$ ESCs (Figure 2.4B). We subsequently centered H3K4me3 peaks from WT and *Set1A*$^{\Delta SET}$ ESCs at Set1A binding regions. Specifically, we used K-means clustering to partition H3K4me3 occupancy at Set1A sites into three clusters for the parental cells and the *Set1A*$^{\Delta SET}$ clones (Figure 2.3C). The first cluster, containing loci with strongest Set1A binding, exhibits a detectable reduction in H3K4me3 peak occupancy in both *Set1A*$^{\Delta SET}$ mutants compared to WT cells, which is further illustrated in the differential heatmap for H3K4me3 occupancy (Figure 2.3C; Figure 2.4C). We examined the expression of the nearest genes for all three peak- clusters, and noted that cluster 1 coincides with the highest nearest gene expression (Figure 2.4D). Genomic Regions Enrichment of Annotations Tool (GREAT) (*161*) analysis showed that cluster 1 peaks are present primarily at regions closest to transcription start sites (TSS) (Figure 2.4E). Furthermore, despite the observed H3K4me3 decrease of cluster 1 peaks, there were no compelling changes in RNA polymerase II (Pol II) occupancy (Figure 2.3D). Representative track examples in Figure 2.3E provide a clear visualization of H3K4me3 decrease at the

promoters, though the accompanying RNA sequencing data show no substantial difference in the pattern of gene expression. We speculate the following: 1) residual H3K4me3 at these specific loci is adequate to maintain gene expression; or 2) Set1A$^{\Delta SET}$ may be affecting recruitment of other H3K4 methyltransferases in the COMPASS family, which we plan to explore in future studies. We currently have reason to believe that Set1B, a close homolog of Set1A, does not sufficiently compensate for Set1A$^{\Delta SET}$ ESCs to sustain global H3K4me3, because Set1B overexpression was unable to rescue defects caused upon loss of Set1A in ESCs (*94*). Nevertheless, the findings thus far communicate that the catalytic SET domain of Set1A and its associated H3K4me3 is dispensable for ESC survival and identity.

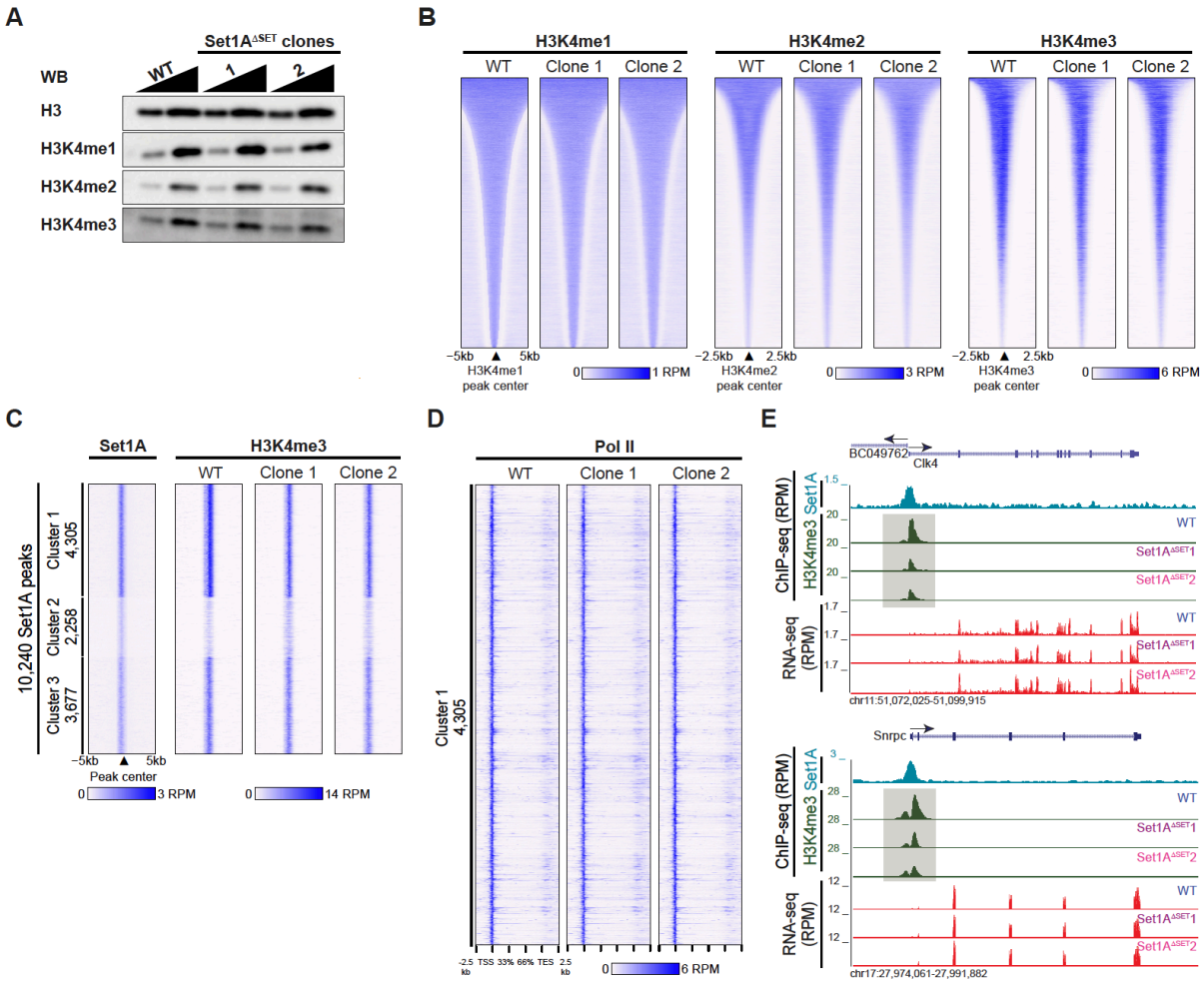

Figure 2.3. SET domain deletion of Set1A in undifferentiated ESCs resulted in decrease of H3K4me3 at specific sites.

(A) Western blot comparing H3K4 methylation levels in *Set1A*$^{\Delta SET}$ to parental WT cells. H3 served as the loading control. Samples were loaded at a 1:2. (B) Heatmaps of H3K4me1, H3K4me2, and H3K4me3 ChIP-seq occupancy levels in WT and *Set1A*$^{\Delta SET}$ cells. Occupancy levels were aligned to WT peaks sorted by decreasing peak width for each individual modification. (C) 10,240 Set1A peaks were called and partitioned into three groups by K-means clustering, and the corresponding H3K4me3 occupancy at the Set1A peaks were plotted for WT and mutant ESCs. (D) Pol II occupancy at the cluster 1 sites for WT and *Set1A*$^{\Delta SET}$ ESCs. (E) Genome browser track examples with corresponding RNA-seq. H3K4me3 and Set1A ChIP-seq tracks are shown. Decrease in H3K4me3 as a result of SET domain deletion of Set1A highlighted in gray boxes.

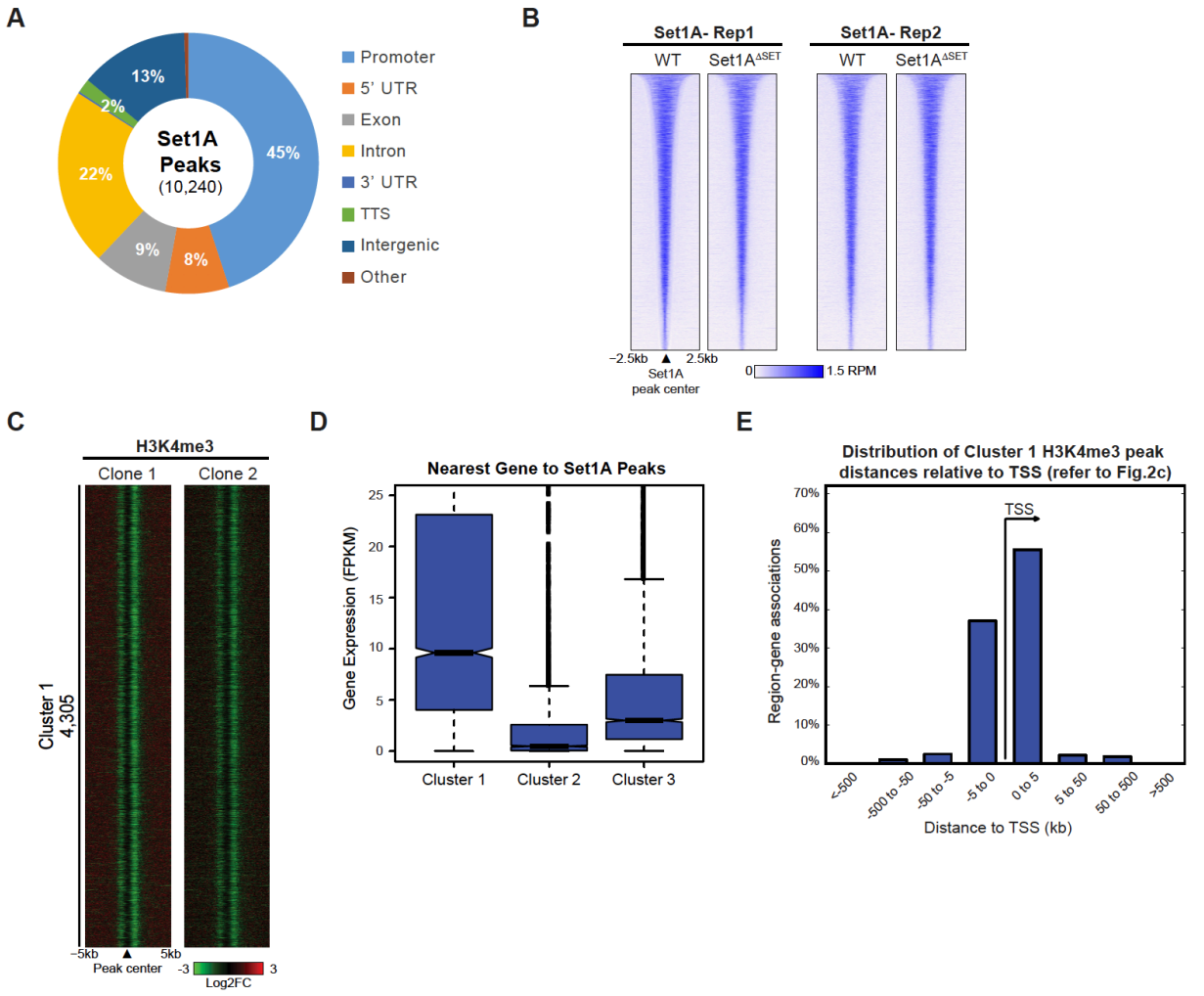

Figure 2.4. *Set1A*$^{\Delta SET}$ ESCs exhibit decreased H3K4me3 at certain sites corresponding with highest nearest gene expression despite no change in Pol II occupancy.

(A) Genome-wide distribution of Set1A binding in ESCs relative to gene structure as determined by ChIP-seq and HOMER annotation. (B) Heatmaps of Set1A binding in WT and *Set1A*$^{\Delta SET}$ cells, with occupancy levels centered at WT peaks sorted by decreasing peak width. (C) Log2 fold changes in H3K4me3 occupancy were determined in *Set1A*$^{\Delta SET}$ relative to WT cells for cluster 1 peaks identified and ordered in Figure 2.3C. (D) Box plot showing the distribution of expression level for genes nearest to Set1A peaks corresponding to clusters presented in Figure 2.3C. (E) GREAT analysis used to display distribution of H3K4me3 peak distances relative to TSSs for cluster 1 peaks identified in Figure 2.3C.

Set1A is required for early embryogenesis (*94*); however, our data show that ablation of Set1A's catalytic activity does not perturb ESC self-renewal. Therefore, we investigated whether our mutant ESCs exhibit defective differentiation by culturing WT and *Set1A*$^{\Delta SET}$ ESCs to form embryoid bodies (EBs). By day 6 of the EB generation process, spheroids generated from *Set1A*$^{\Delta SET}$ ESCs were significantly smaller than those derived from WT cells (Figure 2.5A-B). We performed RNA sequencing of day 6 EBs to compare gene expression profiles of WT and *Set1A*$^{\Delta SET}$ EBs. Differential expression analyses revealed that 649 genes are significantly downregulated in both mutant EBs compared to WT EBs (Figure 2.6A), suggesting that *Set1A*$^{\Delta SET}$ cells are undergoing defective differentiation. We subsequently identified that 447 of these 649 downregulated genes are those that fail to be normally upregulated during mutant *Set1A*$^{\Delta SET}$ EB formation (Figure 2.5C; Figure 2.6B). Gene ontology (GO) annotation shows that these 447 genes are enriched for biological processes linked to tissue specification and development, especially an enrichment for genes primarily involved in mesoderm differentiation (Figure 2.6C). Track examples of *Bmp5* and *Epor*, two genes involved in mesoderm differentiation, exhibit such diminished gene expression in both *Set1A*$^{\Delta SET}$ EBs compared to WT EBs (Figure 2.5D; Figure 2.6D). Consistent with the notion that *Set1A*$^{\Delta SET}$ ESCs display impaired differentiation, we observed elevated expressions of *Nanog*, *Oct4*, and *Sox2* in mutant EBs compared to WT (Figure 2.6E-F), indicating that these pluripotency factors are not properly extinguished in the *Set1A*$^{\Delta SET}$ mutant ESCs during EB formation, which may suggest a role for H3K4 methylation in transcriptional repression either directly or indirectly. To validate our findings that deletion of Set1A SET domain impairs ESC differentiation, we tested monolayer differentiation by culturing the cells in N2B27 media without 2i/LIF, which would promote ESC

differentiation towards the neuronal lineage (*162*). The 2i/LIF withdrawal in the first two days promoted exit from the self-renewal state in both WT and *Set1A*$^{\Delta SET}$ cells (Figure 2.6G). However, differentiation of *Set1A*$^{\Delta SET}$ cells was radically disrupted by Day 3, for the *Set1A*$^{\Delta SET}$ clones were unable to further generate neurite projections as seen in WT cells (Figure 2.6G). We also scrutinized effects on *HoxA* activation upon 24-hour retinoic acid (RA) treatment of WT vs. *Set1A*$^{\Delta SET}$ ESCs, since we had previously reported that *HoxA* gene activation is Set1A-dependent (*98*). Decreased expressions of *Hoxa4*, *Hoxa5*, and *Hoxa7* genes were detected in *Set1A*$^{\Delta SET}$ cells compared to WT cells upon RA induction (Figure 2.6H), reinforcing both our previous findings and that Set1A SET domain-dependent H3K4me3 is key factor during differentiation (*98*). Our results thus far evince that Set1A's H3K4 tri-methylase catalytic activity is required for normal ESC differentiation but is dispensable for ESC self-renewal, suggesting multiple roles for Set1A and its catalytic activity in ESC pluripotency.

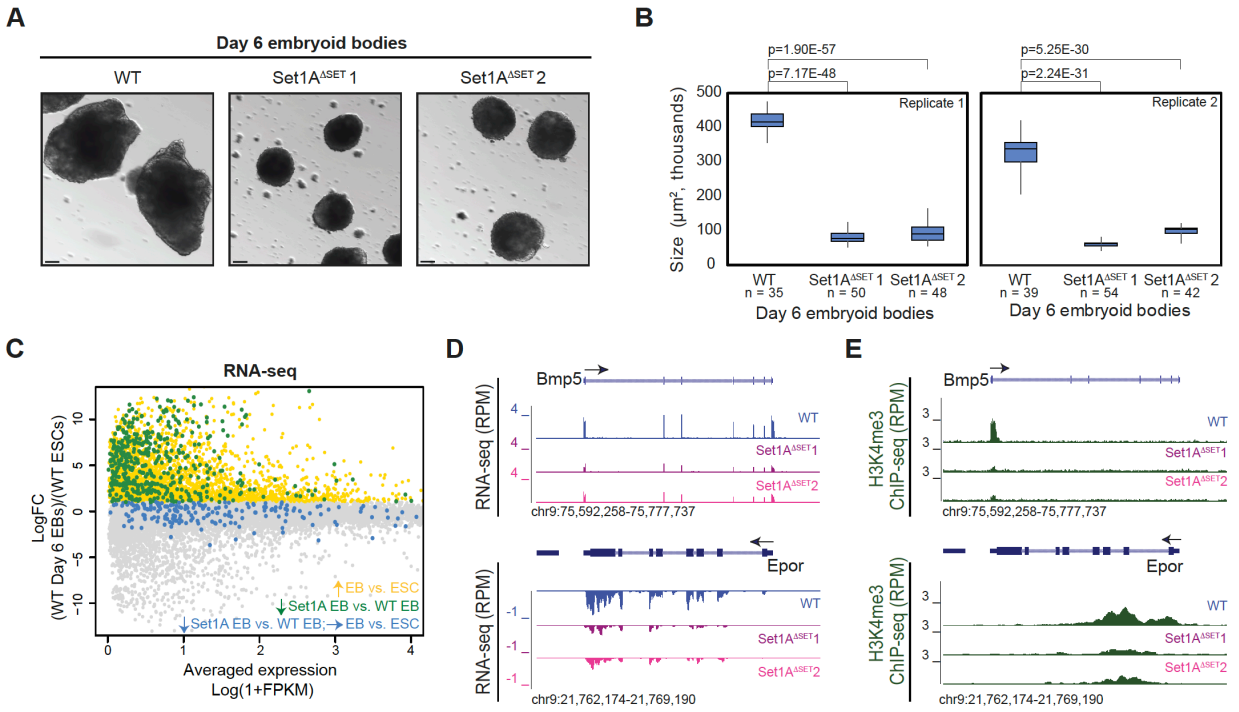

Figure 2.5. *Set1A*$^{\Delta SET}$ mutants exhibit defective embryoid body (EB) differentiation.

(A) WT and *Set1A*$^{\Delta SET}$ ESCs were induced to form EBs. Emboss contrast images were taken at day 6 of EB formation. Black scale bar = 100µm. (B) Box plot contrasting the size (µm^2) of WT day 6 EBs with that of *Set1A*$^{\Delta SET}$ EBs. Number of EBs measured per sample per replicate is indicated. P-values were calculated using the Student's t-test. (C) MA plot of gene expression changes between undifferentiated WT ESCs and WT day 6 EBs. Yellow = upregulated genes during differentiation; blue = genes downregulated in *Set1A*$^{\Delta SET}$ EBs not pertinent to differentiation; green = genes downregulated in *Set1A*$^{\Delta SET}$ EBs but activated during normal differentiation. Colors correspond to the categories illustrated in Figure 2.6B. (D) Two genome browser track examples of genes *Bmp5* and *Epor* downregulated in *Set1A*$^{\Delta SET}$ EBs compared to WT EBs. (E) Example genome tracks of decreased H3K4me3 occupancy in *Set1A*$^{\Delta SET}$ EBs compared to WT EBs for the two genes shown in panel (D).

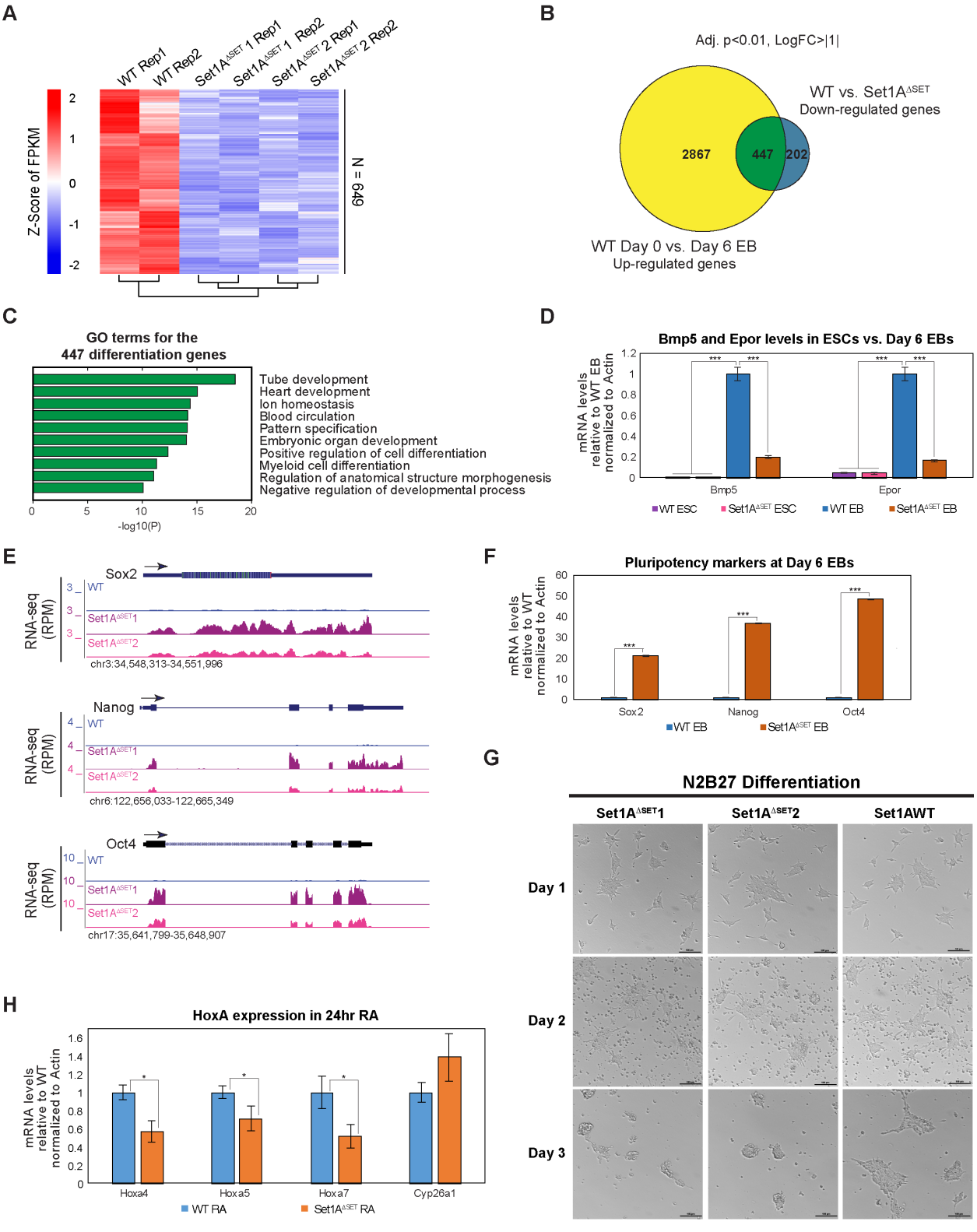
A
WT Rep1
WT Rep2
Set1A^ΔSET 1 Rep1
Set1A^ΔSET 1 Rep2
Set1A^ΔSET 2 Rep1
Set1A^ΔSET 2 Rep2
Z-Score of FPKM
2
1
0
-1
-2
N = 649
B
Adj. p<0.01, LogFC>|1|
WT vs. Set1A^ΔSET
Down-regulated genes
2867
447
202
WT Day 0 vs. Day 6 EB
Up-regulated genes
C
GO terms for the
447 differentiation genes
Tube development
Heart development
Ion homeostasis
Blood circulation
Pattern specification
Embryonic organ development
Positive regulation of cell differentiation
Myeloid cell differentiation
Regulation of anatomical structure morphogenesis
Negative regulation of developmental process
0
5
10
15
20
-log10(P)
D
Bmp5 and Epor levels in ESCs vs. Day 6 EBs
mRNA levels
relative to WT EB
normalized to Actin
1.2
1
0.8
0.6
0.4
0.2
0
***
Bmp5
Epor
WT ESC
Set1A^ΔSET ESC
WT EB
Set1A^ΔSET EB
E
Sox2
Nanog
Oct4
RNA-seq
(RPM)
WT
Set1A^ΔSET1
Set1A^ΔSET2
chr3:34,548,313-34,551,996
chr6:122,656,033-122,665,349
chr17:35,641,799-35,648,907
F
Pluripotency markers at Day 6 EBs
mRNA levels
relative to WT
normalized to Actin
60
50
40
30
20
10
0
***
Sox2
Nanog
Oct4
WT EB
Set1A^ΔSET EB
G
N2B27 Differentiation
Set1A^ΔSET1
Set1A^ΔSET2
Set1AWT
Day 1
Day 2
Day 3
H
HoxA expression in 24hr RA
mRNA levels
relative to WT
normalized to Actin
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
*
Hoxa4
Hoxa5
Hoxa7
Cyp26a1
WT RA
Set1A^ΔSET RA

Figure 2.6. Genes downregulated in *Set1A$^{\Delta SET}$* day 6 EBs compared to WT EBs linked to differentiation and development.

(A) Expression heatmap of genes downregulated in *Set1A$^{\Delta SET}$* EBs relative to WT EBs. Two replicates of day 6 EBs were generated and harvested for RNA-seq for WT and mutant EBs. (B) Venn diagram illustrating that of the 3,314 genes upregulated during wildtype differentiation (yellow and green sections), 447 genes are significantly downregulated (green section) in *Set1A$^{\Delta SET}$* EBs compared to WT EBs. Differentially expressed genes were identified using the criteria as shown. Colors also correspond to categories described in Figure 2.5C. (C) GO analysis for the 447 differentially expressed genes identified in panel (B) and in Figure 2.5C as determined by Metascape. (D) qRT-PCR analysis of expression levels of the two example genes (see Figure 2.5D) downregulated in *Set1A$^{\Delta SET}$* EBs compared to WT EBs. Both ESC and day 6 EB expressions were examined. P-value was determined by the Student's t-test, with significance level of 0.001 (denoted by ***). (E) Expressions of pluripotency factors *Nanog*, *Oct4*, and *Sox2* are noticeably increased in the mutant EBs vs. WT EBs. (F) qRT-PCR analysis of expression levels of pluripotency factors in day 6 EBs. P-value was determined by the Student's t-test, with significance level of 0.001 (denoted by ***). (G) WT and *Set1A$^{\Delta SET}$* ESCs were cultured in N2B27 media without 2i/LIF and induced towards neuronal lineage. Images were taken at day 1, day 2, and day 3 of N2B27 culturing. Black scale bar = 100μm. (H) qRT-PCR analysis of expression levels of *Hoxa4*, *Hoxa5*, and *Hoxa7* genes in cells treated for 24 hours with RA. *Cyp26a1* served as a control for RA induction. P-value was determined by the Student's t-test, with significance level of 0.05 (denoted by *).

To further determine if the enzymatic activity of Set1A regulates H3K4 methylation in *Set1A*$^{\Delta SET}$ EBs, we sought to ascertain H3K4me3 levels via ChIP-seq analyses. We noticed that reduced H3K4me3 occupancy in *Set1A*$^{\Delta SET}$ EBs is evident at *Bmp5* and *Epor* (Figure 2.5E), and therefore surveyed H3K4me3 levels at the 447 mis-regulated genes in *Set1A*$^{\Delta SET}$ EBs. Decreased H3K4me3 was observed in *Set1A*$^{\Delta SET}$ EBs compared to WT EBs (Figure 2.7A), correlating with downregulated expression of these genes in mutant EBs. To examine the extent of H3K4me3 changes attributed to Set1A activity, we first identified sites of differentially increased H3K4me3 occupancy during differentiation (Figure 2.8A). At the sites of increased H3K4me3 peaks during differentiation, we observed a compelling reduction of H3K4me3 occupancy in *Set1A*$^{\Delta SET}$ EBs compared to the occupancy in WT EBs (Figure 2.7B-C; Figure 2.8B), in line with our findings that the catalytic activity of Set1A is necessary for proper stem cell differentiation.

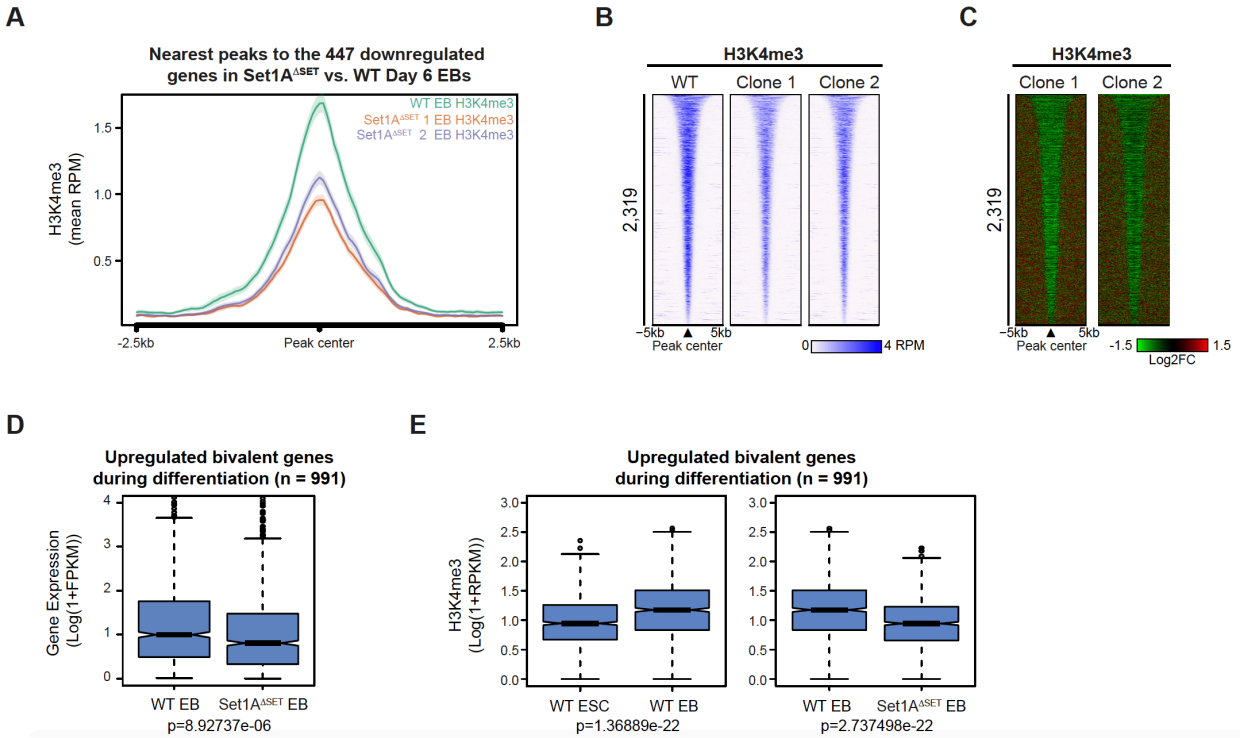

Figure 2.7. Set1A catalytic activity is required for H3K4me3 implementation during differentiation.

(A) Metaplot of H3K4me3 level for the 447 genes significantly downregulated (as described in Figure 2.5C and Figure 2.6B) in *Set1A*$^{\Delta SET}$ day 6 EBs relative to WT EBs. Peaks were centered at EB H3K4me3 peaks. (B) Heatmap of H3K4me3 occupancy in WT and *Set1A*$^{\Delta SET}$ day 6 EBs. Peaks were centered at increased H3K4me3 peaks during differentiation (refer to Figure 2.8B). Peaks are rank ordered by H3K4me3 peak width. (C) Log2 fold changes in H3K4me3 binding for peaks ordered in panel (B). (D) Box plot gene expression analysis of the 991 bivalent genes activated during differentiation in mutant EBs vs. WT EBs. List of bivalent genes was from Galonska et al. (*163*). P-value was determined by the Welch Two Sample t-test. (E) Box plot representation of levels of H3K4me3 peaks nearest to the 991 bivalent genes in WT ESCs vs. WT EBs (left) and in WT EBs vs. *Set1A*$^{\Delta SET}$ EBs (right). P-values were determined by the Welch Two Sample t-test. RPKM = reads per kilobase of transcript per million mapped reads.

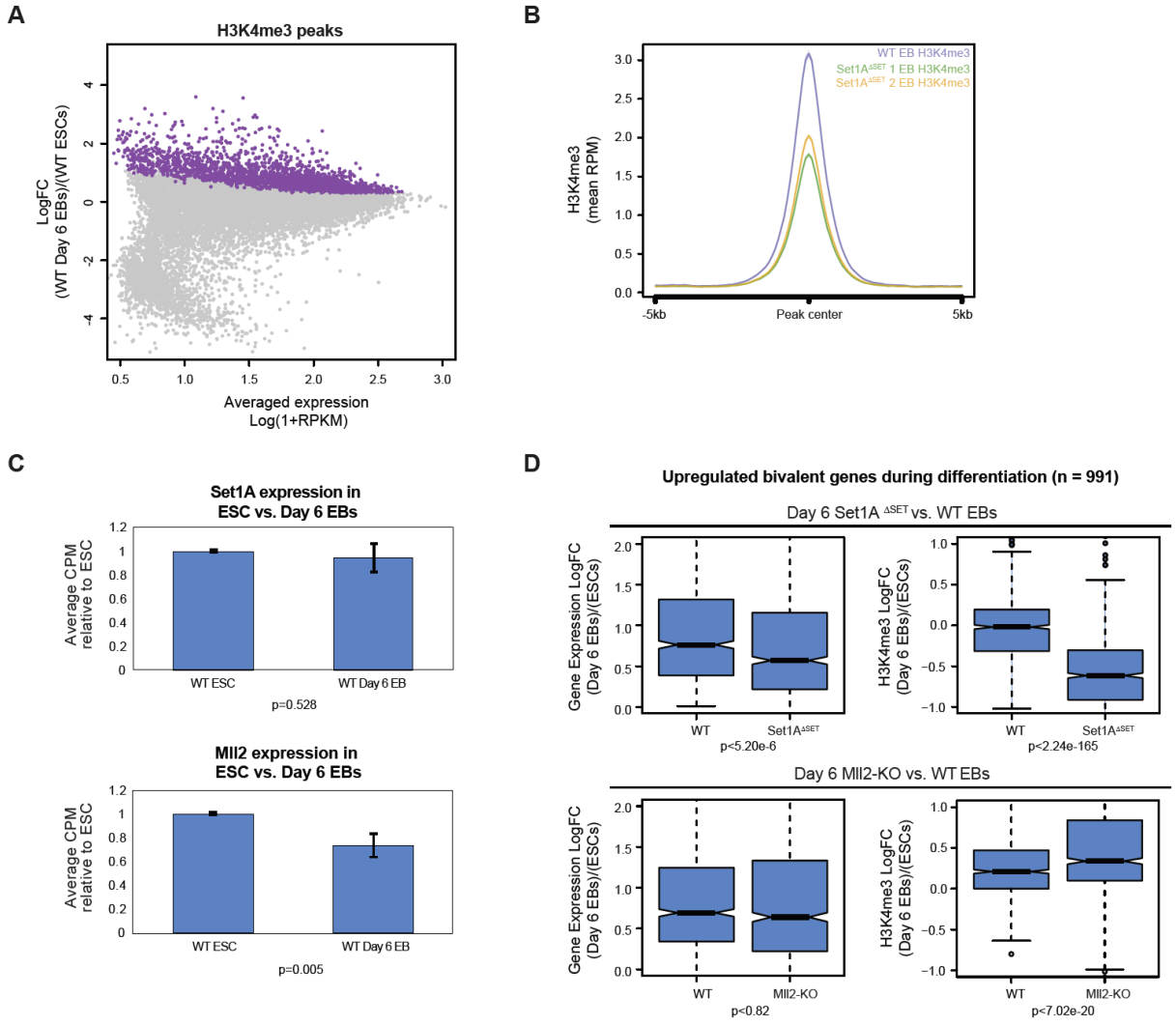

Figure 2.8. *Set1A*$^{\Delta SET}$ EBs have decreased H3K4me3 relative to WT EBs.

(A) MA plot showing differential H3K4me3 occupancy levels during differentiation. Log fold changes of H3K4me3 signals between ESCs and EBs are plotted against average expression levels in all samples. 2,319 H3K4me3 peaks (purple) increased during differentiation were defined by adj.p<0.01. (B) Metaplot of H3K4me3 levels in *Set1A*$^{\Delta SET}$ EBs vs. WT EBs aligned to peaks where H3K4me3 occupancy significantly increased during differentiation identified as purple dots in panel (A). (C) *Set1A* and *Mll2* mRNA levels in ESCs vs. day 6 EBs. CPM (counts per million) was determined using edgeR, and averaged among replicates for corresponding cell state. P-values were determined using the Student's t-test. (D) Box plot gene expression (left) and H3K4me3 (right) analyses of the 991 activated bivalent genes for *Set1A*$^{\Delta SET}$ EBs (top) and *Mll2*-KO EBs (bottom). P-values were determined by the t-test.

Another member of the COMPASS family of H3K4 methyltransferases, Mll2, deposits H3K4me3 at bivalent promoters in ESCs (*70*). However, upon differentiation of *Mll2*-depleted cells, increased gene expression and H3K4me3 were still observed at bivalent genes (*105, 161*), indicating another enzyme might be responsible for activating and catalyzing H3K4me3 at these loci. We analyzed expression of bivalent genes upregulated during differentiation and found a significant reduction in their expression with a coincident decrease in H3K4me3 occupancy in *Set1A*$^{\Delta SET}$ EBs compared to WT EBs (Figure 2.7D-E). These data suggest that Set1A catalytic activity is key in driving transcriptional activation of bivalently marked genes, which denote poised developmental regulators in pluripotent stem cells (*16, 164*), during cellular differentiation. Without Set1A methyltransferase activity, these regulators could not be normally expressed, rendering them unable to govern proper differentiation.

Taken together, our data demonstrate that while Set1A H3K4 methyltransferase activity is dispensable for ESC self-renewal and maintenance, it is essential for properly coordinating gene expression during ESC differentiation. Our findings substantially differ from a recent publication that reported the negligible impact of catalytic-deficient Mll3 and Mll4, COMPASS relatives of Set1A, on ESC viability and transcriptional program (*160*). Unlike *Mll3*/*Mll4*-KO ESCs, *Set1A* deletion in ESCs renders the cells nonviable (*89, 94, 160*). Therefore, the revelation that the enzymatic activity of Set1A/COMPASS is not essential for ESC self-renewal is remarkable. On the basis of our findings, we propose the following: in self-renewing ESCs, TSSs at lineage-specific genes harbor low levels of H3K4me3 implemented by Mll2/COMPASS (*70, 105*); however, during differentiation, Set1A/COMPASS, potentially acting in concert with other co-activating complexes, promotes H3K4me3 deposition at these developmental genes, thus designating them for transcriptional activation. Our data collectively suggest that there is a

potential functional switch between COMPASS family members, here Mll2/COMPASS and Set1A/COMPASS, in stage-specific epigenetic regulation of transcriptional outputs. When we compared mRNA levels of *Set1A* and *Mll2* in ESCs and day 6 EBs, we found a drop in *Mll2* expression while *Set1A* levels remained constant in the EB state, signifying Set1A's functional activity in EBs is correspondingly more impactful in transcriptional regulation during differentiation (Figure 2.8C). Additionally, upon generating *Mll2*-KO (*75*) day 6 EBs and assessing changes in H3K4me3 and gene expression at the same set of activated bivalent genes, we found that unlike *Mll2*-KO, we see a significant decrease in both H3K4me3 deposition as well as gene expression in $Set1A^{\Delta SET}$ EBs at the indicated bivalent genes (Figure 2.8D). The data suggest that Set1A but not Mll2 is responsible for H3K4 methylation and expression of bivalent genes during differentiation, and support our proposal of a functional switch in transcriptional regulation between Mll2 to Set1A during the self-renewing to differentiation transition. In other words, our findings suggest that there might be both loci-specific and context-specific pattern of H3K4 methylation that directs transcriptional regulation during ESC pluripotency.

The foregoing findings in ESCs and EB differentiation, combined with published data suggesting Set1A being necessary for gastrulation (*94*), justify further investigation into the role of Set1A's catalytic activity in regulating development. To define the *in vivo* effects of deleting the SET domain of Set1A, we generated mutant mice harboring the $Set1A^{\Delta SET}$ mutation via pronuclear injection of the same CRISPR sgRNAs (Figure 2.1C; Appendix 7.1). After obtaining 34 resulting F0 founder mice, we attempted to use PCR to identify the intended ~1.7kb deletion from the CRISPR sgRNAs (Figure 2.9A-B). When initial PCR efforts failed to show *in vivo* the intended deletion of the ~1.7kb region (Figure 2.9B), we resorted to next generation sequencing (NGS) to identify F0 mice with the $Set1A^{\Delta SET}$ mutation by designing uniquely barcoded primers

around the gRNA cut site in exon 17 of *Set1A* for possible frameshift indel mutations. One male F0 mouse was found to harbor an allele containing a two-nucleotide insertion at the start of the SET domain of Set1A, resulting in an early stop codon that removes the majority of the SET domain and leads to early protein truncation (Figure 2.9C-D).

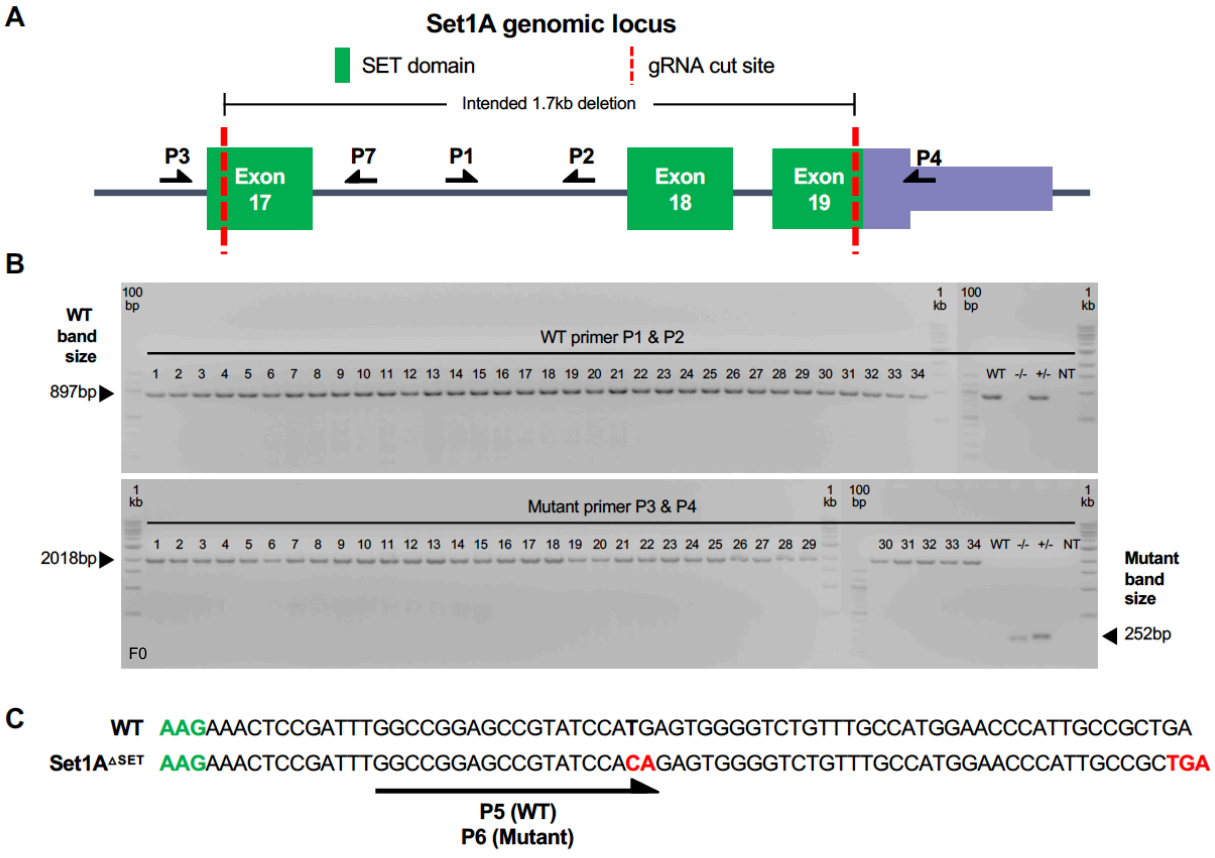
A
Set1A genomic locus
SET domain
gRNA cut site
Intended 1.7kb deletion
P3
P7
P1
P2
P4
Exon 17
Exon 18
Exon 19
B
WT band size
897bp
WT primer P1 & P2
Mutant primer P3 & P4
2018bp
Mutant band size
252bp
WT -/- +/- NT
F0
C
WT AAGAAACTCCGATTTGGCCGGAGCCGTATCCATGAGTGGGGTCTGTTTGCCATGGAACCCATTGCCGCTGA
Set1A^ΔSET AAGAAACTCCGATTTGGCCGGAGCCGTATCCACAGAGTGGGGTCTGTTTGCCATGGAACCCATTGCCGCTGA
P5 (WT)
P6 (Mutant)
D
Set1A^ΔSET ESC C1
Set1A ΔSET/+ mouse
WT V6.5
Set1A^ΔSET ESC C2
GSARSEGYYPISKKEKDKYLDVCPVSARQLEGGDTQGTNRVLSERRSEQRRLLSAIGTSA 1560
IMDSDLLKLNQLKFRKKKLRFGRMRRSPTTTSSH*---------- 1594
IMDSDLLKLNQLKFRKKKLRFGRSRIHRVGSVCHGTHCR*---------- 1599
IMDSDLLKLNQLKFRKKKLRFGRSRIHEWGLFAMEPIAADEMVIEYVGQNIRQMVADMRE 1620
IMDSDLLKLNQLKFRKKKLRFGRSRIH*---------- 1587
KRYVQEGIGSSYLFRVDHDTIIDATKCGNLARFINHCCTPNCYAKVITIESQKKIVIYSK 1680
QPIGVDEEITYDYKFPLEDNKIPCLCGTESCRGSLN* 1716

Figure 2.9. Generating mice harboring *Set1A*$^{\Delta SET}$ mutation by CRISPR/Cas9.

(A) Schematic of the *Set1A* genomic locus and two CRISPR cut sites targeting the SET domain. CRISPR sgRNAs are the same as those used to generate ESC clones (refer to Figure 2.1 and Appendix 7.1), intending for the 1.7kb deletion of genomic sequence. Genotyping primers are indicated (black arrows) and numbered. (B) 34 F0 founder mice were genotyped using the noted primer pair. WT = WT ESCs, -/- = homozygous *Set1A*$^{\Delta SET}$ ESCs, +/- = confirmed heterozygous clone, NT = no template control. 100bp and 1kb ladders are indicated. (C) Next generation sequencing of uniquely barcoded PCR products amplified around the cut site in exon 17 revealed one F0 mouse to harbor an allele containing a two-nucleotide insertion (highlighted in red) at the cut site. Early stop codon TGA was consequently introduced (highlighted in red) in the mutant allele. Start codon of SET domain is indicated in green. PCR primers to distinguish between the WT vs. mutant allele are illustrated. (D) Predicted protein sequence of the *Set1A*$^{\Delta SET}$ mutation in the identified F0 mouse compared to the two previously characterized *Set1A*$^{\Delta SET}$ ESC clones and WT V6.5 ESCs.

We subsequently confirmed that the mutant allele was germline transmissible, noting that the heterozygous mice appear to be physically healthy and fertile like their WT counterparts (Figure 2.10A). After having established the colony via heterozygous breeding, we attempted to obtain viable progeny with homozygous *Set1A*$^{\Delta SET}$ mutation (hereby known as *Set1A*$^{\Delta SET/\Delta SET}$) via heterozygous intercrosses. After evaluating 7 litters of 42 weaned mice from multiple pairs of heterozygous intercrosses, we were unable to obtain any surviving postnatal *Set1A*$^{\Delta SET/\Delta SET}$ mice (Figure 2.10B). The absence of *Set1A*$^{\Delta SET/\Delta SET}$ mutants in litters indicates that the homozygous *Set1A*$^{\Delta SET}$ mutation induces embryonic lethality (Figure 2.10 B-C).

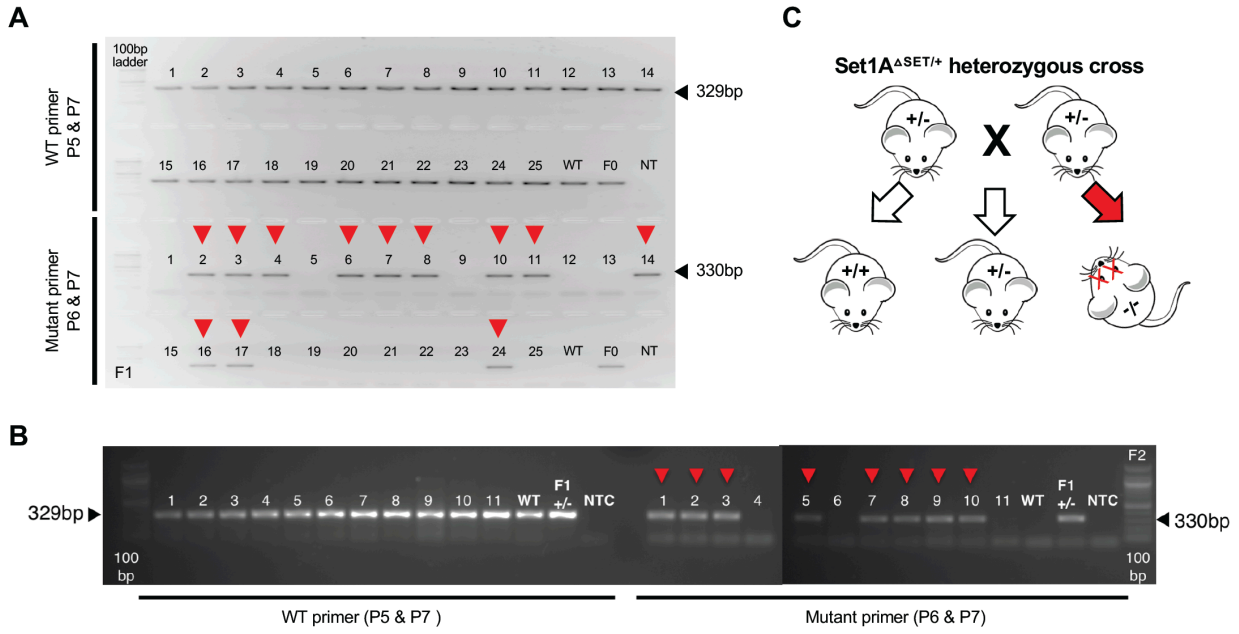

Figure 2.10. Homozygous *Set1A*$^{\Delta SET}$ mutation results in embryonic lethality.

(A) Genotyping using the noted primer pair confirmed germline transmission of mutant allele in F1 progeny. Primer numbering corresponds to the numbered primers in Figure 2.9A, C. (B) Genotyping gel shows 2 representative sets of progenies from heterozygous intercrosses, which did not result in any viable, postnatal homozygous mutants. +/- = confirmed heterozygous mouse, NTC = no template control. (C) Schematic of heterozygous *Set1A*$^{\Delta SET/+}$ mice intercross; homozygous *Set1A*$^{\Delta SET}$ mutation is lethal.

To determine the onset of lethality, developmentally-staged embryos were dissected, genotyped, and characterized for abnormalities. Since *Set1A*-KO mice die ~E7.5 (*94*), we began obtaining embryos at E8.5 to determine if our *Set1A*$^{\Delta SET/\Delta SET}$ mutants would exhibit longer survival. Indeed, we were able to detect *Set1A*$^{\Delta SET/\Delta SET}$ mutants beyond the reported lethality of E7.5 for *Set1A*-KO mutants (*94*) (Figure 2.11). We noticed that mutant embryos were increasingly smaller than WT and heterozygous littermates from E8.5 to E13.5 (Figure 2.11B). During earlier stages of gestation examined, homozygous mutants exhibited grossly delayed and

aberrant growth; by E13.5, every single homozygous mutant identified was retrieved from shrunken decidua, showing these embryos were undergoing resorption (Figure 2.11B).

A

Stage	(WT) Set1A$^{+/+}$	Set1A$^{+/\Delta SET}$	Set1A$^{\Delta SET/\Delta SET}$	Total # of embryos
E8.5	21%	54%	25%	76
E9.5	16%	49%	36%	45
E10.5	26%	63%	11%	38
E11.5	21%	47%	32%	38
E12.5	32%	48%	20%	44
E13.5	22%	61%	17%	46

B

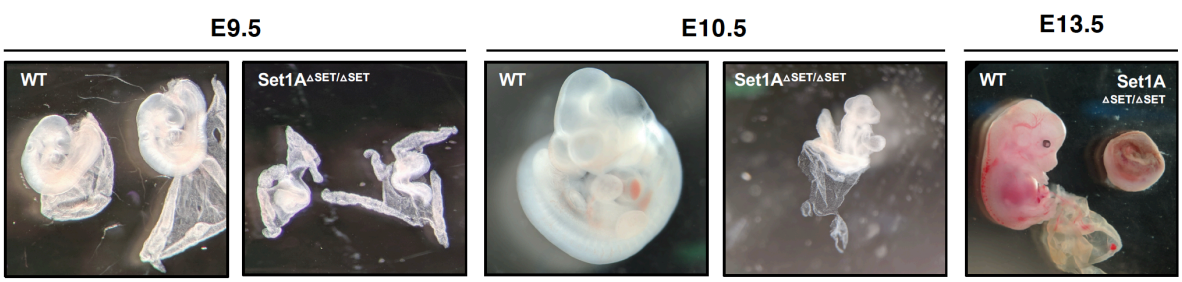

Figure 2.11. Genotypes of embryos from heterozygous *Set1A$^{\Delta SET/+}$* intercrosses.

(A) Genotypes of resulting embryos for each of the specified stage during embryogenesis. A few cases of resorption in which decidua were very small and extremely minute amounts of embryo material were recovered for genotyping were included in the total #. These cases were determined to be homozygous mutants from PCR genotyping. (B) Representative images of dissected littermates WT vs. homozygous mutants at E9.5, E10.5, and E13.5. Note the severe delay in development of the homozygous mutant compared to its WT littermate.

We delved into investigating the underlying transcriptomic changes of mutant embryos compared to their developmentally normal littermates by extracting RNA from whole E8.5 embryos from 3 separate litters for NGS. Differential expression analyses demonstrated that unlike the comparison between WT and heterozygotes, there are significant differences in gene

expression between WT and homozygotes (Figure 2.12 A-B). Overlapping the significantly differentially expressed genes from the 3 litters revealed 313 down-regulated and 352 up-regulated genes (Figure 2.12C). GO annotation shows that these down-regulated genes are especially enriched for processes involved in mesoderm differentiation, while the 352 up-regulated genes include enrichment for negative regulation of cell proliferation (Figure 2.12D). Track examples of *Ankrd1* and *Mybpc3*, which are highly expressed in cardiac muscle cells, exhibit such reduced gene expression in $Set1A^{\Delta SET/\Delta SET}$ mice compared to either WT or $Set1A^{\Delta SET/+}$ E8.5 embryos (Figure 2.12E). It is possible that Set1A may regulate mesoderm differentiation via its SET domain by depositing H3K4 methylation at the promoters of mesodermal genes, or that the SET domain serves as a scaffold for other interacting lineage-specific transcription factors involved in mesodermal signaling. Meanwhile, expressions of *Ccng1* and *Trp53inp1*, known to function in DNA damage response, are dramatically increased in the homozygous mutants vs. their WT or heterozygous counterparts (Figure 2.12E). While is it conceivable that the abnormal up-regulation of proteins involved in p53 signaling may be an indirect effect, several studies have in fact implicated Set1A in orchestrating DNA damage signaling and repair to help maintain genome stability (*100, 122, 123, 165*). Overall, our findings strongly highlight the essential role of Set1A in governing proper cell viability and proliferation, as well as mesoderm development during early embryogenesis.

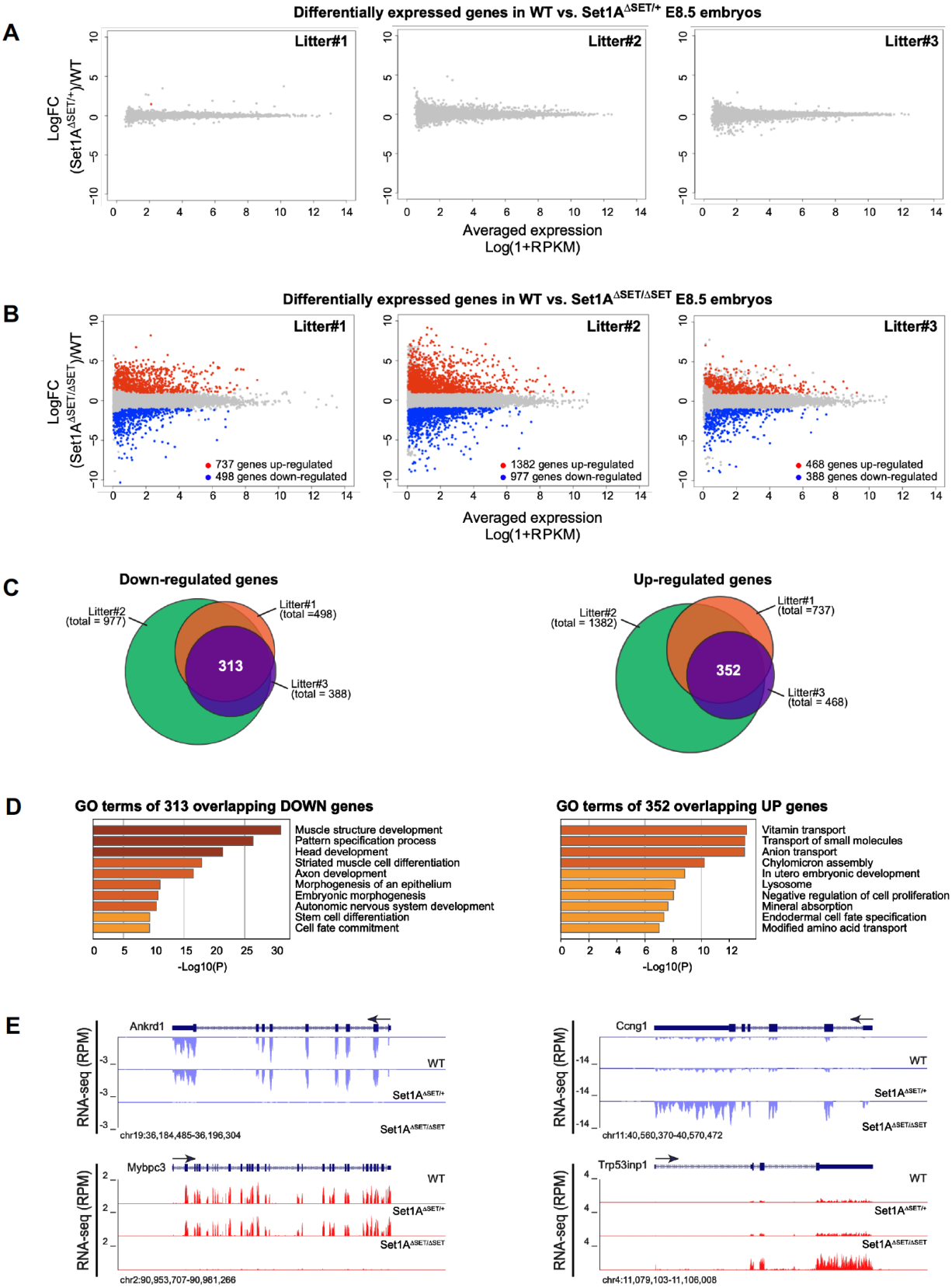
A
Differentially expressed genes in WT vs. Set1A^ΔSET/+ E8.5 embryos
Litter#1
Litter#2
Litter#3
LogFC (Set1A^ΔSET/+)/WT
Averaged expression Log(1+RPKM)
B
Differentially expressed genes in WT vs. Set1A^ΔSET/ΔSET E8.5 embryos
737 genes up-regulated
498 genes down-regulated
1382 genes up-regulated
977 genes down-regulated
468 genes up-regulated
388 genes down-regulated
C
Down-regulated genes
Litter#2 (total = 977)
Litter#1 (total =498)
313
Litter#3 (total = 388)
Up-regulated genes
Litter#2 (total = 1382)
Litter#1 (total =737)
352
Litter#3 (total = 468)
D
GO terms of 313 overlapping DOWN genes
Muscle structure development
Pattern specification process
Head development
Striated muscle cell differentiation
Axon development
Morphogenesis of an epithelium
Embryonic morphogenesis
Autonomic nervous system development
Stem cell differentiation
Cell fate commitment
-Log10(P)
GO terms of 352 overlapping UP genes
Vitamin transport
Transport of small molecules
Anion transport
Chylomicron assembly
In utero embryonic development
Lysosome
Negative regulation of cell proliferation
Mineral absorption
Endodermal cell fate specification
Modified amino acid transport
-Log10(P)
E
Ankrd1
RNA-seq (RPM)
WT
Set1A^ΔSET/+
Set1A^ΔSET/ΔSET
chr19:36,184,485-36,196,304
Ccng1
chr11:40,560,370-40,570,472
Mybpc3
chr2:90,953,707-90,961,266
Trp53inp1
chr4:11,079,103-11,106,008

Figure 2.12. RNA-seq analyses of E8.5 embryos revealed underlying mesodermal defects in homozygous mutants vs. WT littermates.

(A) MA plots comparing the transcriptome between WT and heterozygous littermates for three individual litters of E8.5 embryos. Criteria used to determine differentially expressed genes are: 1) logFC>|1|, 2) log(avgCPM)>1, and 3) Benjamini-Hochberg adjusted p-values<0.01. RPKM = Reads per kilobase of transcript per million mapped reads. (B) MA plots comparing the transcriptome between WT and homozygous littermates from the same three independent litters as in panel (A). Differentially expressed genes were identified using the criteria stated in panel (A). (C) Venn diagram illustrating the overlapping up-regulated and down-regulated genes from 3 independent litters of E8.5 embryos. (D) GO analysis for the overlapping 313 down-regulated genes and 352 up-regulated identified in panel (C) as determined by Metascape. (E) Genome browser track examples of 2 down-regulated and 2 up-regulated genes in the homozygous mutants compared to WT and heterozygous littermates.

Using our mouse model, we have demonstrated that the SET domain of Set1A plays an integral role in mesoderm differentiation, corroborating earlier RNA-seq findings from comparing *Set1A*$^{\Delta SET}$ EBs vs. WT EBs derived from ESCs in culture (Figure 2.6C). In a broader perspective, deleting both alleles encoding the SET domain of Set1A yields a divergent developmental phenotype from that of removing the full Set1A protein: defects from *Set1A*-KO embryos suggested that Set1A functions shortly after inner cell mass formation but pre-gastrulation (*94*), while our data imply that the SET domain of Set1A may function during gastrulation, possibly to help maintain the underlying morphogenetic activity of mesoderm formation (Figure 2.12). From these findings, we could also deduce that the non-catalytic function of Set1A *in vivo* is to ensure proper asymmetric cell division and unrestricted developmental potential of the cells in the inner cell mass prior to the formation of the three germ layers. To date, the other COMPASS methyltransferase known to exhibit distinct domain-specific function *in vivo* is Mll1. *Mll1*-null results in embryonic lethality at E10.5 due to the requirement for Mll1 in sustaining *Hox* gene expression and definitive hematopoiesis during

embryogenesis (*103, 166, 167*). However, targeting the Mll1 SET domain unexpectedly resulted in viable mice with skeletal deformities, unveiling the *in vivo* dependence and independence of the Mll1 SET domain (*104*). Taken together, while these findings collectively contribute to the field's increased understanding of COMPASS function beyond its catalytic SET domain, additional research to elucidate the mechanisms underlying the diverse roles of COMPASS in governing proper development is essential.

3 Coordinated regulation of cellular identity-associated H3K4me3 breadth by the COMPASS family

Majority of the work from this chapter, including figures, has been reproduced, with or without modifications, from my recently accepted manuscript: Christie C. Sze, Patrick A. Ozark, Kaixiang Cao, Michal Ugarenko, Siddhartha Das, Lu Wang, Stacy A. Marshall, Emily J. Rendleman, Caila A. Ryan, Didi Zha, Delphine Douillet, Fei Xavier Chen, and Ali Shilatifard. Sci Adv. 2020. In press.

3.1 Introduction

Epigenetic post-translational modifications on histones are highly dynamic; extensive changes of these histone states can impact recruitment of effector proteins that fundamentally shape gene expression programs underlying processes that govern development and disease (*168-173*). Specifically, methylation at histone lysine residues has received considerable attention due to the important involvement with transcriptional regulation. One such mark, histone H3 lysine 4 trimethylation (H3K4me3), is an evolutionarily conserved modification consistently found at transcription start sites (TSS) and serves as a hallmark of active gene promoters (*174, 175*). Studies have conferred that H3K4me3 interacts with chromatin remodelers (*176, 177*) and promotes recruitment of basic transcription factors to facilitate transcriptional activation (*178, 179*). Other functional implications ascribed to H3K4me3 include DNA damage repair response (*180, 181*), marking of bivalent genes co-marked with the repressive H3K27me3 modification (*16*), and cell identity specification (*182, 183*). Thus, there is a consensus that H3K4me3 is critically linked to transcriptional output and cellular response.

Deposition of methylation on H3K4 is catalyzed by lysine methyltransferases (KMTs), namely the family of COMPASS (COMplex of Proteins ASsociated with Set1) complexes, which is widely conserved from yeast to human (*52, 58, 184*). While Set1/COMPASS is solely

responsible for implementing all three methylation patterns in yeast (*52, 54, 184*), our laboratory and others have described the division of labor in methylating H3K4 among the COMPASS members in higher organisms (*53, 54, 64, 67-70, 72*). In mammals, there are six Set1-related enzymes that reside in COMPASS-like complexes: Set1A, Set1B, Mll1, Mll2, Mll3, and Mll4 (*56, 58, 73*). Evolutionary expansion of the COMPASS family in higher metazoans denotes functional diversification of H3K4 methylation, showcasing the underlying complexity of epigenetic regulation (*59*). Mammalian Set1A/Set1B have been accredited as mainly responsible for bulk H3K4me3 genome-wide (*56, 67, 185, 186*); Mll1/Mll2 catalyze H3K4me3 in a locus-specific manner (e.g. targeting Hox genes) (*69, 70, 75, 187, 188*); and Mll3/Mll4 are key H3K4 mono-methyltransferases at enhancers (*72-74*). Recent studies have begun to reassess the biological significance of the catalytic activity as the primary function of COMPASS. For instance, the enzymatic SET domain of Set1A was shown to be nonessential for ESC self-renewal (*99*), although deletion of the full-length protein impairs viability (*89, 94*). Likewise, inactivating mutations of the SET domain of Trr, the Drosophila homolog of Mll3/Mll4, does not result in clear developmental defects, and Mll3/Mll4 catalytically-deficient ESCs have less transcriptional aberrations than seen in cells with total protein loss (*115, 160*). Our laboratory also determined Mll4/COMPASS could regulate enhancers without its enzymatic activity (*26*). These studies suggest that the COMPASS family function independent of H3K4 methylation is context-dependent. However, deletion of any individual COMPASS member in mice results in embryonic or prenatal lethality with distinct phenotypes (*74, 94, 189-191*), signifying these proteins have partially nonredundant functions. Although research has provided insight into how each COMPASS member operates in unique contexts, mechanisms underlying their roles in cellular and developmental regulation remain elusive.

We previously reported that ablating the SET domain of Set1A (*Set1A*$^{\Delta SET}$) does not disrupt bulk H3K4me3 in ESCs (*99*), indicating the likelihood of other COMPASS members having functionally redundant roles to sustain global H3K4me3 in *Set1A*$^{\Delta SET}$ ESCs. Multiple lines of evidence thus far have suggested that both Set1A and its paralog Set1B contribute to genome-wide H3K4me3 deposition (*67, 185, 186*). However, recent studies suggest Set1A and Set1B have functionally distinct roles, as Set1B overexpression could not mitigate proliferation defects caused by loss of Set1A protein in ESCs (*94*), and Set1B is localized mostly in the cytoplasm (*101*). Given the existing perplexity in their function, we sought to elucidate the extent to which Set1A and Set1B, as well as their familial relatives Mll1 and Mll2, may have distinct versus overlapping responsibilities in H3K4me3 regulation in ESCs in the current study. Through generating an array of ESC lines containing compounding mutations of the COMPASS family, we discovered that Set1A, Set1B, and Mll2 engage in an epigenetic collaborative circuit to modulate the H3K4me3 signature and breadth in ESCs. Our findings provide evidence for both functional specialization and redundancy of the mammalian COMPASS family members to direct transcriptional regulation and cell identity in a context-specific manner, shedding novel insights into mechanisms underlying disease pathogenesis associated with mutations of such critical epigenetic modifiers of chromatin.

3.2 Results & discussion

Our previous study demonstrated that deleting the C-terminal catalytic SET domain of Set1A does not disrupt global H3K4 methylation in ESCs, which suggested that other enzymes implement bulk H3K4 methylation in the absence of Set1A activity (*99*). Since Set1B is structurally homologous to Set1A (Figure 3.1A), we investigated whether Set1B could compensate for Set1A$^{\Delta SET}$ function to sustain global H3K4 methylation in ESCs. Therefore, we generated *Set1B* knockout (*Set1B*KO) ESCs by deleting the first four exons of the *Set1B* genomic locus via CRISPR/Cas9 gene editing. PCR genotyping and RNA sequencing (RNA-seq) confirmed the deletion of the intended genomic region and *Set1B* transcript (Figure 3.1B-C), and Western blotting substantiated the complete loss of the Set1B protein in ESCs (Figure 3.2A). Ablating Set1B did not alter bulk H3K4 methylation as observed by Western blotting (Figure 3.1D; Figure 3.2B). ChIP-seq (chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing) analyses exhibited similar H3K4me3 levels in wild-type and *Set1B*KO ESCs (Figure 3.2C). In addition, the expressions of key pluripotency factors *Sox2*, *Nanog*, and *Oct4* remain invariable when comparing wild-type and *Set1B*KO cells (Figure 3.1E). These findings thus far agree with prior studies demonstrating that Set1B is not essential for ESC self-renewal, and that Set1B removal does not affect bulk H3K4 methylation (*94*).

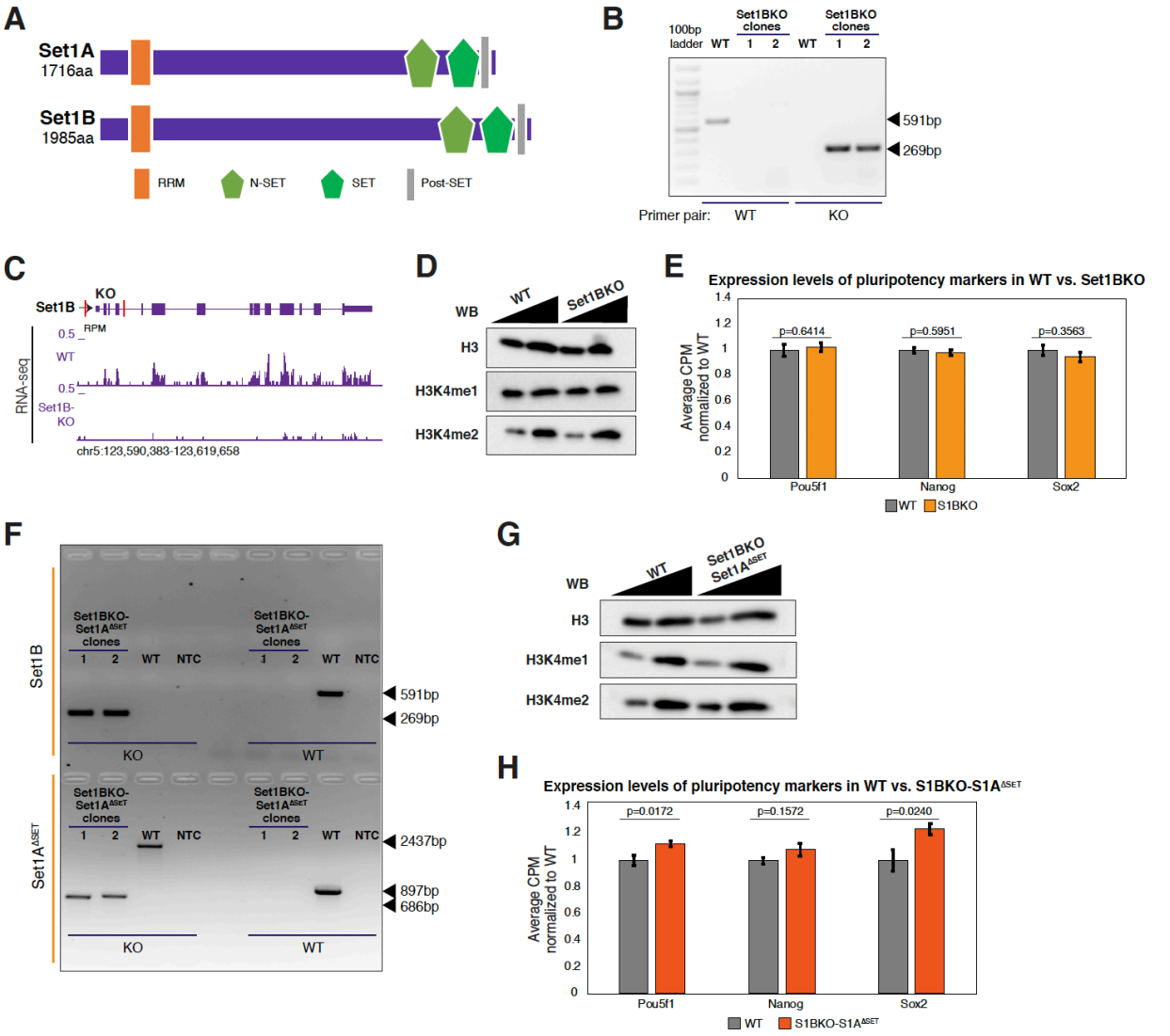
A
Set1A
1716aa
Set1B
1985aa
RRM
N-SET
SET
Post-SET
B
100bp ladder
Set1BKO clones
WT
1
2
WT
1
2
591bp
269bp
Primer pair:
WT
KO
C
Set1B
KO
RPM
0.5
WT
0.5
Set1B-KO
RNA-seq
chr5:123,590,383-123,619,658
D
WB
WT
Set1BKO
H3
H3K4me1
H3K4me2
E
Expression levels of pluripotency markers in WT vs. Set1BKO
Average CPM normalized to WT
p=0.6414
p=0.5951
p=0.3563
Pou5f1
Nanog
Sox2
WT
S1BKO
F
Set1B
Set1A^ΔSET
Set1BKO-Set1A^ΔSET clones
1
2
WT
NTC
591bp
269bp
2437bp
897bp
686bp
KO
WT
G
WB
WT
Set1BKO Set1A^ΔSET
H3
H3K4me1
H3K4me2
H
Expression levels of pluripotency markers in WT vs. S1BKO-S1A^ΔSET
Average CPM normalized to WT
p=0.0172
p=0.1572
p=0.0240
Pou5f1
Nanog
Sox2
WT
S1BKO-S1A^ΔSET

Figure 3.1. Generation and characterization of *Set1B*KO and *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs.

(A) Schematic of known functional domain organization of mouse Set1A and Set1B proteins. (B) PCR genotyping results to identify WT vs. *Set1B*KO ESCs. Arrowheads indicate sizes of PCR products in base pairs (bp). (C) RNA-seq tracks confirming deletion of the intended *Set1B* transcript. Vertical red bars indicate targeted genomic region in *Set1B*KO ESCs. RPM: reads per million. (D) Western blot (WB) of H3K4me1 and H3K4me2 in *Set1B*KO vs. WT ESCs. H3 served as loading control. Samples were loaded at a 1:2 ratio. (E) Average expression level of key pluripotency genes *Pou5f1* (*Oct4*), *Nanog*, and *Sox2* in WT vs. *Set1B*KO ESCs. CPM reads were averaged and normalized relative to WT for each gene. P-values were determined using the Student's t-test. (F) PCR genotyping of WT vs. *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs. Targeted genes *Set1B* and *Set1A* are indicated, with the resulting PCR sizes shown. NTC is non-template water control. (G) Western blot of H3K4me1 and H3K4me2 in *Set1B*KO-*Set1A*$^{\Delta SET}$ vs. WT cells, with total H3 as the loading control. (H) Average expressions of pluripotency factors *Pou5f1*, *Nanog*, and *Sox2* in *Set1B*KO-*Set1A*$^{\Delta SET}$ relative to WT ESCs. The Student's t-test was used to calculate p-values.

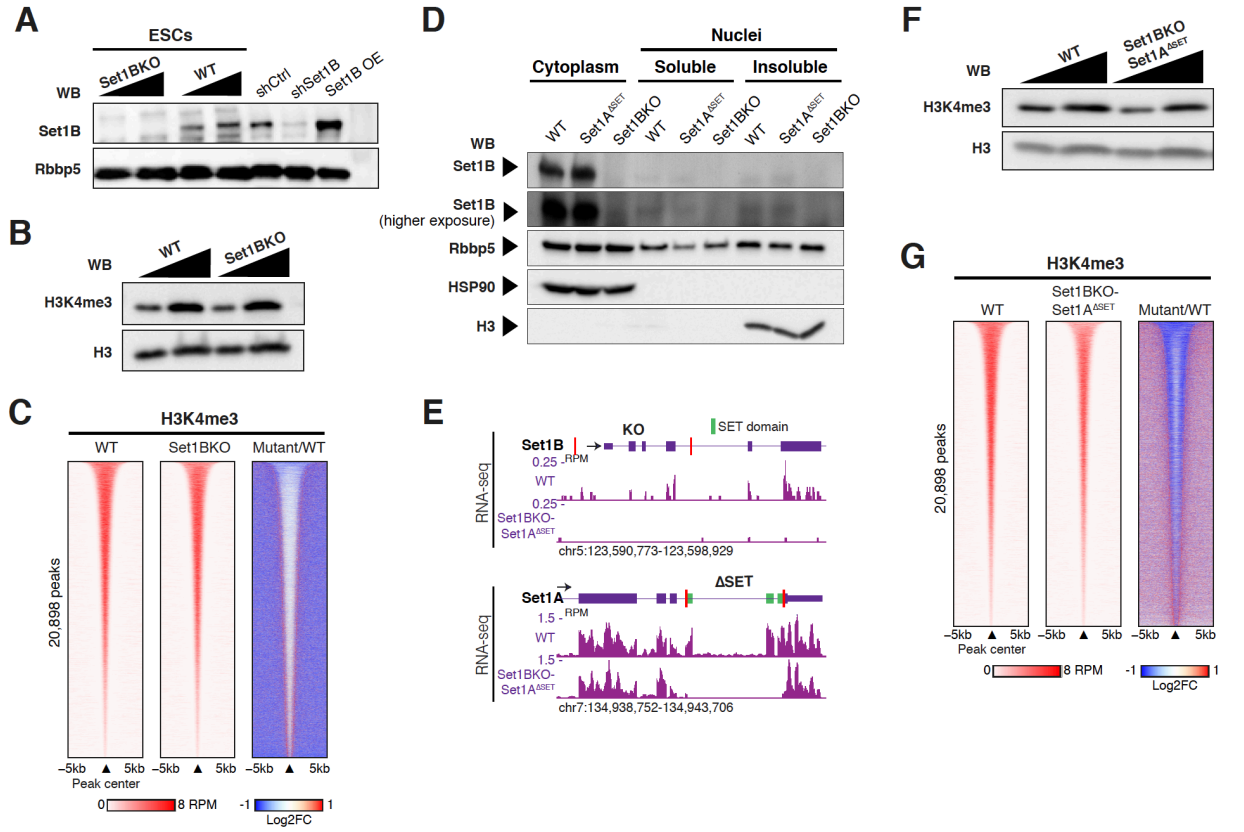

Figure 3.2. Set1B compensates Set1A in depositing H3K4me3 in ESCs.

(A) Western blot of Set1B in wild-type (WT) and *Set1B*KO ESCs, with Rbbp5 as the loading control. Samples were loaded at a 1:2 ratio. Whole cell extract lysates of MCF7 cells with *Set1B* knockdown and overexpression (47) were used to identify the correct Set1B band. (B) Western blot comparing H3K4me3 levels in *Set1B*KO with WT cells. Total H3 was used as the loading control. Samples were loaded at a 1:2 ratio. (C) Left: Heatmaps of H3K4me3 ChIP-seq levels in WT and *Set1B*KO ESCs. Occupancy levels were aligned to WT peaks sorted by decreasing peak width. Right: Log2 fold changes (Log2FC) in H3K4me3 binding comparing *Set1B*KO and WT ESCs for the same ordered peaks. Two biological replicates of each genotype were analyzed. (D) Cytoplasmic and nuclear fractions were extracted from WT, *Set1A*$^{\Delta SET}$, and *Set1B*KO ESCs, and endogenous levels of Set1B and Rbbp5 were detected by Western blot. HSP90 and total H3 served as cytoplasmic and insoluble nuclear fractionation marker respectively. (E) RNA-seq results validate CRISPR/Cas9-mediated deletion of exons 1-4 of Set1B (top) and the sequence coding for the SET domain of Set1A (bottom) in ESCs. Vertical red bars indicate targeted genomic regions in *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs. RPM: reads per million. (F) Western blot of H3K4me3 in WT and *Set1B*KO-*Set1A*$^{\Delta SET}$ cells, with H3 serving as loading control. Samples were loaded at a 1:2 ratio. (G) Heatmaps showing H3K4me3 levels in WT and *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs. Occupancy was aligned to WT peaks, sorted in descending width, and displayed on the left. Log2FC of H3K4me3 between double-mutant and WT cells is shown on the right. Two biological replicates of each genotype were analyzed.

Our laboratory has recently established that Set1B, unlike the rest of its COMPASS family relatives, resides predominantly in the cytoplasm (*101*); however, we find that a small fraction of Set1B still binds to chromatin in ESCs, prompting us to determine the potential role of Set1B in H3K4 methylation (Figure 3.2D). Since independent abrogation of Set1B or the catalytic SET domain of Set1A did not change overall H3K4 methylation levels (*99*), we used CRISPR/Cas9 to establish *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs. Successful generation of homozygous double-mutant ESCs was verified by PCR genotyping and RNA-seq (Figure 3.1F; Figure 3.2E). Expressions of pluripotency markers *Oct4*, *Nanog*, and *Sox2* are comparable between wild-type and *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs at the significance level of $p<0.01$ (Figure 3.1H). We assessed H3K4 methylation in the double-mutant cells compared with that in wild-type ESCs and found that while bulk H3K4me1/me2 levels did not differ, H3K4me3 was somewhat decreased in *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs by Western blotting (Figure 3.1G; Figure 3.2F). The reduction in H3K4me3 levels in the double-mutant compared to wild-type ESCs was also consistently detected by ChIP-seq (Figure 3.2G). The diminished H3K4me3 in the combined deletions of the SET domain of Set1A and knockout of Set1B compared to either mutation alone, suggests functional redundancy between the two Set1 paralogs in regulating H3K4 methylation in ESCs. It is possible that both Set1A and Set1B co-localize to the same genomic loci to compensate for each other in stress conditions, since both methyltransferases reside within COMPASS complexes with overlapping subunit composition. Currently, it is challenging to determine genome-wide occupancy of Set1B given the lack of ChIP-grade antibody against Set1B, combined with the fact that majority of Set1B exists in the cytoplasm (Figure 3.2D) (*101*).

Several studies have implicated a role for H3K4me3 in transcriptional activation by RNA polymerase II (RNAP II) (*178, 179*). Given the noticeable decrease in H3K4me3 in *Set1B*KO-

Set1A$^{\Delta SET}$ compared to wild-type cells, we used ChIP-seq to evaluate RNAP II occupancy genome-wide. Reduced RNAP II is evident at sites with H3K4me3 loss, and accompanying reduction in gene expression can be seen in representative track examples (Figure 3.3A). Such decrease in H3K4me3 and RNAP II is not due to the loss of Set1A SET domain or Set1B alone (*99*), further supporting their functionally redundant roles in H3K4me3 implementation and transcription regulation at these regions (Figure 3.4A-B). Genome-wide analyses in the *Set1B*KO-*Set1A*$^{\Delta SET}$ double-mutant cells revealed that H3K4me3 is lost across regions proximal to transcription start sites (TSSs), with a corresponding decrease in RNAP II occupancy compared to wild-type ESCs (Figure 3.4C). Partitioning of these H3K4me3 peaks using K-means clustering revealed that H3K4me3 loss in the double-mutant occurs in the second and third clusters, where Set1A binding is the strongest (Figure 3.3B-C). Both clusters also exhibit somewhat diminished RNAP II levels in *Set1B*KO-*Set1A*$^{\Delta SET}$ compared to wild-type cells (Figure 3.3C). The accompanying RNA-seq data demonstrate moderate differences in gene expression pattern, with a stronger decrease in expression pertinent to cluster 3, which has a relatively more severe H3K4me3 loss (Figure 3.3C). Cluster 3 also contains genes that are more highly expressed in wild-type ESCs (Figure 3.4D), and is especially enriched for housekeeping factors such as ribosomal-related proteins as shown by gene ontology (GO) annotation (Figure 3.4E), as well as key ESC pluripotency markers (e.g., Oct4, Nanog, and Sox2). Collectively, these data signify that H3K4me3 at promoters of more highly expressed genes are more perturbed by the combinatorial deletions of Set1B and the SET domain of Set1A in ESCs. These findings are consistent with previous studies reporting that H3K4me3 at genes with higher expression are more sensitive to loss of Cxxc1, a key subunit of the Set1/COMPASS complexes (*192, 193*).

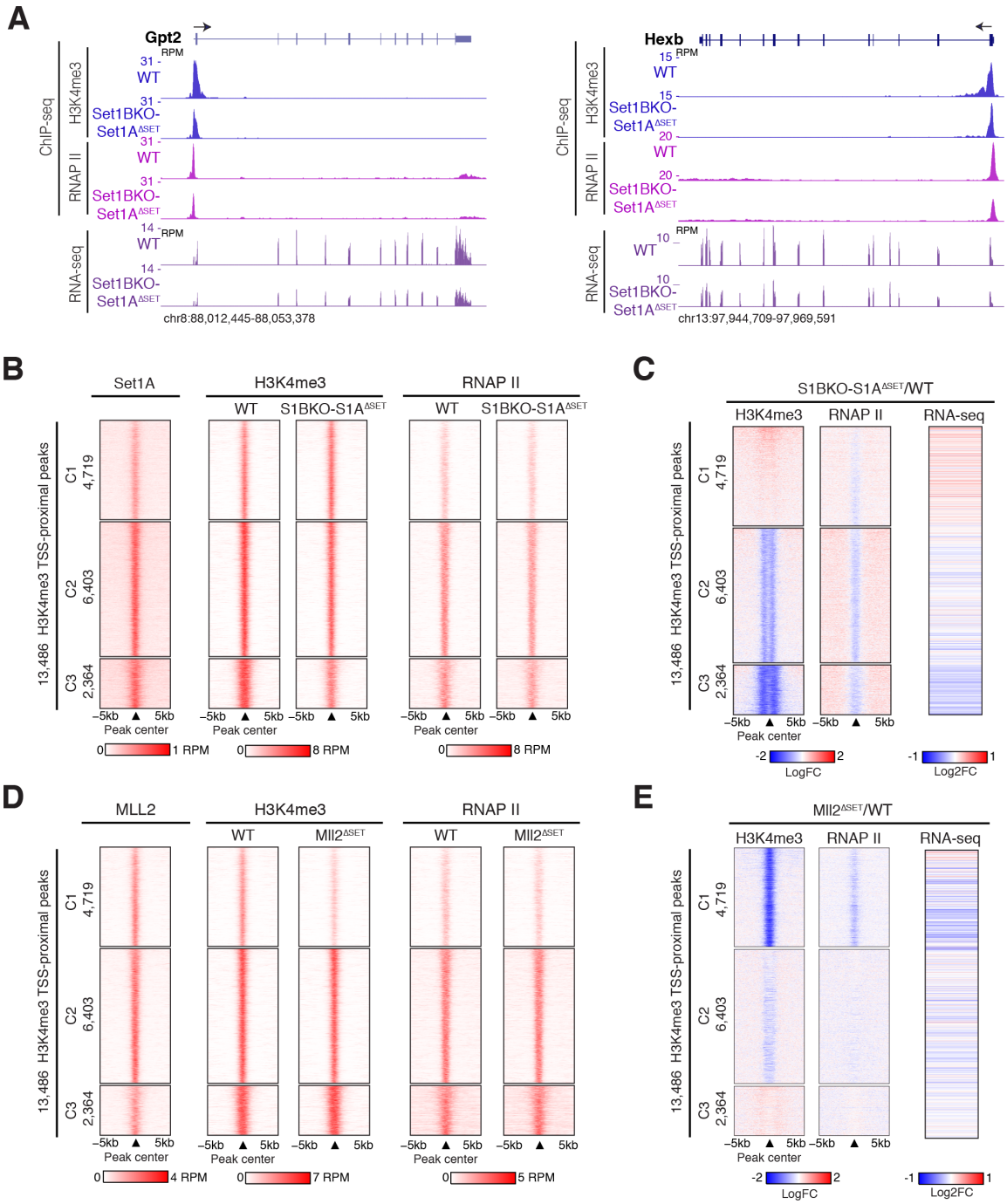
A
Gpt2
Hexb
ChIP-seq
RNA-seq
H3K4me3
RNAP II
WT
Set1BKO-Set1A^ΔSET
RPM
chr8:88,012,445-88,053,378
chr13:97,944,709-97,969,591
B
Set1A
H3K4me3
RNAP II
S1BKO-S1A^ΔSET
13,486 H3K4me3 TSS-proximal peaks
C1 4,719
C2 6,403
C3 2,364
Peak center
1 RPM
8 RPM
C
S1BKO-S1A^ΔSET/WT
RNA-seq
LogFC
Log2FC
D
MLL2
MII2^ΔSET
4 RPM
7 RPM
5 RPM
E
MII2^ΔSET/WT

Figure 3.3. H3K4me3 is implemented by Set1/COMPASS at more transcriptionally active promoters, while Mll2 catalyzes H3K4me3 at lowly expressed genes.

(A) Genome browser track examples showing ChIP-seq tracks of H3K4me3 and RNAP II, with corresponding RNA-seq tracks for *Gpt2* (left) and *Hexb* (right). (B) 13,486 TSS-proximal H3K4me3 peaks were identified and partitioned into three groups via K-means clustering, and corresponding Set1A (left), H3K4me3 (middle), and RNAP II (right) levels were plotted for WT and double-mutant ESCs. (C) Log2FC changes in H3K4me3 (left) and RNAP II (middle) occupancy, as well as in RNA-seq (right), were determined in *Set1B*KO-*Set1A*$^{\Delta SET}$ relative to WT cells for the three clusters of peaks identified and ordered in (B). (D) Mll2 (left), H3K4me3 (middle), and RNAP II (right) occupancy levels were plotted for WT and *Mll2*$^{\Delta SET}$ ESCs at the clustered 13,486 H3K4me3 peaks identified in (B). (E) Log2FC changes in H3K4me3 (left) and RNAP II (middle) occupancy, as well as in RNA-seq (right), were assessed in *Mll2*$^{\Delta SET}$ compared to WT ESCs for the three clusters of peaks determined and ordered in (B).

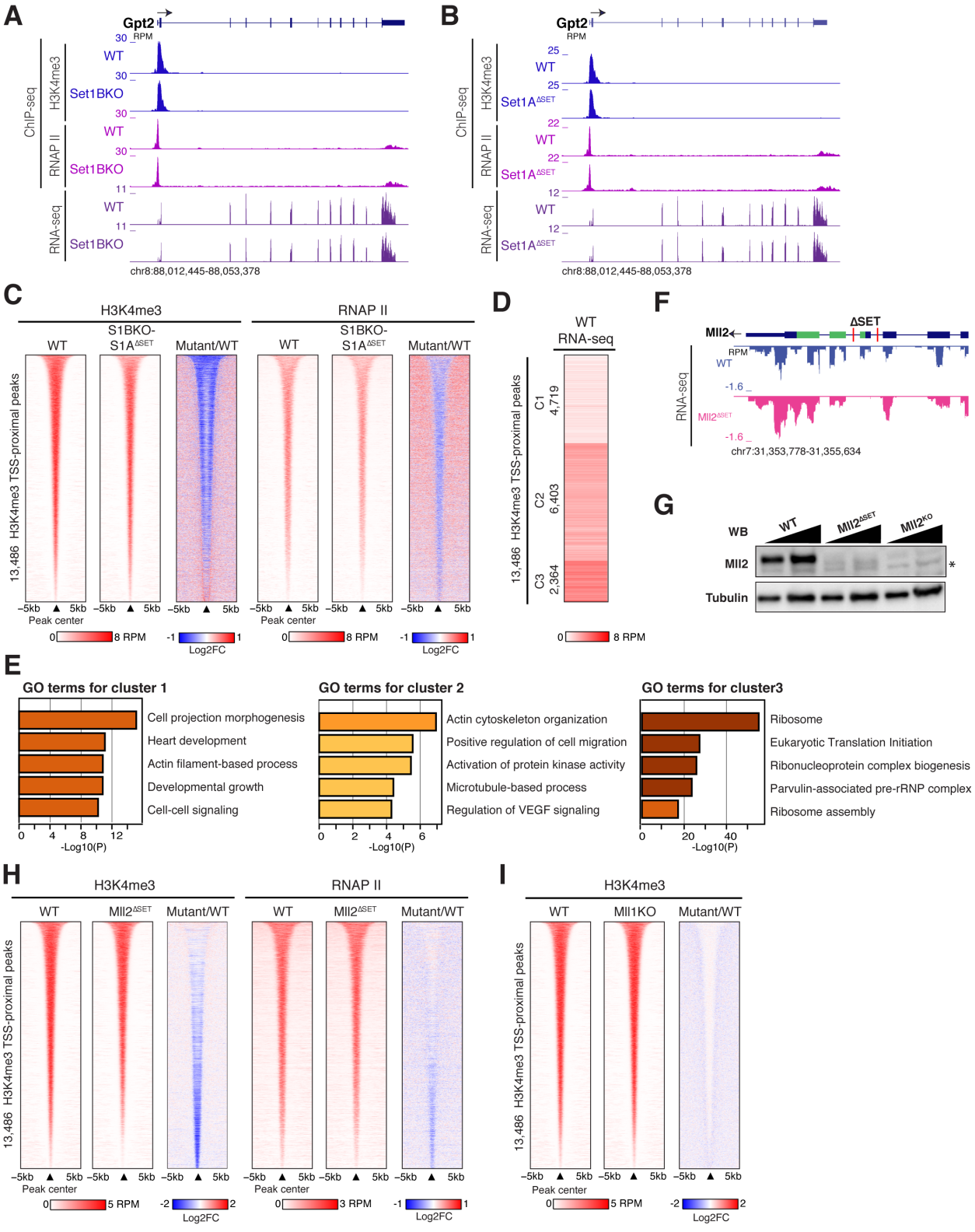
A
Gpt2
ChIP-seq
H3K4me3
RNAP II
RNA-seq
WT
Set1BKO
chr8:88,012,445-88,053,378
B
Set1A^ΔSET
C
H3K4me3
RNAP II
S1BKO-S1A^ΔSET
Mutant/WT
13,486 H3K4me3 TSS-proximal peaks
Peak center
Log2FC
D
WT RNA-seq
C1 4,719
C2 6,403
C3 2,364
F
Mll2
ΔSET
Mll2^ΔSET
chr7:31,353,778-31,355,634
G
WB
Mll2
Tubulin
E
GO terms for cluster 1
Cell projection morphogenesis
Heart development
Actin filament-based process
Developmental growth
Cell-cell signaling
-Log10(P)
GO terms for cluster 2
Actin cytoskeleton organization
Positive regulation of cell migration
Activation of protein kinase activity
Microtubule-based process
Regulation of VEGF signaling
GO terms for cluster3
Ribosome
Eukaryotic Translation Initiation
Ribonucleoprotein complex biogenesis
Parvulin-associated pre-rRNP complex
Ribosome assembly
H
Mll2^ΔSET
Mll1KO
I

Figure 3.4. Set1 and Mll2/COMPASS catalyze H3K4me3 at different genomic regions.

(A, B) Genome browser track examples showing ChIP-seq tracks of H3K4me3 and RNAP II, with corresponding RNA-seq tracks, for *Set1B*KO (A) or *Set1A*$^{\Delta SET}$ (B) ESCs compared to WT. H3K4me3 ChIP-seq data for *Set1A*$^{\Delta SET}$ ESCs were previously published (*99*). (C) Left: Heatmaps displaying the TSS-proximal H3K4me3 peaks in WT and *Set1B*KO-*Set1A*$^{\Delta SET}$ cells. Occupancy (left) was aligned to WT peaks, sorted in decreasing peak width. H3K4me3 Log2FC (right) comparing *Set1B*KO-*Set1A*$^{\Delta SET}$ to WT cells is determined accordingly. Right: Corresponding occupancy and log2FC of RNAP II levels in *Set1B*KO-*Set1A*$^{\Delta SET}$ vs. WT ESCs at the identified TSS-proximal H3K4me3 peaks were plotted. (D) Expression level heatmap for genes nearest to the H3K4me3 TSS-proximal peaks corresponding to the clusters presented in Figure 3.3, B to E. (E) GO analysis for the genes in each of the three clusters from panel (D) using the Metascape software. (F) RNA-seq tracks confirming deletion of the intended SET domain of Mll2. Vertical red bars indicate targeted genomic region in *Mll2*$^{\Delta SET}$ ESCs. (G) Western blot of Mll2 in WT, *Mll2*$^{\Delta SET}$, and *Mll2*KO ESCs, with tubulin as the loading control. Samples were loaded at a 1:2 ratio. Asterisk denotes a non-specific band. (H) Occupancy and log2FC heatmaps of H3K4me3 and RNAP II for WT and *Mll2*$^{\Delta SET}$ ESCs. Peaks are sorted as in panel (C). At least two biological replicates of each genotype were analyzed. (I) Occupancy and log2FC heatmaps of the TSS-proximal H3K4me3 peaks in WT and *Mll1*KO ESCs, which were previously generated in our laboratory (*98*). Peaks are sorted as in panel (C).

We previously found that Mll2, another member of the COMPASS family of H3K4 methyltransferases, also catalyzes H3K4me3 in ESCs (*73*); therefore, we deleted the SET domain of Mll2 using CRISPR/Cas9 (Figure 3.4F). Through western blotting, we noticed that disrupting the SET domain of Mll2 adversely affects the protein's stability (Figure 3.4G), indicating that Mll2 protein level in *Mll2*$^{\Delta SET}$ ESCs is quite comparable to Mll2KO cells, which were previously generated in our laboratory (*75*). Histone H3K4me3 ChIP-seq analysis in *Mll2*$^{\Delta SET}$ ESCs revealed loss of H3K4me3 at sites distinct from those seen in *Set1B*KO-*Set1A*$^{\Delta SET}$ when compared to wild-type cells (Figure 3.4C, H). Particularly, the greatest H3K4me3 reduction in *Mll2*$^{\Delta SET}$ cells is confined mainly to cluster 1 peaks, where there is also a corresponding decrease in RNAP II level and overall expression of genes nearest to these sites (Figure 3.3D-E). Cluster 1 contains the TSS of genes that are typically less transcriptionally active and are linked to function in proper development (Figure 3.4D-E), concordant with previous findings that Mll2/COMPASS implements H3K4me3 primarily at specific loci such as bivalent genes (*70*). Analysis of H3K4me3 ChIP-seq in ESCs without Mll1 (*98*), also known to deposit H3K4me3 in mammals, showed comparable levels of global H3K4me3 between *Mll1*KO and wild-type cells (Figure 3.4I), signifying that Mll1/COMPASS is not a crucial regulator of bulk H3K4me3 in ESCs.

The K-means clustering of H3K4me3 in Figure 3.3 effectively divided peaks by their width, which also correlated with gene expression, such that wider H3K4me3 coincided with higher nearest gene expression, and narrower H3K4me3 concurred with lower nearest gene expression (Figure 3.3B-E; Figure 3.4D), suggesting an underlying biological significance to H3K4me3 breadth. In fact, multiple studies have recently explored such functional implications, reporting a positive relationship between H3K4me3 breadth and gene expression (*183, 194*) and

attributing the role of breadth to determining specific cell identity during development and disease (*182, 195, 196*). We noticed the predilection of Set1/COMPASS mutant cells to present H3K4me3 loss in clusters 2 and 3, which have greater H3K4me3 breadth, while mutant Mll2/COMPASS primarily affected cluster 1 with the narrowest breadth (Figure 3.4B-E). To investigate the function of H3K4me3 breadth further, we classified broad and narrow wild-type peaks, annotated to the nearest TSS, and sorted these regions from wide to narrow. We retrieved 500 sites with either the broadest or narrowest peaks and examined H3K4me3 levels in wild-type, *Set1B*KO-*Set1A*$^{\Delta SET}$, and *Mll2*$^{\Delta SET}$ ESCs. In comparison to wild-type, *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs show diminished H3K4me3 at genes primarily with broader peaks, while *Mll2*$^{\Delta SET}$ cells exhibit abated H3K4me3 levels at genes with mainly narrower peaks (Figure 3.5A). These unequivocal differences of altered ChIP density by peak width in the two mutant cell lines is further illustrated quantitatively in composite profiles (Figure 3.5B) and representative track examples (Figure 3.5C). GO term analyses indicate that genes with broader H3K4me3 peaks are enriched for biological processes that are relatively more important for proper maintenance of stem cell identity (Figure 3.6A), supporting previous studies associating H3K4me3 breadth with cell identity specification (*182, 183*). In addition, narrow peaks are enriched for processes that are more pertinent to neuronal development (Figure 3.6A), consistent with published studies showing Mll2 implementing H3K4me3 at developmental-related genes in ESCs (*70, 105, 197*).

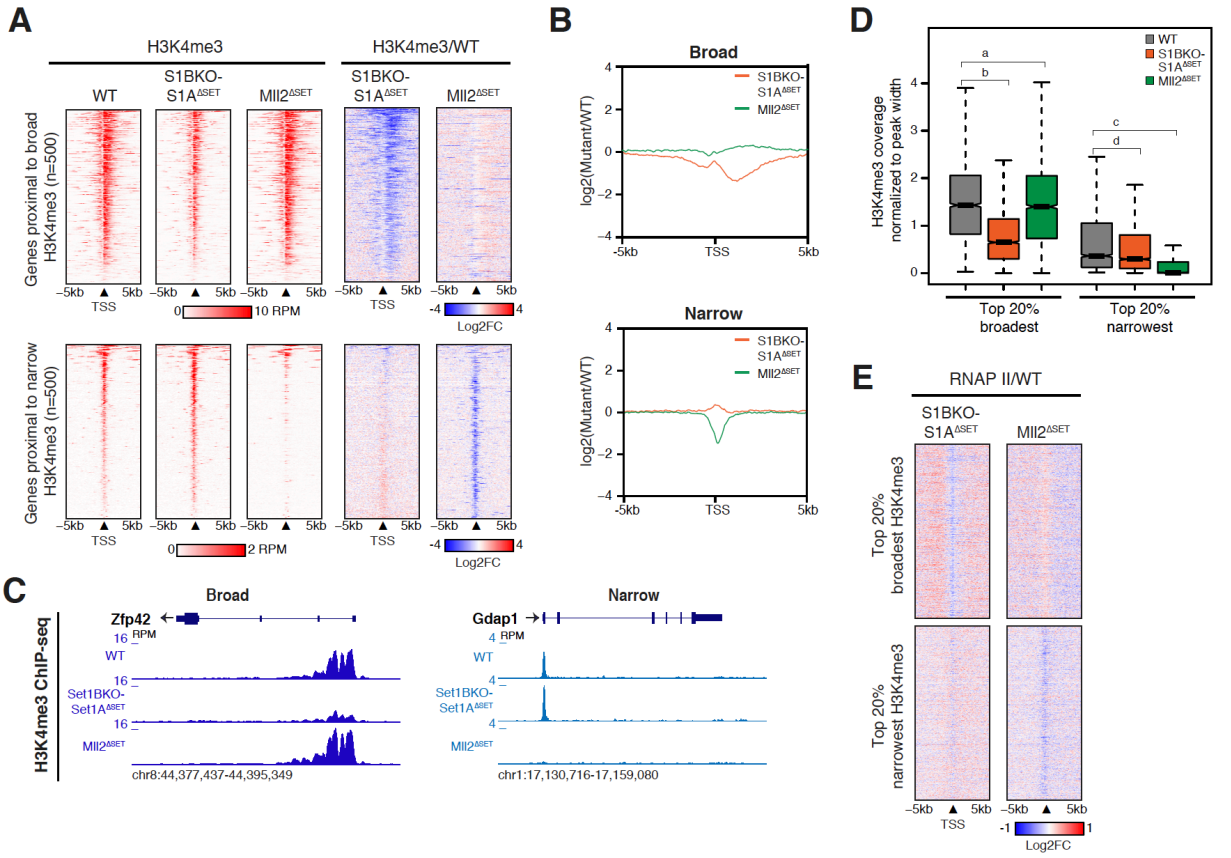

Figure 3.5. H3K4me3 peak breadth is coordinately controlled by Set1 and Mll2/COMPASS.

(A) Left: Heatmaps of H3K4me3 occupancy in WT, *Set1B*KO-*Set1A*$^{\Delta SET}$, and *Mll2*$^{\Delta SET}$ ESCs at 500 genes with the broadest (top) and 500 genes with the narrowest (bottom) H3K4me3 peaks in WT. TSSs are ordered by descending peak width of their nearest H3K4me3 peak. Right: Log2FC in H3K4me3 levels comparing either *Set1B*KO-*Set1A*$^{\Delta SET}$ or *Mll2*$^{\Delta SET}$ relative to WT cells for the same ordered set of most broad (top) and narrow (bottom) peaks. (B) Profiles of log2FC of average H3K4me3 signal density (RPM) at the loci with either 500 most broad (top) or narrow (bottom) H3K4me3 peaks in *Set1B*KO-*Set1A*$^{\Delta SET}$ or *Mll2*$^{\Delta SET}$ compared to WT ESCs. (C) Example genome tracks of broad (right) and narrow (left) H3K4me3 peaks in WT, *Set1B*KO-*Set1A*$^{\Delta SET}$, and *Mll2*$^{\Delta SET}$ ESCs. (D) Box plot contrasting normalized H3K4me3 ChIP-seq coverage for the top 20% broadest H3K4me3 peaks (N=2,626) and top 20% narrowest H3K4me3 peaks (N=2,627) in WT, *Set1BKO-Set1A*$^{\Delta SET}$, and *Mll2*$^{\Delta SET}$ ESCs. Coverage was normalized to peak width. P-values (indicated by lower case letters) were determined using the Wilcoxon rank sum test with continuity correction (paired). Calculated p-values are as follows: a, p=0.01901; b, p<6.173E-282; c, p=6.173E-282; d, p=1.7914E-74. (E) Log2FC heatmaps of RNAP II levels at H3K4me3-proximal TSSs for each indicated mutant ESC line relative to WT cells for the top 20% broadest (top) or narrowest (bottom) H3K4me3 peaks initially defined in (D).

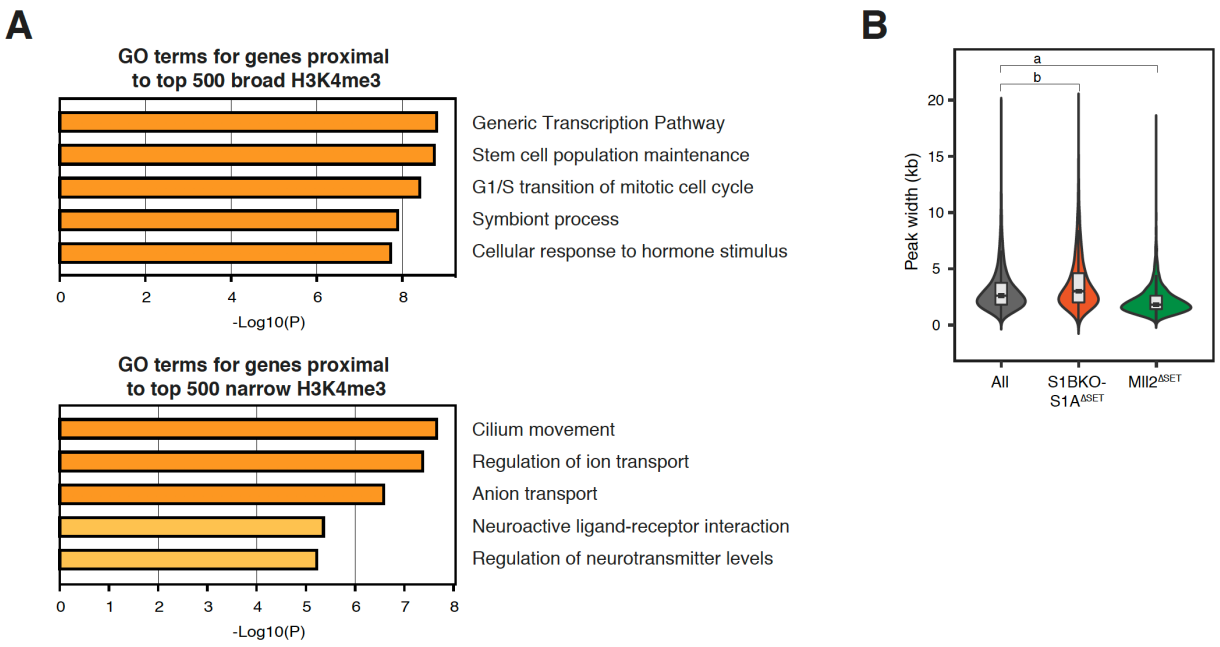

Figure 3.6. H3K4me3 peak breadth is determined by Set1 and Mll2/COMPASS.

(A) GO terms for the genes nearest to the 500 most broad (top) and 500 most narrow (bottom) H3K4me3 peaks using Metascape. (B) Differential H3K4me3 peaks at TSSs were determined between *Set1B*KO-*Set1A*$^{\Delta SET}$ or *Mll2*$^{\Delta SET}$ vs. WT ESCs. Peak breadth of differential H3K4me3 peaks in *Set1B*KO-*Set1A*$^{\Delta SET}$ (N=4,467 peaks) or *Mll2*$^{\Delta SET}$ (N=4,821 peaks) mutants were plotted against the breadth of all H3K4me3 peaks in WT ESCs (N=13,124 peaks). P-values (indicated by lower case letters) were calculated using the Wilcoxon rank sum test with continuity correction. Calculated p-values are as follows: a, p=9.392E-266; b, p=3.551E-48.

To further investigate the consequence of H3K4me3 peak width, the top and bottom breadth quintiles were evaluated. For the top 20% H3K4me3 peaks, the *Set1B*KO-*Set1A*$^{\Delta SET}$ mutant featured a significantly drastic decrease in H3K4me3 coverage compared to wild-type, while the *Mll2*$^{\Delta SET}$ mutant exhibited comparable H3K4me3 occupancy relative to wild-type (Figure 3.5D). In contrast, *Mll2*$^{\Delta SET}$ ESCs showed a more significant reduction in H3K4me3 for the 20% narrowest H3K4me3 peaks compared to wild-type, while the *Set1B*KO-*Set1A*$^{\Delta SET}$ mutant is much less altered for H3K4me3 at these narrow peak regions (Figure 3.5D). Furthermore, when assessing peak width of differential H3K4me3 TSS-proximal peaks for each mutant, Set1/COMPASS mainly affected wider peaks, while *Mll2*$^{\Delta SET}$ mutant primarily affected narrower peaks (Figure 3.6B). By examining the levels of RNAP II at H3K4me3-proximal TSSs in *Set1B*KO-*Set1A*$^{\Delta SET}$ and the *Mll2*$^{\Delta SET}$ mutants compared to wild-type, we observe decreased RNAP II levels at broader peaks for the *Set1B*KO-*Set1A*$^{\Delta SET}$ mutant and reduced RNAP II levels at narrower peaks for the *Mll2*$^{\Delta SET}$ mutant (Figure 3.5E). Based on these findings, we speculate the following: 1) transcription of cell identity genes, which are marked by broad H3K4me3 (*182*), is regulated by Set1/COMPASS, and thus illuminates the importance of Set1/COMPASS in maintaining stem cell viability (*94, 99*); and 2) the observed phenomenon of Mll2 affecting narrow H3K4me3 peaks, which mark developmental genes, supports the critical role of Mll2 in development (*75, 198, 199*). There are several plausible explanations for the differential recruitment and activity of these COMPASS methyltransferases at distinct groups of loci. Studies have shown that Set1, the yeast homolog to mammalian Set1A and Set1B, is recruited to chromatin through its association with elongating RNAP II (*83-85*). One study has recently proposed that repeated passaging of elongation complexes containing Set1/COMPASS

contribute to the widening or broadening of H3K4me3 levels, which correlate with increased transcription frequency (*86*). While this model has yet to be demonstrated for Mll2/COMPASS, it is possible that Mll2 is recruited to chromatin by other factors that include Menin, which also resides in the Mll/COMPASS complexes, and Ledgf (*69, 90-93*). In addition, different DNA-binding specificities of the CXXC motif found in the Cxxc1 protein, a key subunit in the Set1/COMPASS complex, and in Mll2, have been reported to contribute to selective targeting of various COMPASS complexes to their respective genomic loci (*76-78*). In sum, these findings demonstrate that the difference in H3K4me3 patterns established by Set1/COMPASS and Mll2/COMPASS are indicative of functional significance in stem cells.

Interestingly, we noted that Set1A and Mll2 are both highly enriched at H3K4me3 TSS-proximal peaks depicted in the K-means clustered heatmaps, indicating co-localization of Set1A and Mll2 on chromatin (Figure 3.3B, D). In addition, at annotated TSS sites, there is >98% overlap in Mll2 and Set1A binding regions (Figure 3.7A). This co-localization could be explained by the fact that both Cxxc1, a key subunit in the Set1A/COMPASS complex as discussed earlier, and Mll2 harbor a CXXC motif that recognizes and binds to unmethylated CpG-containing DNA (*75*). It is therefore possible that Mll2 could compensate for Set1/COMPASS loss in H3K4me3 deposition. Consequently, we depleted *Mll2* using a short hairpin RNA (shRNA) in *Set1B*KO-*Set1A*$^{\Delta SET}$ lines and compared to wild-type ESCs (Figure 3.7B). Knockdown of *Mll2* resulted in a dramatic decrease of H3K4me3 in *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs (Figure 3.8A-B). To investigate this observation in further detail, we therefore generated a triple-mutant cell line, where we deleted the SET domain of Mll2 in *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs by CRISPR/Cas9 (Figure 3.7C-D). Simultaneously, we removed *Mll1* in *Set1B*KO-*Set1A*$^{\Delta SET}$

ESCs using previously reported gRNAs for targeting *Mll1* (*98*), in the event that Mll1 may manifest a compensatory role in implementing H3K4me3 under conditions of Set1/COMPASS mutation (Figure 3.7C-D). Remarkably, we were able to successfully retrieve homozygous triple-mutant ESCs harboring the intended *Set1B*KO-*Set1A*$^{\Delta SET}$ -*Mll2*$^{\Delta SET}$ and *Set1B*KO-*Set1A*$^{\Delta SET}$ -*Mll1*KO mutations, as validated by PCR genotyping and RNA-seq (Figure 3.7C-D). However, we noticed that the *Set1B*KO-*Set1A*$^{\Delta SET}$ -*Mll2*$^{\Delta SET}$ cells proliferate more slowly than wild-type, *Set1B*KO-*Set1A*$^{\Delta SET}$, and *Mll2*$^{\Delta SET}$ cells (data not shown). To ascertain if the appended deletion of either SET domain of Mll2 or of Mll1 affected H3K4 methylation, we performed Western blotting and ChIP-seq analyses to evaluate H3K4me3 levels. Consistent with the effects seen in knocking down *Mll2* in the double-mutant ESCs (Figure 3.8A-B), bulk H3K4me3, including the level at TSS-proximal regions, is substantially lowered in the *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ cells (Figure 3.7E-G). However, no additive perturbation effect on H3K4me3 levels was observed in the *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll1*KO triple-mutant relative to the double-mutant (Figure 3.7E-G), affirming the minimal contribution to H3K4me3 deposition by Mll1 in ESCs.

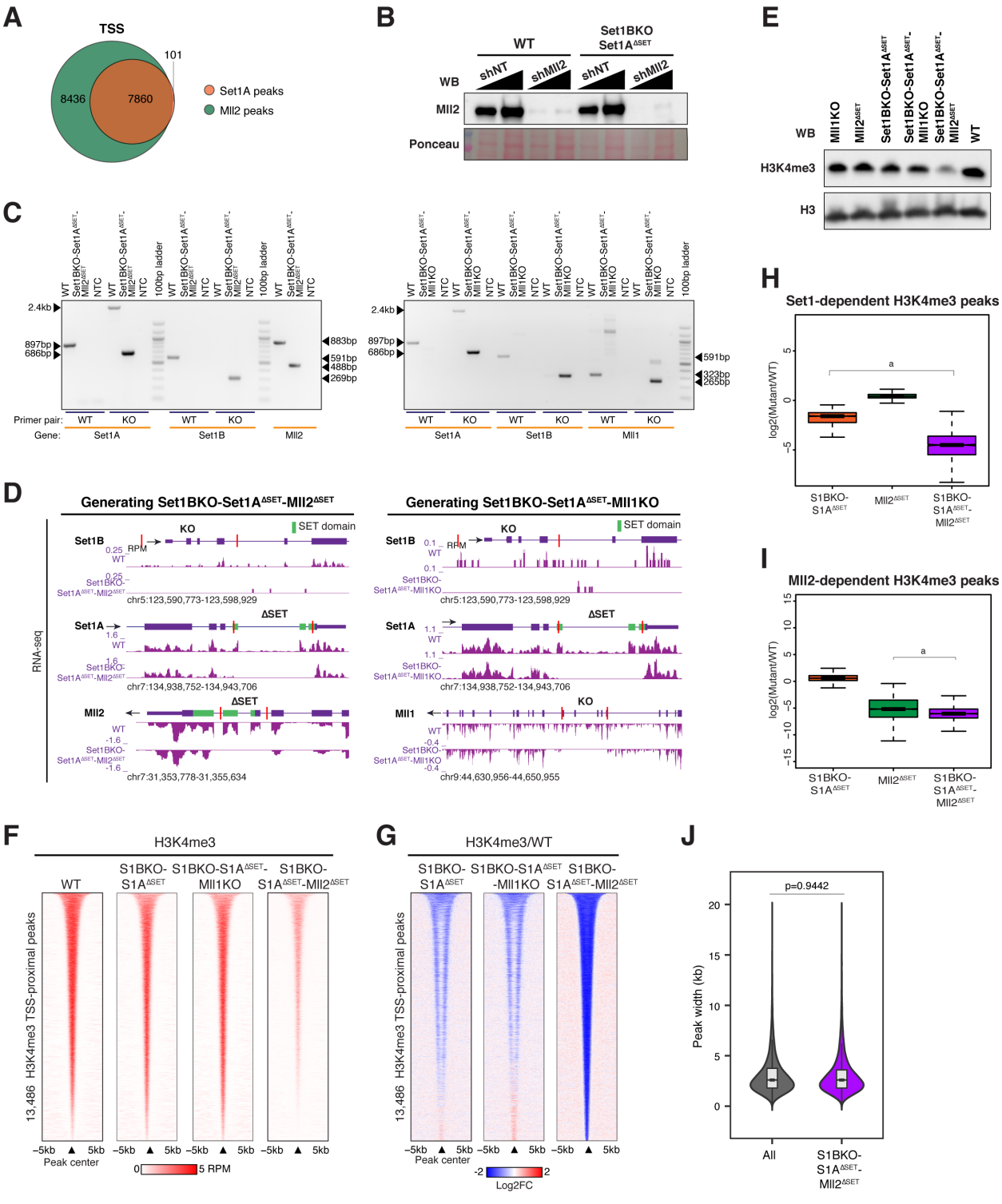
A
TSS
101
8436
7860
Set1A peaks
Mll2 peaks
B
WT
Set1BKO
Set1A^ΔSET
WB
shNT
shMll2
shNT
shMll2
Mll2
Ponceau
E
WB
Mll1KO
Mll2^ΔSET
Set1BKO-Set1A^ΔSET-
Set1BKO-Set1A^ΔSET-
Mll1KO
Set1BKO-Set1A^ΔSET-
Mll2^ΔSET
WT
H3K4me3
H3
C
2.4kb
897bp
686bp
883bp
591bp
488bp
269bp
100bp ladder
Primer pair:
WT
KO
Gene:
Set1A
Set1B
Mll2
591bp
323bp
265bp
Mll1
H
Set1-dependent H3K4me3 peaks
log2(Mutant/WT)
a
S1BKO-S1A^ΔSET
Mll2^ΔSET
S1BKO-S1A^ΔSET-Mll2^ΔSET
D
Generating Set1BKO-Set1A^ΔSET-Mll2^ΔSET
Generating Set1BKO-Set1A^ΔSET-Mll1KO
RNA-seq
Set1B
KO
SET domain
RPM
WT
Set1BKO-Set1A^ΔSET-Mll2^ΔSET
Set1BKO-Set1A^ΔSET-Mll1KO
chr5:123,590,773-123,598,929
Set1A
ΔSET
chr7:134,938,752-134,943,706
Mll2
chr7:31,353,778-31,355,634
Mll1
chr9:44,630,956-44,650,955
I
Mll2-dependent H3K4me3 peaks
F
H3K4me3
WT
S1BKO-S1A^ΔSET
S1BKO-S1A^ΔSET-Mll1KO
S1BKO-S1A^ΔSET-Mll2^ΔSET
13,486 H3K4me3 TSS-proximal peaks
−5kb
5kb
Peak center
0
5 RPM
G
H3K4me3/WT
-2
2
Log2FC
J
p=0.9442
Peak width (kb)
All

Figure 3.7. Generation and characterization of the two triple-mutant ESCs.

(A)Venn diagram of the overlap of Set1A and Mll2 peaks at TSS-proximal peaks. (B) Western blot of Mll2 in WT and *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs upon control (shNT) and *Mll2* (sh*Mll2*) knockdowns, with ponceau showing total protein loaded per sample. Samples were loaded at a 1:2 ratio. (C) Left: PCR genotyping results of WT vs. *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ ESCs, with arrowheads indicating sizes of PCR products. Right: PCR genotyping results of WT vs. *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll1*KO cells. (D) RNA-seq tracks confirming deleted genomic regions (target locations indicated by vertical red bars) at the *Set1B*, *Set1A*, and *Mll2* (left) or *Mll1* (right) loci, showing successful generation of the two triple-mutant ESC lines. (E) Western blot comparing H3K4me3 levels across the following genotypes: *Mll1*KO, *Mll2*$^{\Delta SET}$, *Set1B*KO-*Set1A*$^{\Delta SET}$, *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll1*KO, *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ and WT. Total H3 served as the loading control. (F) Occupancy heatmaps of the TSS-proximal H3K4me3 peaks ranked in descending peak width in WT, *Set1B*KO-*Set1A*$^{\Delta SET}$, *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll1*KO, and *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ ESCs. (G) Corresponding log2FC heatmaps of the H3K4me3 TSS-proximal peaks for each of the indicated mutant lines relative to WT cells. At least two biological replicates of each genotype were analyzed. (H) Log2FC box plot of H3K4me3 levels at Set1-dependent H3K4me3 peaks (n=1,897) in *Set1B*KO-*Set1A*$^{\Delta SET}$, *Mll2*$^{\Delta SET}$, and *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ relative to WT. Indicated p-value was calculated using the Wilcoxon rank sum test with continuity correction (paired). P-value: a, p=8.688E-299. (I) Log2FC box plot of H3K4me3 at Mll2-dependent H3K4me3 peaks (n=5,163) in the same mutant genotypes as (H) relative to WT. P-value was calculated as in panel (H) P-value: a, p=2.851E-177. (J) Differential H3K4me3 peaks at TSSs between *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ (N=13,122 peaks) and WT (N=13,134 peaks) cells were visualized against all WT H3K4me3 peaks. The plot confirms that the appended *Mll2*$^{\Delta SET}$ mutation in the *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs severely affected both broad and narrow H3K4me3 peak breadth. P-value was determined as in Figure 3.6B.

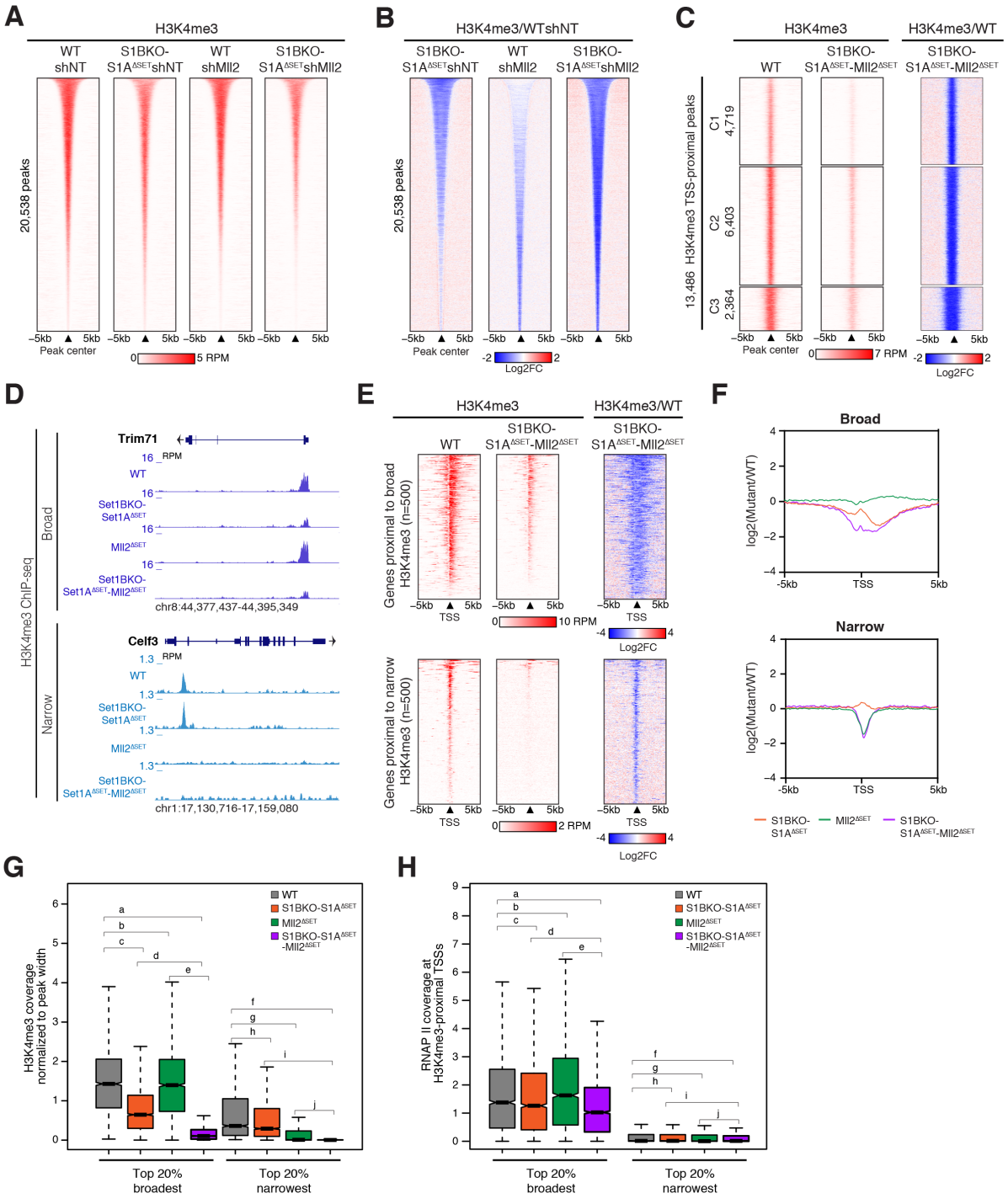
A
H3K4me3
WT shNT
S1BKO-S1A^ΔSETshNT
WT shMll2
S1BKO-S1A^ΔSETshMll2
20,538 peaks
Peak center
0 5 RPM
B
H3K4me3/WTshNT
20,538 peaks
Log2FC
C
H3K4me3
H3K4me3/WT
WT
S1BKO-S1A^ΔSET-Mll2^ΔSET
H3K4me3 TSS-proximal peaks
C1 4,719
C2 6,403
C3 2,364
13,486
Peak center
0 7 RPM
Log2FC
D
H3K4me3 ChIP-seq
Broad
Narrow
Trim71
Celf3
WT
Set1BKO-Set1A^ΔSET
Mll2^ΔSET
Set1BKO-Set1A^ΔSET-Mll2^ΔSET
chr8:44,377,437-44,395,349
chr1:17,130,716-17,159,080
E
Genes proximal to broad H3K4me3 (n=500)
Genes proximal to narrow H3K4me3 (n=500)
TSS
0 10 RPM
0 2 RPM
Log2FC
F
Broad
Narrow
log2(Mutant/WT)
-5kb
TSS
5kb
S1BKO-S1A^ΔSET
Mll2^ΔSET
S1BKO-S1A^ΔSET-Mll2^ΔSET
G
H3K4me3 coverage normalized to peak width
WT
S1BKO-S1A^ΔSET
Mll2^ΔSET
S1BKO-S1A^ΔSET-Mll2^ΔSET
Top 20% broadest
Top 20% narrowest
H
RNAP II coverage at H3K4me3-proximal TSSs
Top 20% broadest
Top 20% narrowest

Figure 3.8. Mll2 is functionally redundant to Set1/COMPASS in sustaining global H3K4me3 level and breadth.

(A) Heatmaps of H3K4me3 occupancy in WT and *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs upon control (shNT) and *Mll2* (sh*Mll2*) knockdowns. Profiles are plotted in decreasing H3K4me3 peak width. Two replicates of each genotype were analyzed. (B) Log2FC heatmaps comparing differences in H3K4me3 levels in WT-sh*Mll2*, *Set1B*KO-*Set1A*$^{\Delta SET}$ -shNT, or *Set1B*KO-*Set1A*$^{\Delta SET}$ -sh*Mll2* and WT-shNT conditions for the same ordered peaks as in (A). (C) Heatmaps of clustered H3K4me3 occupancy (left) and corresponding log2FC (right) comparing *Set1B*KO-*Set1A*$^{\Delta SET}$ -*Mll2*$^{\Delta SET}$ to WT cells at the 13,486 H3K4me3 TSS-proximal peaks initially determined in Figure 3.3B. (D) Track examples of broad (top) and narrow (bottom) H3K4me3 peaks in WT, *Set1B*KO-*Set1A*$^{\Delta SET}$, *Mll2*$^{\Delta SET}$, and *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ ESCs. (E) Heatmaps of H3K4me3 occupancy (left) and corresponding log2FC (right) at genes with the 500 most broad (top) and 500 most narrow (bottom) H3K4me3 peaks in *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ and WT ESCs. Broad and narrow peaks were initially determined in Figure 3.5A. (F) Average log2FC H3K4me3 signal density plots (RPM) at genes with the 500 most broad (top) and 500 most narrow (bottom) H3K4me3 peaks in *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$, *Set1B*KO-*Set1A*$^{\Delta SET}$, or *Mll2*$^{\Delta SET}$ relative to WT ESCs. (G) Box plot comparing H3K4me3 ChIP-seq density for the top 20% broadest or narrowest H3K4me3 peaks in WT, *Set1B*KO-*Set1A*$^{\Delta SET}$, *Mll2*$^{\Delta SET}$, and *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ cells. Normalized coverage, peak number per group, and p-values (indicated by lower case letters) were calculated as in Figure 3.5D. Calculated p-values are as follows: a, c, d, e, f, i, p<6.173E-282; b, p=0.01901; g, p=6.1730E-282; h, p=1.7914E-74; j, p=1.1087E-206. (H) Box plot showing RNAP II coverage at H3K4me3-proximal TSSs for the top 20% broadest or narrowest H3K4me3 peaks in WT, *Set1B*KO-*Set1A*$^{\Delta SET}$, *Mll2*$^{\Delta SET}$, and *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ ESCs. Peak number per group and p-values (indicated by lower case letters) were calculated as in panel (G). Calculated p-values are as follows: a, p=9.0609E-256; b, p=7.5707E-219; c, p=5.0143E-54; d, p=4.0960E-204; e, p<9.0609E-256; f, p=1.7272E-21; g, p=1.0845E-6; h, p=0.0017; i, p=6.9636E-38; j, p=0.4571.

Extending our analyses by evaluating H3K4me3 changes in the previously defined clusters, we observe a robust and synergistic decrease in H3K4me3 occupancy in all three clusters in the *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ triple-mutant cells (Figure 3.8C), evincing that Mll2/COMPASS is functionally redundant to Set1/COMPASS in sustaining global H3K4me3 levels in ESCs. In fact, when we analyzed H3K4me3 changes at Set1- vs. Mll2-dependent H3K4me3 regions in *Set1B*KO-*Set1A*$^{\Delta SET}$, *Mll2*$^{\Delta SET}$, and *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ compared to wild-type, we noticed the following: the significant decrease in H3K4me3 in the triple-mutant compared to that in the double-mutant at Set1-dependent H3K4 methylation is unquestionably greater than the significant reduction in H3K4me3 in the triple-mutant compared to that in the *Mll2*$^{\Delta SET}$ at Mll2-dependent sites (Figure 3.7H-I). This supports the interpretation that while Mll2 is functionally redundant to Set1 at Set1-controlled regions, Set1 does not appear to reciprocate redundancy to Mll2 at Mll2-controlled sites, suggesting a unidirectional compensatory relationship between Mll2 and Set1. The added mutation of *Mll2*$^{\Delta SET}$ in *Set1B*KO-*Set1A*$^{\Delta SET}$ ESCs also further perturbed H3K4me3 breadth, such that the reduction in H3K4me3 density is seen at both genes with broad or narrow breadth, with a more radical decrease at broader peaks in the triple-mutant (Figure 3.7J; Figure 3.8D-G; also compare Figure 3.5A to Figure 3.8E). These data pinpoint the role of Mll2/COMPASS to be a key compensator for Set1/COMPASS function in bolstering overall H3K4me3 level and breadth in ESCs.

Based on global analysis, we noticed an additional decrease in RNAP II at broader and narrower H3K4me3 peaks for the *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ versus *Set1B*KO-*Set1A*$^{\Delta SET}$ cells (Figure 3.8H). The severe loss of H3K4me3 in the *Set1B*KO-*Set1A*$^{\Delta SET}$-*Mll2*$^{\Delta SET}$ mutant at these regions correlates with a significant decrease in the transcription of cell identity genes compared

with wild-type and *Set1B*KO-*Set1A*$^{\Delta SET}$ cells, and as a result may contribute to the proliferation defect in these triple-mutant cells. Thus, in retrospect, the less dramatic decrease of RNAP II levels in the double-mutant at these broader peak regions depicted in Figure 3.5E may indicate that the remaining H3K4me3 at these loci is adequate to maintain expression of such cell identity genes to the levels that cell proliferation is not impaired. Furthermore, we previously demonstrated that proper ESC differentiation requires the catalytic activity of Set1A (*99*), which is a key contributor to the formation of broad H3K4me3. Therefore, we deduce that loss of H3K4me3 at cell identity genes in Set1-mutant cells have a direct effect on cellular differentiation. Since self-renewal and differentiation are hallmarks of pluripotent stem cells, the loss of differentiation potential in Set1-mutant ESCs reflects the impairment of their identity. Altogether, these data signify that H3K4me3 levels at broad domains play an instructive role in maintaining ESC identity.

An intriguing question stemming from the findings thus far pertains to the genomic localization of Set1/COMPASS and Mll2/COMPASS—particularly, whether their genome-wide occupancy changes upon perturbation of the individual methyltransferases. When we analyzed Set1A and Mll2 ChIP in WT ESCs, we noticed that Set1A occupancy is lower than Mll2 occupancy at Mll2-dependent H3K4me3 peaks, whereas both binding levels are similar at Set1-dependent H3K4me3 peaks (Figure 3.9A, C). This further supports the proposition of a unidirectional functionally redundant role between Mll2 and Set1 discussed earlier (Figure 3.7H-I). We then examined Mll2 binding in *Set1B*KO-*Set1A*$^{\Delta SET}$ compared to WT at TSS-proximal regions where Set1A and Mll2 overlap (“Overlapping”) or at Mll2 only regions (previously characterized in Figure 3.7A). As displayed in Figure 3.9B, Mll2 levels in the double-mutant vs. WT are comparable at both sets of genomic regions, indicating that Mll2 binding is independent

of Set1/COMPASS. We also analyzed Set1A binding in the context of Mll2$^{\Delta SET}$ by performing Set1A ChIP in WT and *Mll2$^{\Delta SET}$* ESCs. While we do not see changes in Set1A occupancy at Mll2 only TSS-proximal regions, we do observe a slight increase in Set1A binding at the overlapping sites (Figure 3.9D). One interpretation of this finding is that these overlapping sites are primarily dependent on Set1/COMPASS, and the slight increase in Set1A occupancy at these regions is needed to help maintain cellular identity by preserving H3K4me3 levels in the absence of Mll2 activity. Furthermore, by having both protein complexes co-localize to the same sites, ESCs are able to employ a "fail-safe mechanism" to ensure cell viability and maintenance.

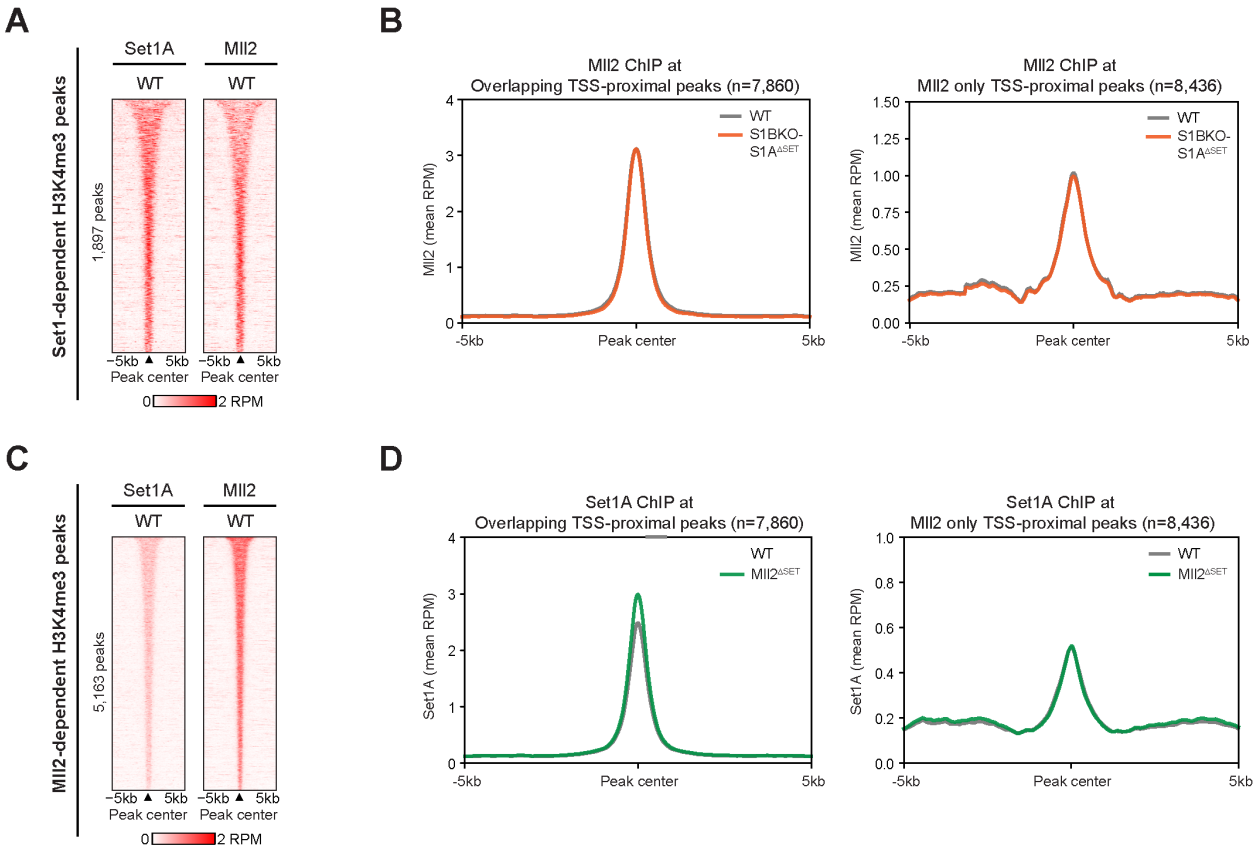

Figure 3.9. Set1A and Mll2 binding in the designated COMPASS mutant ESC lines compared to WT ESCs at the indicated peaks.

(A) Heatmaps of Set1A and Mll2 ChIP at Set1-dependent H3K4me3 peaks (n=1,897) in WT ESCs. Occupancy levels were plotted at WT peaks and sorted by decreasing peak width. (B) Metaplot of Mll2 binding in WT and *Set1BKO-Set1A*$^{\Delta SET}$ cells for the following peaks: 7,860 TSS-proximal peaks where Set1A and Mll2 binding overlap (left); 8,436 TSS-proximal peaks where only Mll2 binding occurs (right). Peaks were centered at WT Mll2 peaks. (C) Heatmaps of Set1A and Mll2 ChIP at Mll2-dependent H3K4me3 peaks (n=5,163) in WT ESCs. Occupancy was plotted as in panel (A). (D) Metaplot of Set1A binding in WT and *Mll2*$^{\Delta SET}$ cells for the same set of peaks as in panel (B). Peaks were centered at WT Set1A peaks.

In summary, we present findings delineating the functionally redundant roles of the COMPASS family members in H3K4me3 implementation in mammalian stem cells. By investigating H3K4me3 enrichment and breadth in a series of ESC lines harboring compounding mutations of COMPASS enzymes relative to wild-type cells, we report that COMPASS members Set1A, Set1B, and Mll2 coordinate regulation of H3K4me3 level and peak breadth across the mammalian genome. We also establish that Mll2 plays an important compensatory role in sustaining global H3K4me3 level and breadth in the absence of Set1/COMPASS. This study enhances our current knowledge, shedding new light on the extraordinary ability of our cells to adapt to contextual changes for sustainment. Moreover, with numerous studies indicating the role of COMPASS in disease pathogenesis—Set1A/Set1B in cancer (*100, 122*), Set1A in schizophrenia (*118, 127*), and Mll2 in childhood-onset dystonia (*133*)—our work offers insight into discovering potential targets for future therapy against these relevant diseases through characterization and assessment of the broad/narrow H3K4me3 epigenetic signature.

4 Guardians of pluripotency: Set1A teams up with other chromatin modifiers to protect the self-renewal state of embryonic stem cells

Work from this chapter is currently ongoing at the time of writing this dissertation and will be a published manuscript in the near future.

4.1 Introduction

Embryonic stem cells (ESCs) have two major defining features: 1) the extensive ability to self-renew, and 2) the potential to differentiate into all cell lineages, known as pluripotency. These unique characteristics bestow ESCs the fortune to serve as an effective platform for disease modeling and drug discovery research. In addition to a core network of transcription factors, self-renewing and pluripotent states are mediated by various epigenetic factors that encompass both chromatin remodelers and histone modifiers that promote a transcriptionally permissive environment (*31, 47, 200*). For that reason, it is necessary to tightly regulate the expressions and interactions of numerous proteins to prevent untimely differentiating events. However, how these proteins impact one another and the transcriptional program to jointly control ESC pluripotency remains to be thoroughly understood.

Among the COMPASS family of six histone H3 lysine 4 (H3K4) methyltransferases identified in mammals, Set1A is the only member whose full genetic knockout has been shown to be essential for ESC proliferation and self-renewal (*89, 94*). We previously demonstrated that although removing the catalytic SET domain of Set1A via CRISPR/Cas9 resulted in defective ESC differentiation, $Set1A^{\Delta SET}$ ESCs remained viable and could undergo proper self-renewal (*99*). It is likely that Set1A coordinates with other proteins and/or downstream effectors to regulate pluripotency; however, these additional functional interactors are currently undefined.

By leveraging a CRISPR/Cas9-based dropout screen approach, we sought to identify these novel factors that genetically interact with *Set1A*$^{\Delta SET}$ in an unbiased manner. Our genome-wide screen revealed several candidates that are synthetic perturbations to *Set1A*$^{\Delta SET}$ in ESCs, one of which is *Ing5*. Ing5 is a core subunit of three histone acetyltransferase (HAT) complexes (MOZ vs. MORF vs. HBO1) responsible for lysine acetylation on histones H3 and H4 (*201*). To date, only a handful of studies have implicated the role of Ing5 in preserving self-renewal, although all three HAT complexes show context-specific functional relevance in maintaining stem cell self-renewal. While our findings suggest that Set1A and Ing5 are cooperative gatekeepers of pluripotency, additional investigation is needed to elucidate the underlying mechanisms of how they jointly regulate ESC self-renewal. Investigating how epigenetic regulation governs ESC pluripotency is crucial for developing viable cell replacement therapies.

4.2 Results

Since deleting the enzymatic SET domain of Set1A does not perturb ESC self-renewal and viability (*99*), we performed a genome-wide CRISPR/Cas9 drop-out (negative selection) screen to identify genes that are required for cell viability of $Set1A^{\Delta SET}$ ESCs (Figure 4.1A). Wildtype (WT)- and $Set1A^{\Delta SET}$-Cas9-expressing cell pools were transduced with the Brie knockout library and cultured for 21 days to permit adequate time for the depletion of cells with sgRNA-perturbed essential genes. Genomic DNA from the initial and terminal populations (Day 3 and Day 21 post-transduction respectively) from both WT and $Set1A^{\Delta SET}$ lines was extracted, and sgRNA composition was measured by deep sequencing. Each sgRNA with a log2 fold change (log2FC) > -1 in the terminal population vs. initial population was considered as depleted, and the total number of depleted sgRNAs was computed. We found that essential genes such as *Pcna* and *Ctcf* were commonly dropped out in both cell lines during the 3-week screen and eliminated them from further candidate identification analyses. We outlined the workflow for sample processing following the dropout screen in Figure 4.1B.

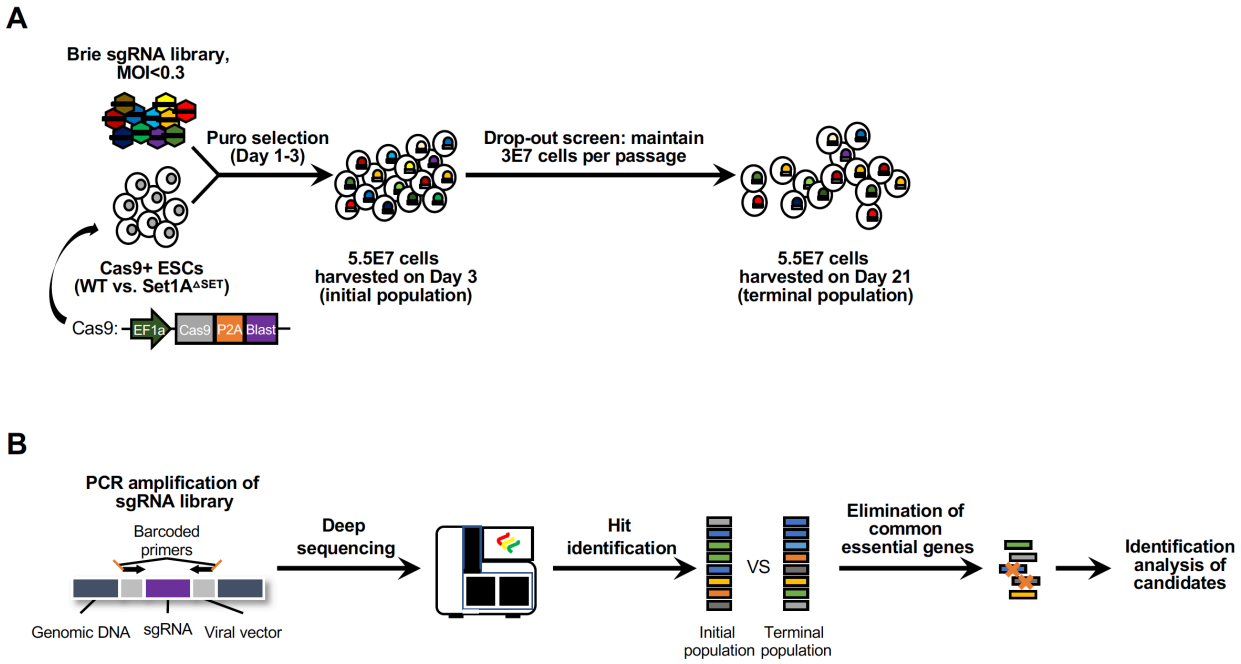

Figure 4.1. Genome-wide CRISPR dropout screen overview.

(A) Pooled Cas9-expressing WT vs. *Set1A*$^{\Delta SET}$ ESCs were transduced with Brie sgRNA library at MOI <0.3. Under puromycin selection, cells were passaged every 2-3 days and maintained at 3E7 cells per passage to ensure sufficient sgRNA representation. 5.5E7 cells were harvested at Day 3 (initial population) and Day 21 (terminal population) post-transduction. (B) Depleted candidate targets were identified following library amplification, Illumina sequencing, and elimination of common essential genes.

To retrieve the list of putative candidates synthetically lethal to *Set1A*$^{\Delta SET}$, we focused on genes that were depleted only in *Set1A*$^{\Delta SET}$ cells. Our screen identified an initial list of 1,425 targets that had at least 2 sgRNAs depleted out of the 4 sgRNAs represented in the library (Figure 4.2A). Gene ontology (GO) analysis of the 1,425 dropout genes revealed the most significant enrichment in pathways involved in DNA damage response & repair and chromatin regulation (Figure 4.2B). To identify the most essential genetic dependencies of *Set1A*$^{\Delta SET}$ ESCs, each target candidate was assigned a "Differential Dependency Score" (DDS), which reflects

both the degree and consistency of dropout across multiple replicates of the 21-day screen, and subsequently ranked. As shown in Figure 4.2A, 6 targets were found to be depleted in at least 3 out of 4 replicates. These targets, *Ing5*, *Uhrf1*, *Ints6*, *Ccnf*, *Pias4*, and *Dnmt1*, are all known to play a role in chromatin regulation and/or DNA damage response. As a result, we decided to focus our subsequent target validation efforts on the genes in the top two significant GO term categories (Figure 4.2B). Of the 1,425 ranked targets, 139 genes had at least 3 out of 4 sgRNAs depleted in a minimum of 2 replicates, from which 39 genes were selected for further validation based on their role in DNA-damage response and/or chromatin modification. We listed these select targets and color-coded them by their known link in Figure 4.3.

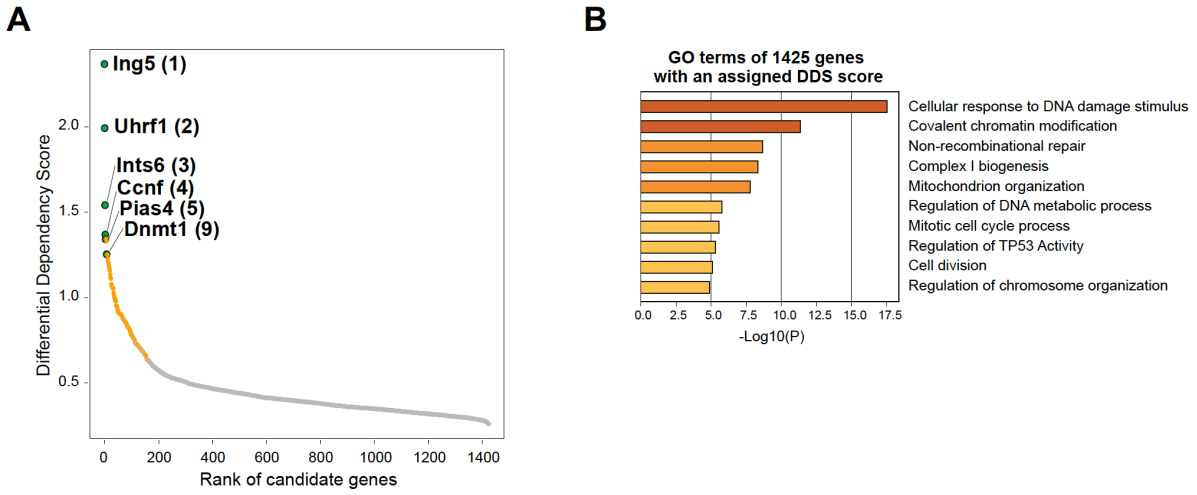

Figure 4.2. Identified genetic dependencies of *Set1A*$^{\Delta SET}$ ESCs.

(A) 1,425 dropout targets were ranked by their assigned "Differential Dependency Score" (DDS), which encompasses both the magnitude and reproducibility of dropout depletion across four replicates of the dropout screen. Ranked targets had at least 2 sgRNAs depleted out of 4 sgRNAs. Genes in green were depleted in at least 3 out of 4 replicates, and genes in orange were depleted in at least 2 out of 4 replicates. The 6 labeled targets correspond to the genes in green, with their ranking as determined by the DDS score in parentheses. (B) GO analysis for the 1,425 dropout targets in panel (A) using the Metascape software.

Rank	Gene	Rank	Gene	Rank	Gene	Rank	Gene
1	Ing5	20	Jade1	52	Stag1	93	Suv39h1
2	Uhrf1	27	Hmces	55	Dclre1b	95	Spidr
3	Ints6	28	Brcc3	72	Ascc1	112	Bcor
4	Ccnf	37	Lig4	74	Asxl1	117	Mettl4
5	Pias4	38	Smarcal1	77	Bcl7b	132	Tasp1
9	Dnmt1	40	Nfrkb	78	Nrd1	143	Kdm8
11	Fam98b	41	Brpf1	79	Hat1	148	Cdkn2aip
12	Supt20	43	Hand2	80	Thrap3	152	Phf12
13	Fam208a	44	Hus1	85	Ncoa3	164	Pphln1
19	Ep300	47	Actr8	91	Ddi2		

Legend
Chromatin
DNA-damage
Both

Figure 4.3. Select targets from the CRISPR dropout screen for validation.

Targets were selected based on the following criteria: 1) at least 3 sgRNAs depleted out of 4 sgRNAs in at least 2 out of 4 replicates, and 2) link to DNA-damage response (orange), chromatin modification (green), or both (yellow). Ranking as determined by their DDS score is also shown.

We individually targeted each of the 39 selected genes for functional perturbation via lentiviral sgRNA transduction in WT-Cas9 and *Set1A*$^{\Delta SET}$-Cas9 ESCs, followed by selection for 10 days. Following this, to alkaline-phosphatase (AP) staining was performed to assess cell colony growth and viability (Figure 4.4). One sgRNA, obtained from the Brie library, was used for each selected target. sgRNAs against *Pcna* and a non-targeting sequence (NTC) served as positive and negative controls respectively. We initially employed AP staining to perform a swift qualitative evaluation of target validation before resorting to a more robust validation assay. Specifically, we compared each target to the NTC for both WT and *Set1A*$^{\Delta SET}$ cells and then examined the extent of growth hindrance in *Set1A*$^{\Delta SET}$ over WT. Based on our assessment of the AP staining results, the following targets appeared to result in decreased proliferation of

Set1A$^{\Delta SET}$ ESCs compared to WT cells: *Ing5*, *Pias4*, *Bcor*, *Fam98b*, and *Nfrkb*. Consequently, these targets were selected for further subsequent validation.

Upon closer scrutiny of the 39 selected candidates (Figure 4.3), we noticed two targets known to be in complex with Ing5: Brpf1 and Jade1. Ing5 is a core component of two different histone acetyltransferase (HAT) complexes: MOZ/MORF or HBO1 complexes. These complexes include alternate scaffold proteins Brpf1/2/3 or Jade1/2/3 that help dictate acetylation specificity on the histone H3 or H4 tail (*201, 202*). The identification of *Brpf1* and *Jade1* as dropout targets in this negative selection screen increases the level of confidence that *Ing5* and its associated factors are genetic dependencies to *Set1A*$^{\Delta SET}$ in the context of ESC maintenance and viability. Therefore, we included *Brpf1* and *Jade1* in the list of select targets for further validation.

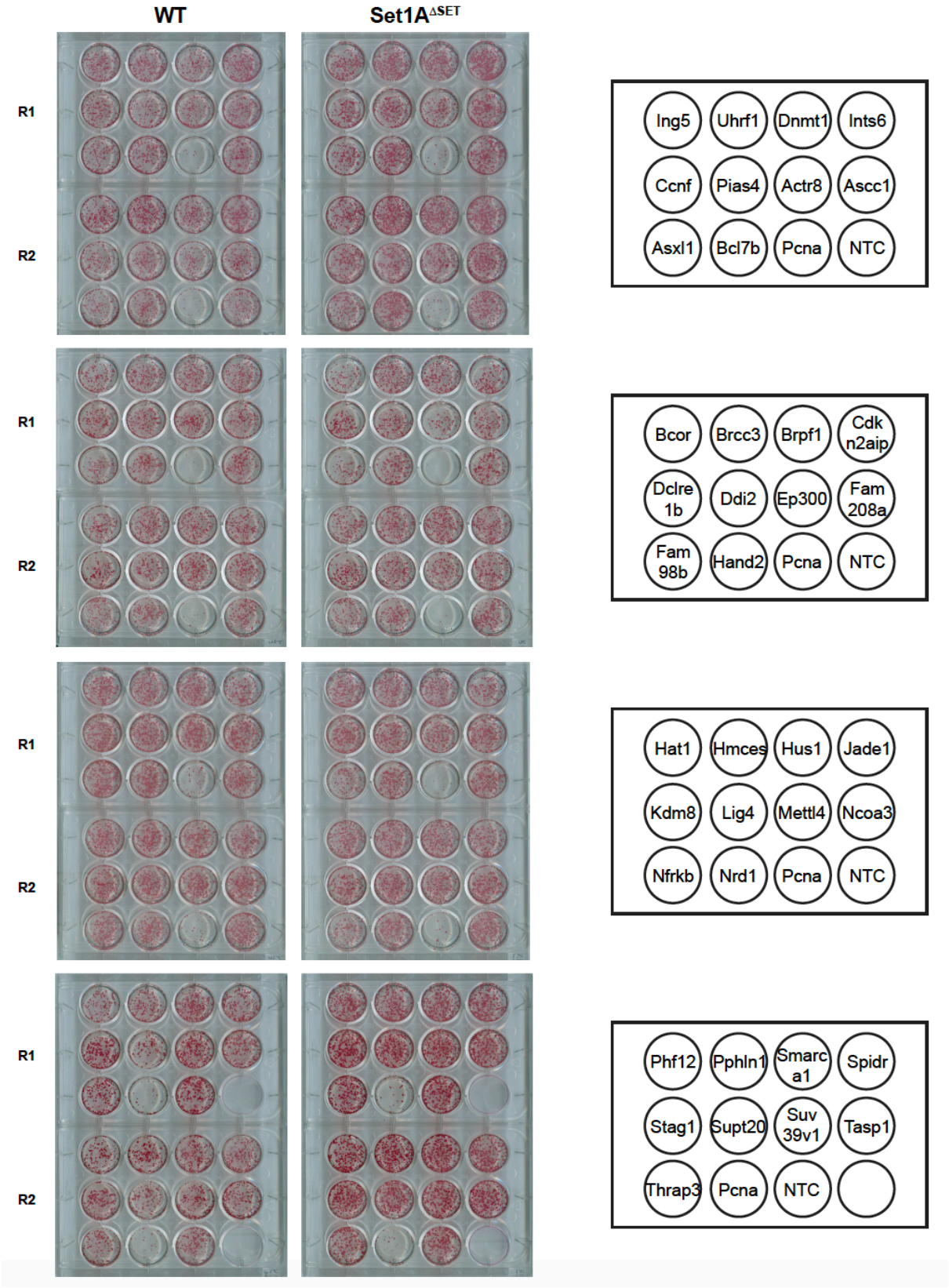
WT
Set1A^ΔSET
R1
R2
R1
R2
R1
R2
R1
R2
Ing5
Uhrf1
Dnmt1
Ints6
Ccnf
Pias4
Actr8
Ascc1
Asxl1
Bcl7b
Pcna
NTC
Bcor
Brcc3
Brpf1
Cdk n2aip
Dclre 1b
Ddi2
Ep300
Fam 208a
Fam 98b
Hand2
Pcna
NTC
Hat1
Hmces
Hus1
Jade1
Kdm8
Lig4
Mettl4
Ncoa3
Nfrkb
Nrd1
Pcna
NTC
Phf12
Pphln1
Smarc a1
Spidr
Stag1
Supt20
Suv 39v1
Tasp1
Thrap3
Pcna
NTC

Figure 4.4. Alkaline-phosphatase (AP) staining of WT- and *Set1A*$^{\Delta SET}$-Cas9 ESC colonies 10 days post-transduction with individual targeted sgRNAs.

One sgRNA (obtained from the Brie library) for each selected target was used. A PCNA sgRNA and non-targeting sgRNA (NTC) served as pan-essential vs. non-essential controls respectively. Two replicates (R1 vs. R2) per targeted sgRNA were performed per cell line. A map of the targeted genes is shown in the diagram on the right for each set of replicates.

We subjected *Ing5*, *Pias4*, *Bcor*, *Fam98b*, *Nfrkb*, *Brpf1*, and *Jade1* to additional validation by applying a CRISPR/Cas9-based cell competition assay (Figure 4.5A). For this assay, we first fluorescently labeled WT-Cas9 and *Set1A*$^{\Delta SET}$-Cas9 ESCs with mCherry and eGFP respectively, and then mixed the two labeled lines at a 1:1 ratio prior to transducing the mixed population with individual sgRNA of the select targets (Figure 4.5B). If disrupting a gene by sgRNA debilitates the fitness of *Set1A*$^{\Delta SET}$ ESCs but not WT ESCs, then the WT cells in the mixed population should outcompete the *Set1A*$^{\Delta SET}$ cells, and the fraction of eGFP+ cells should reduce over time. In this cell competition assay, we also included two genes, *Phf12* and *Pphln1*, which did not appear to be perturbed in the *Set1A*$^{\Delta SET}$ vs. WT from the AP staining assay (Figure 4.4), to serve as negative controls. Transduction with lentiGuide-Puro served as the non-targeting internal control. The cell competition assay was performed over 21 days, and fractions of mCherry+ vs. eGFP+ cells for each gRNA perturbation were scored every 2-3 days.

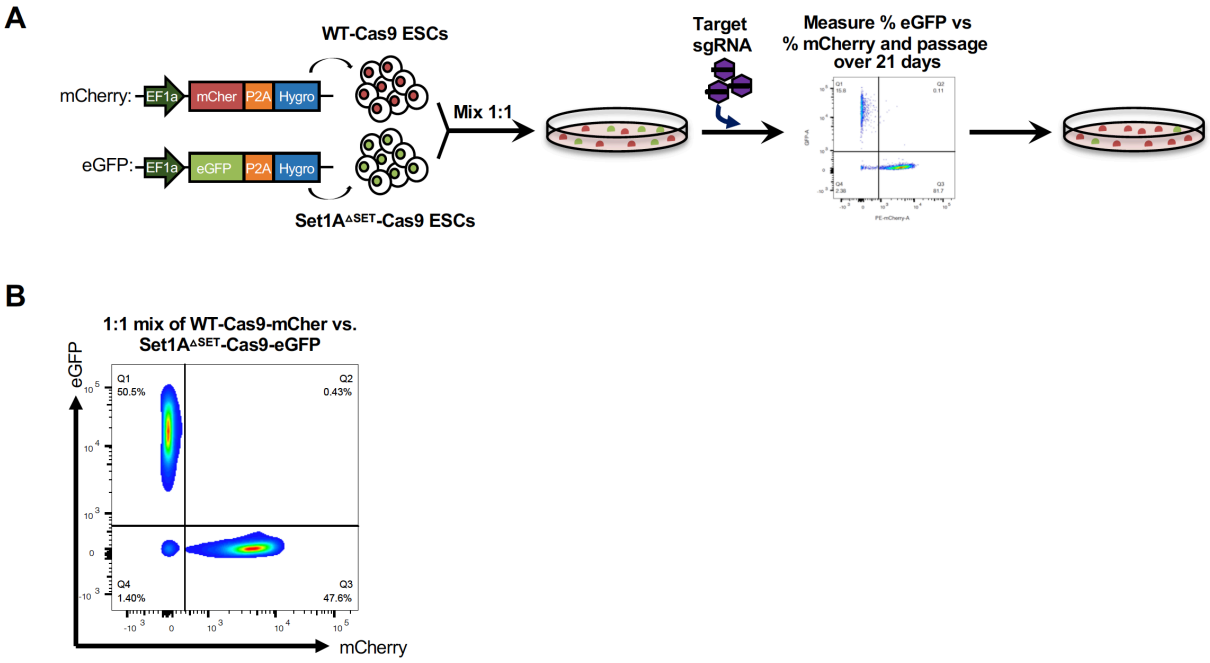

Figure 4.5. CRISPR-based competitive growth assay used to validate dropout candidates.

(A) Schematic overview of competitive growth assay. WT-Cas9 and *Set1A*$^{\Delta SET}$-Cas9 cells were labeled with mCherry and eGFP respectively, mixed at a 1:1 ratio prior to gRNA lentivirus transduction the next day. During the 21 day-screen, fractions of eGFP+ vs. mCherry+ were measured by flow cytometry. (B) Flow analysis showing 1:1 mix of WT-Cas9-mCherry and *Set1A*$^{\Delta SET}$-Cas9-eGFP at time of transduction.

To confirm which identified target is a true genetic dependency to *Set1A*$^{\Delta SET}$ in maintaining ESC viability, we calculated dropout ratios, which correspond to the percentage of *Set1A*$^{\Delta SET}$ cells over the percentage of WT cells for each sgRNA and normalized to Day 3, for each indicated timepoint (Figure 4.6). Despite the guide targeting *Brpf1* failing to drop out, guides targeting *Ing5* and *Jade1* were notably depleted in *Set1A*$^{\Delta SET}$ ESCs relative to WT ESCs. In addition, sgRNAs against *Pias4*, *Nfrkb*, and *Bcor* also dropped out in *Set1A*$^{\Delta SET}$ vs. WT cells. Consistent with AP staining findings, guides targeting *Phf12* and *Pphln1* did not reduce the

fitness of *Set1A*$^{\Delta SET}$ cells vs. WT cells in the competition assay. We do notice an interesting phenomenon with the *Fam98b* sgRNA, which initially appeared to promote competitive fitness of *Set1A*$^{\Delta SET}$ relative to WT cells before pivoting in the opposite direction. This demonstrates that a rigorous and methodical validation of each potential target identified in a genome-wide CRISPR screen is necessary. Furthermore, since we only used a single sgRNA per target in this cell competition assay thus far, testing additional sgRNAs to these select targets is warranted.

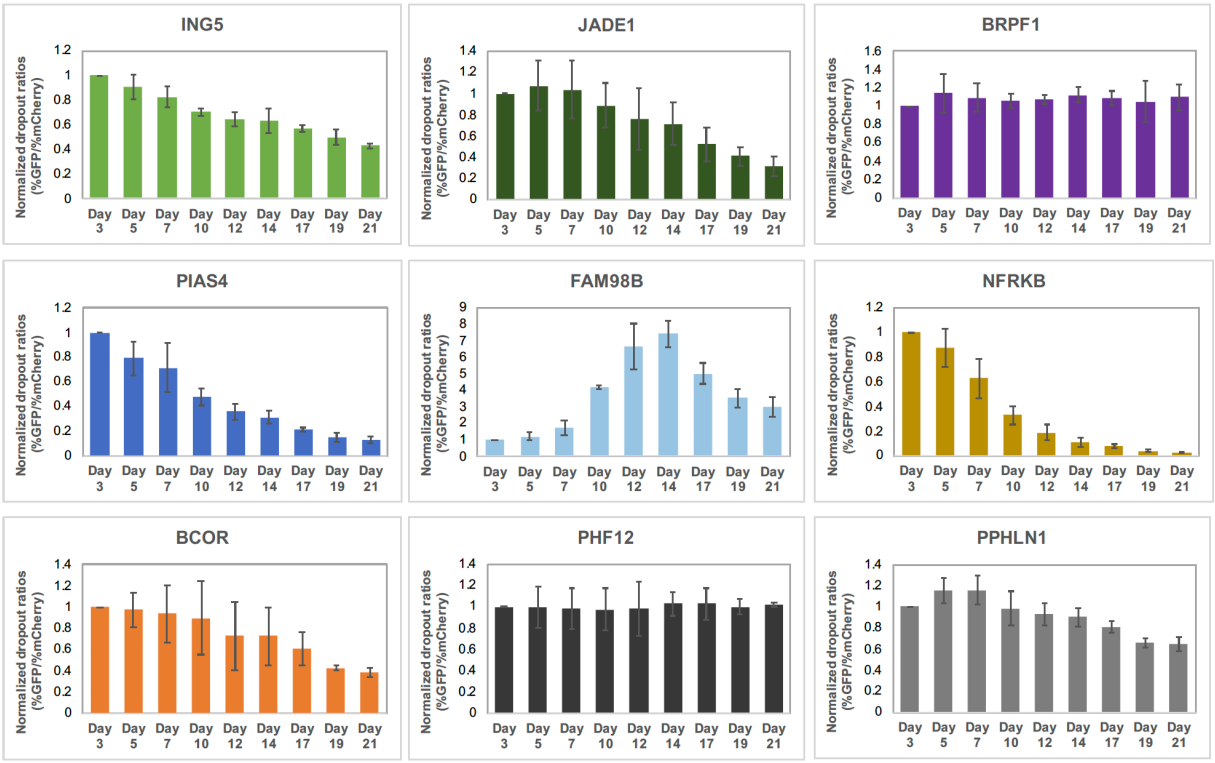

Figure 4.6. Cell competition assay to evaluate the essentiality of putative genetic dependencies to *Set1A*$^{\Delta SET}$.

Select targets for cell competition assay were based on initial validation results from AP staining (refer to relevant text for Figure 4.4). Dropout ratios (%eGFP/%mCherry, corresponding to %*Set1A*$^{\Delta SET}$-Cas9/%WT-Cas9) for each day noted were normalized to ratios for Day 3 and non-targeting control. Data are shown as mean ± SD (n = 3).

Moving forward in this study, we are choosing to focus our investigation on *Ing5* as a key genetic dependency to *Set1A*$^{\Delta SET}$ in regulating ESC viability for the following reasons: 1) *Ing5* was identified as the highest ranked target in the initial genome-wide dropout screen, with all 4 sgRNAs depleted across all 4 replicates; 2) its complex scaffolding partners *Jade1* and *Brpf1* were also targets uncovered from the genome-wide screen; and 3) both the AP staining and cell competition assays affirmed that perturbing the *Ing5* locus would reduce growth and fitness of *Set1A*$^{\Delta SET}$ ESCs compared to WT ESCs. As described earlier, Ing5 is a key subunit of the MOZ/MORF and HBO1 HAT complexes, with Moz/Morf (also known as Kat6a/Kat6b) and Hbo1 (also known as Kat7) possessing the enzymatic activity of their respective complexes (Figure 4.7A). To begin investigating the underlying mechanism of how Set1A/COMPASS cooperates with either MOZ/MORF or HBO1 complex to maintain ESC viability, we first assessed the functional effect of genetic depletion of the catalytic subunits. Using the cell competition assay, we found that targeting *Moz* appears to phenocopy the fitness effects of *Ing5* and *Jade1* depletion (Figure 4.6 and 4.7B). However, testing additional sgRNAs to these enzymes and investigating the consequential effects on relevant histone modifications are needed to unravel the mechanisms underlying the complexities of this epigenetic network.

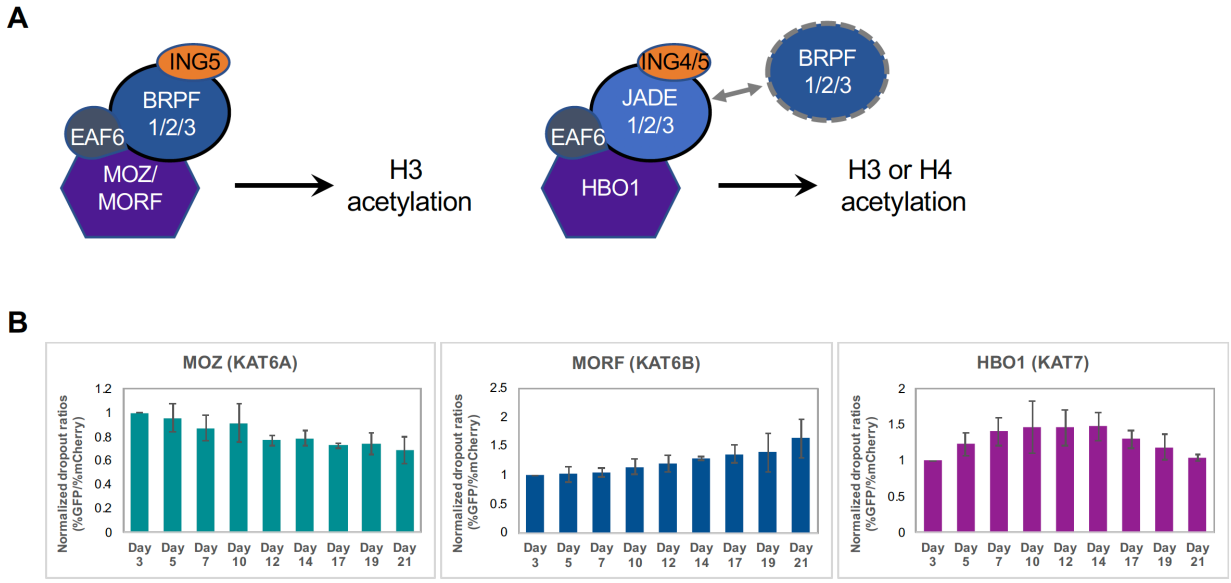

Figure 4.7. *Ing5* is a synthetic perturbation to *Set1A*$^{\Delta SET}$ in ESCs.

(A) Ing5 is known to be a part of the MOZ/MORF and HBO1 complexes, which are involved in acetylation at histone H3 and/or H4. (B) Cell competition assays with gRNAs targeting *Moz* (*Kat6a*), *Morf* (*Kat6b*), or *Hbo1* (*Kat7*) in WT-Cas9 and *Set1A*$^{\Delta SET}$-Cas9 ESCs. Dropout ratios for each indicated day were normalized to ratios for Day 3 and non-targeting control. Data are shown as mean ± SD (n = 3).

4.3 Discussion

The findings reported in this chapter thus far provide initial insight into the intricacy underlying epigenetic regulation of ESC self-renewal. During self-renewal, ESCs proliferate while preserving their replicative and broad developmental potential; as a result, this process is tightly regulated via careful coordination of multiple proteins and signaling pathways. Set1A is one of these proteins crucial for ESC viability (*89, 94, 98*), although removing its catalytic SET domain does not perturb self-renewal (*99*). Using a genome-wide CRISPR/Cas9 negative selection screen, we have identified several plausible genetic dependencies to *Set1A*$^{\Delta SET}$ in maintaining ESC viability.

Our screen revealed *Nfrkb* as a dropout candidate. Nfrkb (or Ino80g) is part of the INO80 SWI/SNF remodeling complex (*203*), which has been shown to be required for ESC self-renewal (*204*). It is postulated that INO80 maintains open chromatin structure at pluripotency gene promoters to recruit RNAP II and Mediator for transcriptional activation, and its genome-wide binding was shown to be dependent on the pluripotency factor Oct4 and COMPASS subunit Wdr5 (*204*). Pinpointing *Nfrkb* as a dependency to *Set1A*$^{\Delta SET}$ provides additional confidence in our screen and subsequent validation experiments. By uncovering the genetic interactions between Set1A/COMPASS and other chromatin modifying complexes, we provide an illustration of the complexity underlying the fail-safe mechanism that ESCs employ at the epigenetic level to preserve self-renewal; that is, ESCs appoint a network of guardians or gatekeepers to ensure tight control of self-renewal.

One genetic dependency identified from our dropout screen is *Ing5*, part of the evolutionarily conserved ING family of proteins that are known H3K4me3 readers targeting

histone acetyltransferase (HAT) or histone deacetylase (HDAC) complexes to chromatin (*205*). Ing5, which was first identified via sequence homology search to Ing4 (*206*), possesses a plant homeodomain (PHD) that recognizes the H3K4me3 mark, thus acting as an anchoring adaptor for HAT complexes at H3K4me3-enriched regions to promote local histone acetylation (*207*). Only a handful of studies to date have implicated the role of Ing5 in various stem cell populations with minimal understanding of the underlying mechanisms pertaining to its function: 1) overexpression of Ing5 in glioblastoma stem cells promotes their self-renewal, while *Ing5* knockdown results in a higher fraction of differentiated cells (*208*); 2) *Ing5* was identified via a focused genetic screen as part of an epigenetic network controlling epidermal stem cell maintenance (*209*); and 3) Ing5-HAT complexes (elaborated further below) have been shown to maintain pluripotency and proliferation of ESCs (*210*), hematopoietic stem cells (HSCs) (*211, 212*), and adult neural stem cells (NSCs) (*213*). Furthermore, Ing5 binding, shown by ChIP-seq, occurs primarily at TSS of transcribed genes and correlates positively with mRNA expression in epidermal stem cells (*209*). These findings collectively suggest that Ing5 may functionally overlap with Set1A to jointly control ESC viability and self-renewal in the current study. Therefore, we would expect that targeting both Set1A and Ing5 would result in a more perturbed expression profile and phenotype than targeting each gene individually, and that the consequences from Ing5 loss would phenocopy the effects from deleting the SET domain of Set1A in ESCs.

As stated previously in Section 4.2, Ing5 is a core component of the MOZ/MORF and HBO1 complexes. MOZ/MORF and HBO1 are three of five mammalian members of the highly conserved MYST family of tetrameric HAT complexes (Figure 4.7A), with subunits Moz

(Kat6a), Morf (Kat6b), and Hbo1 (Kat7) possessing the catalytic MYST domain responsible for acetylation by their respective complexes (*201, 214, 215*). Despite the breadth of published studies, there appears to be some confusion in the field on which specific histone modifications are attributed to MOZ/MORF vs. HBO1. MOZ/MORF complexes are believed to acetylate H3K9/K14/K23 at transcriptionally active promoters (*216-222*), while the HBO1 complex reportedly acetylates H3K14 at promoters and gene bodies (*211, 223-226*) and H4K5/K8/K12 (*202, 219, 227*). It was discovered for HBO1 that the histone tail specificity between H3 and H4 was directed by the precise composition of the HBO1 complex; in other words, native HBO1 complexes can contain either Jade1/2/3 or Brpf1/2/3 scaffold proteins to acetylate either H4 or H3 respectively (*202*). Associated complex proteins and experimental inconsistencies aside, it is possible that the local cellular environment and tissue-specific activities help limit the promiscuous activity of HATs, thereby resulting in conflicting reports in the field. For instance, *Hbo1* siRNA depletion in 293T cells led to decreased acetylation at H4K5/K8/K12 (*219*); however, genetic deletion of *Hbo1* resulted only in H3K14ac reduction in MEFs (*224*). Furthermore, while the MOZ complex has been shown to mediate H3K14ac (*219*), *Moz*-deficient mice exhibited diminished H3K9ac at certain genomic sites (*217*). It is therefore conceivable that there may be additional HAT subcomplexes to be unveiled; our preliminary findings from the competitive growth assay seem to suggest a possible MOZ-Jade complex in regulating ESC viability alongside Set1A (Figure 4.6 and 4.7B). Determining the consequent effects on histone modifications upon targeting *Moz* (and/or *Morf* and *Hbo1*) in $Set1A^{\Delta SET}$ ESCs may also help explain the underlying mechanism of epigenetic regulation.

Despite residing in similar multimeric complexes, each HAT has a unique role in development as reflected by the different phenotypes observed in mutant mice. *Hbo1*-null mice are embryonic lethal at E10.5; deleting *Hbo1* appears to adversely affect embryonic patterning and organogenesis, especially the development of blood vessels and somites (*224*). To date, complete deletion of *Morf in vivo* has yet to be reported, although hypomorphic *Morf* mutants are able to survive to birth with substantial craniofacial and neural defects (*228*). *Moz*-KO mutants have defective hematopoiesis and thus die at E15.5 (*229*), although inserting a *neo* cassette at exon 16 of *Moz* thereby disrupting the C-terminal transcriptional activation domain extends survival until birth, with altered homeotic transformation and segmental identity (*217*). The different *Moz* phenotypes highlight the value in generating mutants that address different functions of HATs (i.e. enzymatic activity vs. scaffolding function) and also points to the importance in evaluating results within specific contexts. It is possible that the catalytic activity of Moz, located in the first-half of the protein, would be functionally relevant in our current study given the more general requirement of the full protein vs. the truncated version without the transactivation domain *in vivo*. Future experimentation may encompass generating domain-specific mutants once the target HAT is identified to be functionally coordinating with Set1A in modulating ESC self-renewal and viability.

Since ESCs propagate quickly and indefinitely, it becomes imperative that they maintain genomic integrity to protect self-renewal and that continual DNA replication does not induce spurious differentiation (*230, 231*). Recently, several studies have demonstrated Set1A's role in orchestrating the DNA damage response and repair pathway to prevent genome instability, especially during replicative stress (*100, 123-125*). Interestingly, Ing5 and its associated

complexes were also reported to participate in DNA replication. In fact, HBO1 was first discovered via its interaction with ORC1 and MCM helicases, which are involved in the pre-replication complex (*232, 233*). Additional studies have linked the function of Ing5 and its associated complexes to p53 signaling, including physically interacting with p53 to activate p53-downstream targets (e.g. p21) in response to DNA damage (*206, 234, 235*). Ing5 has also been reported to regulate cell proliferation in a p53-independent manner (*236*). Given the current literature, it is reasonable to extrapolate that Set1A/COMPASS cooperates with Ing5-associated complexes to maintain ESC self-renewal via mechanisms involving regulating the DNA damage response. Indeed, crosstalk between the two families of chromatin-modifying complexes have been previously described: 1) Mll1/COMPASS and MOZ cooperate to regulate *Hox* genes in human cord blood cells (*237*); and 2) recruitment of HBO1 by Mll1 to regulate the *HoxA* gene cluster in leukemic stem cells (LSCs) (*211*). Interestingly, two independent studies have also implicated the involvement of HBO1 in transcriptional elongation (*211, 238*); in particular, H3K14ac deposition by HBO1 may facilitate RNAP II processivity throughout the coding regions of LSC genes (*211*). Since studies have shown that yeast Set1, homolog to mammalian Set1A and Set1B, could be recruited to chromatin by associating with elongating RNAP II (*83-85*), it is also therefore possible that the connection between Set1A/COMPASS and Ing5-related complexes is via the RNAP II elongation complex.

We present here initial findings and speculation on a previously unrecognized potential functional interaction between Ing5 and Set1A/COMPASS in ESCs, showcasing a sophisticated relationship among different families of epigenetic modifiers in mediating self-renewal. Clearly, much remains to be investigated about these underlying epigenetic mechanisms in governing

ESC pluripotency. Such molecular insights would be highly applicable for understanding the behavior of cancer stem cells, given their shared characteristics with ESCs, as well as stem cell reprogramming, which would greatly advance the field of regenerative medicine.

5 Concluding remarks

Parts of this chapter have been reproduced, with or without modifications, from my published review: Christie C. Sze and Ali Shilatifard. Cold Spring Harb Perspect Med. 2016, 6(11):a026427.

5.1 Summary of key findings and significance

Pluripotency and development are, in essence, governed by epigenetics. Despite the fact that all cells within our body bear identical genetic information encoded by the same set of genes, only a subset of those genes are expressed at a given time in a given cell of a given tissue. This level of meticulous control relies on the intricate interplay between different combinations of transcription factors and chromatin regulators. As differentiation/development advances, the combinations of post-translational modifications differ from pluripotent cells vs. progenitor cells vs. terminally differentiated cells. A thorough understanding of the transcriptional and epigenetic complexities in pluripotency will facilitate efficient use of ESCs, as well as use of induced pluripotent stem cells given their molecular similarity, in the field of regenerative medicine.

The COMPASS family and H3K4 methylation play central roles in dynamic gene expression programs and development, and numerous studies have indicated the role of COMPASS in disease (e.g. various COMPASS subunits in cancer, Set1A in schizophrenia, Set1B in epilepsy, Mll2 in childhood-onset dystonia, and Mll4 in Kabuki syndrome). Hence, key insights into the physiological activity and regulation will assist our understanding of their dysfunction in disease, and ultimately, facilitate the discovery of new targets for future therapy. To improve our knowledge of COMPASS biology, I studied the function of Set1A, one of the less well understood members of the COMPASS family, in stem cell pluripotency and development, which would help shed new insight on its pathogenic role consequent to mutation

and dysregulation. The following key findings were discussed in this dissertation:

1) deletion of the C-terminal enzymatic SET domain of Set1A neither perturbs ESC self-renewal nor bulk H3K4me3; however, H3K4 methylation by Set1A is necessary for proper differentiation of ESCs, showcasing context-dependent H3K4 methylation in regulating transcriptional outputs (Chapter 2);
2) the SET domain of Set1A is essential for embryogenesis; however, having this domain permits an extended period of viability for the embryos compared to *Set1A*-KO mutants (Chapter 2);
3) reflecting the tissue culture findings of embryoid body differentiation, removing the SET domain of Set1A perturbs mesoderm differentiation (Chapter 2);
4) Set1A and Set1B compensate for each other's function in implementing H3K4me3, which tends to have greater breadth and is associated with highly expressed genes (Chapter 3);
5) Depletion of *Mll2*, but not *Mll1*, affects different regions of H3K4me3 deposition associated with narrower breadth (Chapter 3);
6) Mll2 is functionally redundant to Set1/COMPASS in maintaining bulk H3K4me3 level and breadth in ESCs (Chapter 3);
7) Genome-wide dropout screen in WT vs. *Set1A*$^{\Delta SET}$ ESCs revealed that Ing5, a core component of several known complexes involved in histone acetylation, might functionally interact with Set1A$^{\Delta SET}$ in regulating ESC viability, revealing a potential crosstalk between COMPASS and HATs (Chapter 4).

In addition to the studies described in this dissertation, I have also collaborated with multiple colleagues to investigate the role of other COMPASS family members in ESCs,

development, and cancer. Much of our laboratory's research in the past several years has been devoted to elucidate catalytic-dependent vs. catalytic-independent functions of COMPASS and to illustrate context-dependent crosstalk among epigenetic modifiers in transcriptional regulation (*26, 75, 99, 101, 115, 239, 240*), an overarching theme delineated throughout this dissertation. As initially introduced in this dissertation, H3K4 methylation is an evolutionarily conversed histone mark that is prominently distributed throughout the genome. Historically, H3K4 methylation has been considered to be predictive of activating gene expression, although accumulating evidence is pointing to a more convoluted connection. While multiple functions have been ascribed to H3K4 methylation (discussed in Section 3.1), we now know that we cannot interpret the histone mark and its functional impact on transcriptional regulation without regards to the overall genomic and regulatory context.

5.2 Key outstanding questions in the field and future studies

Despite the tremendous amount of progress made over the past 40 years studying the COMPASS family of H3K4 methyltransferases in development and disease, many questions remain outstanding. For instance, are there additional non-catalytic functions of each COMPASS enzyme yet to be identified, and how exactly do they affect the complex's activity in development and disease? One key finding discussed earlier in this dissertation is that removal of Set1A's catalytic activity does not disrupt ESC viability, indicating that Set1A may act as a key scaffold protein, and that its genomic recruitment may serve a critical non-catalytic role in maintaining ESC pluripotency. As discussed in Sections 1.3 and 3.2, Cxxc1 is a known component of the Set1/COMPASS complexes (Figure 1.2). Cxxc1 contains a CXXC motif that recognizes and interacts with unmethylated CpG DNA, which consequently results in the selective targeting of the Set1/COMPASS complexes and limiting their methyltransferase activity to promoters (*75, 193*). Studies have shown that *Cxxc1*-null ESCs are viable and able to self-renew, although these ESCs are unable to differentiate (*193, 241*). The ability of *Cxxc1*-null ESCs to survive and self-renew suggests that there may be other proteins important in the recruitment and non-catalytic function of Set1A in regulating ESC pluripotency.

There are two proteins that are known Set1A interactors that could potentially be involved in Set1A recruitment and non-catalytic role to maintain ESC pluripotency: Wdr82 and Bod1L. For Wdr82, it has been shown that its depletion in HeLa (*67*) and HEK293 (*84*) results in a decrease in Set1A and H3K4me3 levels, and reduces Set1A occupancy at TSSs (*84*). Wdr82 has also been reported to be essential for embryogenesis, for *Wdr82* knockdown down-regulates expression of *Oct4* and induces high apoptotic rates of blastocysts and eventually resulting in

embryonic lethality (*242*). Furthermore, a very recent study reported that the yeast homolog of Wdr82, Swd2 (also known as Cps35), is important for the N-terminal association of yeast Set1 with RNAP II and COMPASS recruitment to chromatin (*87*). Given these data, it would be interesting if Wdr82 cooperates with Set1A to regulate ESC pluripotency. In the case of Bod1L, a Set1A-specific interactor (*101*), its role and relationship with Set1A has never been investigated in ESCs to date. Bod1L is part of the fork protection machinery to safeguard stalled or damaged replication forks from uncontrolled degradation (*165*), thus, playing an integral role in maintaining genomic stability. As mentioned in Section 1.3, it was reported that Set1A orchestrates DNA damage response and repair during replicative stress (*123, 124*). Together with the known fact that ESCs are notably sensitive to DNA damage (otherwise they quickly apoptose) (*231*), it is plausible that Set1A and Bod1L cooperatively regulate DNA damage and repair to help maintain ESC self-renewal.

In addition to elucidating Set1A recruitment and non-catalytic role, another possible future study stemming from this dissertation is to identify molecular targets of Set1A. The findings discussed in Chapter 2 implicate that Set1A is involved in mesoderm formation in the context of ESC differentiation and embryogenesis. However, much remains to be investigated regarding downstream targets of Set1A in ESC self-renewal. The findings in Chapter 4 provide preliminary insights into Set1A regulation of ESC self-renewal, though in the context of interplay with other chromatin regulators. One approach to identify molecular targets of Set1A is utilizing the auxin-inducible degradation (AID) system (*243*) to acutely deplete Set1A. To date, I have generated AID-tagged Set1A ESCs that express the TIR1 protein following a previously described methodology (*244*) (data not shown, and plasmids are gifts from B. Bruneau). By

comparing gene expressions between *Set1A*$^{\Delta SET}$ and AID-tagged Set1A ESCs, one would be able to detect non-overlapping targets unique to AID-tagged Set1A ESCs. These targets would be subjected to future loss-of-function analyses to elucidate the mechanistic environment surrounding Set1A.

Extending beyond Set1A, there are still key outstanding areas of research that would help further advance our understanding of the COMPASS family of methyltransferases. As exhibited in Figure 1.3, the majority of each COMPASS protein structure remains largely uncharacterized, and it is possible that these unannotated regions could confer non-catalytic COMPASS function beyond H3K4 methylation. Furthermore, the divergent mouse phenotypes observed with deleting full-length Set1A or Mll1 vs. solely targeting the SET domain clearly demonstrates domain-specific functions of COMPASS in development, and additional research efforts are necessary to fully elucidate these *in vivo* roles. In addition to defining new structural motifs, there could be additional unidentified factors involved in mobilizing COMPASS to chromatin for transcriptional regulation. With the recent discovery of cytoplasmic Set1B and its function in cancer cell metabolism (*101*), it is possible that cytoplasmic roles of other COMPASS members remain to be determined in both physiological and pathological conditions.

As noted in the multitude of studies elaborated earlier in this dissertation, crosstalk occurs among chromatin complexes and their histone marks to regulate transcription. Consequently, the mechanistic relationship between post-translational modifications and the enzymatic complexes responsible for their deposition is increasingly being examined in a context-dependent manner. However, additional research is necessary to provide a comprehensive map and interpretation of these complex interactions. Understanding, for example, how COMPASS complexes

functionally coordinate with other chromatin regulators, histone marks, and transcription factors, and the consequential impact from these interactions would illuminate how our cells employ intricate mechanisms to ensure homeostasis and viability. The rapid advancement of next-generation sequencing tools, combined with high-throughput CRISPR/Cas9-based screening approaches to target chromatin regulators and their modifications, will enable in-depth analyses of spatiotemporal changes in chromatin dynamics genome-wide in a tissue-specific manner.

Collective understanding of chromatin regulators such as the COMPASS family and their biological roles at the molecular and cellular levels could provide insight into how their mutations aberrantly elicit gene expression to subvert cellular identity and promote disease. For instance, how do the underlying mechanisms of COMPASS mis-regulation differ across cancer and neurodevelopmental disorders? Could mutations at particular domains within the COMPASS proteins have distinct molecular ramifications, such as perturbed recognition of appropriate marks necessary for recruitment or altered catalytic activity? Are there certain interactors that are themselves mutated to dysregulate COMPASS recruitment and activity to cause pathogenesis? The answers to these questions will be crucial to understand the full spectrum of COMPASS contributions to disease and help provide key therapeutic strategies for targeting COMPASS-mediated pathologies. With the rapid development of next-generation sequencing methods, determining novel mutations will be important to elucidate signaling pathways likely deregulated by alterations of these chromatin-modifying enzymes. Altogether, these studies will ultimately provide critical insights to facilitate the identification of therapeutic opportunities for disease.

6 Materials and methods

6.1 Murine embryonic stem cell (ESC) culture, CRISPR/Cas9-mediated gene editing, embryoid body (EB) formation, short hairpin (shRNA) knockdown, and generation of Cas9-expressing lines

Mouse V6.5 ESCs were cultured in N2B27 media supplemented with MEK inhibitor and GSK3 inhibitor and LIF (2i/LIF) as previously described (*36*). EBs were generated via the hanging drop method, where 1,500 ESCs were cultured in 25uL of EB differentiation medium on the lid of 150mm culture plates for six days. EB differentiation medium composed of DMEM supplemented with 15% FBS (Gemini Bio-Products), 1x GlutaMAX (Gibco), 1x MEM nonessential amino acids (Gibco), 1x ß-mercaptoethanol (Gibco), 1x penicillin-streptomycin (Gibco). N2B27 monolayer differentiation were performed as previously described (*162*). For retinoic acid (RA)-induced differentiation, ESCs were grown in N2B27 medium with 1μM all-trans RA (Sigma) without 2i/LIF, and RA treatments were performed for 24 hours. Vectors expressing Cas9 and CRISPR sgRNAs were generated by annealing and cloning oligos encoding the desired sgRNA sequences into the pX459 plasmid following a previously published protocol (*245*). Subsequent ESC transfection was performed as follows: 5 million ESCs were electroporated with the CRISPR sgRNA targeting vectors using Lonza's Nucleofector2b (following manufacturer's instructions). One day post-transfection, ESCs were selected with 2ug/mL puromycin (Life Technologies) for two days and grown in 2i/LIF ESC medium for ten days until clone picking. Targeted ESC clones were screened by PCR and confirmed by Sanger sequencing, RNA-seq, and/or western blotting. CRISPR sgRNA oligo sequences and genotyping primers are listed in Appendix 7.1 and 7.2 respectively. Knockdown was performed as described previously (*26*). The lentiviral construct against *Mll2* was reported previously (*70*). To generate

Cas9-expressing lines, WT V6.5 and *Set1A*$^{\Delta SET}$ ESCs were transduced with lentiCas9-Blast virus (Addgene #52962) (*246*). 24 hours post-transduction, cells were selected in 4μg/mL blasticidin (Life Technologies) for 5 days until no viable cells were observed in an untransduced control plate treated with blasticidin. To test for Cas9 activity efficiency, pooled Cas9 cells were transduced with lentiviral plasmid expressing eGFP and its sgRNA (Addgene #107145) followed by puromycin selection, eGFP expression was measured every 2 days over a period of 7 days post-transduction. Pooled Cas9 cells were then expanded under blasticidin selection up to their subsequent transductions with the Brie mouse library or individual targeted sgRNAs (see Section 6.8 below).

6.2 Antibodies

The following antibodies were generated in house: anti-H3, anti-H3K4me1, anti-H3K4me2, anti-H3K4me3, anti-Set1A(*101*), anti-Mll2, and anti-Set1B (*101*). Other antibodies used in this work are: H3 (CST #1B1B2), H3K4me3 (CST #9727), Rpb1 (CST #D8L4Y), HSP90 (Santa Cruz #7947), and Rbbp5 (Bethyl Labs #A300-109A). Set1A ChIP-seq data in Figure 3.3B were from the Sze et al. 2017 study (*99*).

6.3 Alkaline-phosphatase (AP) staining and imaging

AP staining was performed on ESCs using the Red AP Substrate Kit (Vector Labs) following the manufacturer's directions. ESCs and EBs were captured using the Nikon Eclipse Ts2R microscope and DS-Qi2 camera. EB sizes were measured using Nikon's NIS-Elements Basic Research software.

6.4 Quantitative RT-PCR

Total RNA was extracted using RNeasy mini kit (Qiagen) following manufacturer's instructions, and cDNA was synthesized using the High Capacity RNA-to-cDNA Kit (Applied Biosystems). Resulting cDNA levels were measured on a CFX connect Real-Time PCR detection system (Bio-Rad) using the Maxima SYBR Green/ROX qPCR Master Mix (ThermoFisher), and relative expression to Actin was calculated. Primers used in qRT-PCR assays are listed in Appendix 7.3.

6.5 Cellular fractionation

Harvested cells were pelleted and washed twice with 1xPBS, followed by resuspension in ice-cold Buffer A (10mM Hepes pH7.9, 1.5mM MgCl2, 10mM KCl, 0.5mM DTT, 1:100 Sigma P8340 protease inhibitor cocktail) and incubated on ice for 10 minutes, vortexing every 5 minutes. 10% NP-40 was added to samples followed by another 10-minute incubation on ice and vortexing every 5 minutes. After a 10-minute centrifugation at 400g at 4^{o}C, the supernatant was collected as the cytoplasmic fraction. The remaining nuclei pellet was washed with ice-cold Buffer A before resuspension and 15-minute incubation on ice with ice-cold Buffer B (20mM Hepes pH7.9, 1.5mM MgCl2, 420mM NaCl, 25% (v/v) glycerol, 0.2mM EDTA, 0.5mM DTT, 1:100 Sigma P8340 protease inhibitor cocktail), vortexing every 5 minutes. Samples were then spun down at 4000g for 10 minutes at 4^{o}C, and the supernatant was harvested as the soluble nuclear fraction. The remaining pellet was then washed with ice-cold Buffer B, before being resuspended in 1xPBS containing 1% SDS, sonicated for 10 seconds, and centrifuged at

maximum speed for 10 minutes at 4°C. The resulting supernatant was collected as the insoluble nuclear fraction.

6.6 Western blotting

Cells were lysed directly in 2x sample buffer (with loading dye and 10% 2-mercaptoethanol) and boiled at 95°C for 10 minutes. Lysates were run by SDS-PAGE and transferred to 0.45µm nitrocellulose membranes for 90 minutes at 350 milli-Amps at 4°C. Membranes were then stained with Ponceau S solution (Sigma #P7170), followed by blocking in 5% milk in 1xTBST for 1 hour at room temperature and overnight primary antibody incubation at 4°C. Primary antibodies (as listed in Section 6.2 above) were diluted in 5% milk in 1xTBST. After primary antibody incubation, membranes were washed 3 times for 10 minutes with 1xTBST at room temperature and then incubated with horseradish peroxidase conjugated secondaries for 1 hour at room temperature (Sigma #A4416 and #A6154). After 3 more washes with 1xTBST, membranes were imaged using Immobilon Crescendo Western HRP substrate (ThermoFisher) and ChemiDoc Imaging System (Bio-Rad).

6.7 Generation of *Set1A$^{\Delta SET}$* mouse line and genotyping

Set1A$^{\Delta SET}$ mutant C57BL/6 mice were generated via pronuclear injection of CRISPR sgRNAs to the SET domain of Set1A (refer to Appendix 7.1) with the assistance of the Northwestern University Transgenic and Targeted Mutagenesis Laboratory (TTML). The resulting F0 founder mice were genotyped using PCR and next generation sequencing (NGS) to

identify the F0 mice harboring the *Set1A*$^{\Delta SET}$ mutation. In brief, genomic DNA was extracted from tail snips (provided by TTML) of resulting F0 mice. We designed NGS primers that amplify the intended sgRNA target region of *Set1A* to include Illumina adaptor sequences, a staggering length sequence, and an 8bp barcode for multiplexing of different F0 samples:

Forward primer:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT (1-9bp staggering length sequence) GGAAGAAGAAACTCCGATTTGG

Reverse primer:

CAAGCAGAAGACGGCATACGAGAT (unique 8bp barcode) GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ACCATCTCATCAGCGGCAAT

For each F0 sample, we used Phusion® High-Fidelity DNA Polymerase (NEB #M0530) for PCR amplification (35 cycles) and then combined the resulting amplicons across samples. Pooled PCR reactions were precipitated using isopropanol, gel extracted, and then sequenced using the NextSeq 500 Sequencing platform (Illumina). Raw BCL output files converted into fastq files using bcl2fastq (Illumina, version 2.17.1.14), followed by quality trimming using Trimmomatic (*247*). Trimmed reads were then aligned to the mouse genome (UCSC mm9) using Burrows-Wheeler Aligner (BWA) (*248*). Output BAM files were converted into SAM files using SAMtools (*249*), from which CIGAR strings were retrieved for each F0 sample and subsequently analyzed to determine which F0 mouse harbored an intended mutation in the Set1A SET domain. One F0 mouse harboring a two-nucleotide insertion at the gRNA cut site located at the start of the SET domain of Set1A was identified and subsequently crossed with wild-type (WT) female C57BL/6 mice, and the resulting offspring validated the germline transmission of the mutant

allele. Heterozygous breeding was used to establish and maintain the mouse colony, and heterozygous intercrosses were carried out to obtain progeny with homozygous *Set1A*$^{\Delta SET}$ mutation. Developmentally-staged embryos from heterozygous intercrosses were dissected, genotyped, and characterized for developmental deformities. For mouse genotyping post-colony establishment, mice were ear notched, and the ear notch biopsies were used for genotyping following a previously published protocol (*250*). Genotyping primers are listed in Appendix 7.4.

6.8 Genome-wide CRISPR/Cas9 dropout screen

The Brie mouse library (*251*) was purchased from Addgene. This library targets each of the 19,674 mouse genes with approximately 4 sgRNAs per gene, plus 1,000 non-targeting control sgRNAs (*251*). Brie library amplification, lentiviral production, multiplicity of infection (MOI) determination, and transduction were performed as previously described (*252*). In brief, 3E7 WT V6.5 and *Set1A*$^{\Delta SET}$ ESCs stably expressing Cas9 (hereby known as WT-Cas9 and *Set1A*$^{\Delta SET}$-Cas9 respectively; refer to Section 6.1 above) were transduced with the Brie library at MOI<0.3. 24 hours after transduction, infected cells were treated with puromycin (2μg/mL, Life Technologies), and 5.5E7 cells were pelleted and snap-frozen 2 days later (day 3). Remaining cells were passaged every 2 days for an additional 18 days, during which at least 3E7 cells were maintained per passage to ensure adequate sgRNA representation. On day 21, 5.5E7 cells were pelleted and snap-frozen. Four replicates of this 21-day dropout screen were performed. As described previously (*252*), genomic DNA was extracted from pelleted cells collected on day 3 and day 21, which serve as the initial and terminal populations of transduced cells respectively, and the sgRNA library was amplified from the extracted DNA by PCR with primers containing

adaptors for Illumina sequencing. Deep sequencing on the NextSeq 500 Sequencing System followed by statistical analyses were used to analyze sgRNA library composition in the initial and terminal populations. In brief, identification of dropout candidates was determined by comparing the sgRNA makeup between the initial and terminal populations for each genotype, WT and *Set1A*$^{\Delta SET}$. The dropout of individual guides was calculated as the log2 fold change of the terminal over the initial population, with each sgRNA with a log2FC > -1 considered as depleted. Each gene was then classified by the number of sgRNAs depleted out of 4 sgRNAs. Essential genes (e.g. *Pcna*, *Ctcf*) that were commonly depleted in both cell lines during the 21-day screen were discarded from further candidate analyses. To obtain the initial list of candidates that are synthetically lethal to *Set1A*$^{\Delta SET}$, we focused on genes that were dropped out only in *Set1A*$^{\Delta SET}$ cells. To pinpoint the most important dropout targets in *Set1A*$^{\Delta SET}$ cells, genes were ranked by their assigned "Differential Dependency Score" (*DDS*) as shown below, which reflects both the magnitude and reproducibility of dropout depletion across all four replicates.

$$DDS\ (gene) = -\sum_{i=1}^{4} W_i * \log_2 FC^{\mathrm{Set1A\Delta SET}}$$

where $W_i = 1$ if there are at least two sgRNAs of gene meet the criteria: $-log2FC^{\mathrm{Set1A\Delta SET}} > 1$ and $-log2FC^{\mathrm{WT}} < 1$, otherwise $W_i = 0$.

6.9 Validation of screen candidates: alkaline-phosphatase and cell competition assays

Among the candidate genes identified, 39 genes involved in chromatin regulation and/or DNA damage repair were selected for validation. For each selected gene, individual targeted sgRNAs were cloned into the lentiGuide-Puro plasmid (Addgene #52963) (*246*) as described

previously (*252*). Cloned constructs were used for lentivirus production and transduction of WT-Cas9 and *Set1A*$^{\Delta SET}$-Cas9 ESCs, which then underwent puromycin selection for 10 days prior to AP staining (refer to Section 6.3 above) and qualitatively evaluated for affected cell colony proliferation and viability. Two replicates of AP staining evaluation were performed. Select targets resulting in perturbed *Set1A*$^{\Delta SET}$ proliferation determined by AP staining were subjected to further validation using a cell competition assay. For the cell competition assay, we first labeled WT-Cas9 and *Set1A*$^{\Delta SET}$-Cas9 ESCs with mCherry and eGFP. The plasmid for expressing mCherry in ESCs was purchased from Addgene (#120426) (*253*), into which we cloned the eGFP transgene in place of mCherry to generate the eGFP-expressing plasmid (*253*). mCherry-labeled WT-Cas9 cells and eGFP-labeled *Set1A*$^{\Delta SET}$-Cas9 cells were mixed at a 1:1 ratio and seeded 12 hours before gRNA lentivirus transduction. 24 hours post-transduction (day 1), infected cells were selected with puromycin. The percentages of mCherry+ vs. eGFP+ cells per gRNA perturbation were measured between day 3 and day 21 after transduction via flow cytometry using the BD FACSAria II. At least two replicates of the 21-day cell competition assay were performed as part of the target validation process. Flow cytometry analyses, including gating, were performed on FlowJo v10.6.2.

6.10 RNA-seq, ChIP-seq, and next generation sequencing data processing

RNA from ESCs was isolated using the RNeasy mini kit, and RNA from E8.5 embryos was extracted using the AllPrep® DNA/RNA Micro kit, following Qiagen's instructions. RNA-seq libraries were prepared using the TruSeq Stranded Total RNA Preparation Kit (Illumina). ChIP was performed as previously described (*99, 254*). In brief, ESCs were fixed in 1%

formaldehyde, followed by quenching, cell lysis, and chromatin shearing with an E220 focused-ultrasonicator (Covaris). Sonicated chromatin was subsequently subjected to immunoprecipitation. Immunoprecipitated DNA was then washed, eluted, reverse-crosslinked, and purified prior to library preparation using the KAPA HTP library preparation kit (KAPA Biosystems). ChIP-seq experiments in EBs were performed similarly except EBs were dounce-homogenized prior to sonication. RNA-seq and ChIP-seq libraries were single read sequenced on the NextSeq 500 or NovaSeq 6000 (Illumina). Raw BCL output files were processed using bcl2fastq (Illumina, version 2.17.1.14) prior to quality trimming using Trimmomatic (*247*). Trimmed ChIP-seq and RNA-seq reads were aligned to the mouse genome (University of California at Santa Cruz [UCSC] mm9) using Bowtie v1.1.2 (*255*) and TopHat v2.1.0 (*256*) respectively. Only uniquely mapped reads satisfying the two-mismatch maximum threshold within the entire length of the gene were considered for ensuing analyses. Mapped ChIP-seq reads were extended to 150bp to represent sequenced fragments. Raw read counts from both ChIP-seq and RNA-seq were normalized to total reads per million (RPM). Output BAM files were converted into bigwig coverage plots to generate UCSC genome browser tracks. For RNA-seq, exonic reads were assigned to specific genes from Ensembl release 72 using the python package HTSeq-0.6.1 (*257*). At least two biological replicates were performed for RNA-seq and ChIP-seq under each experimental condition.

6.11 RNA-seq and ChIP-seq analyses

For ChIP-seq: Peaks were called using MACS v1.4.2. with default parameters. Heatmaps and metaplots in Chapter 2 were generated using ngsplot (*258*), and heatmaps and metaplots in

Chapter 3 were generated using deepTools 2.0 v3.1.1 (*259*). Occupancy levels in RPM were observed over peak regions as denoted, and were either ranked by peak width or partitioned by K-means clustering. Differential heatmaps show log2FC in occupancy of mutant cells relative to WT cells. Metaplots illustrate average ChIP-seq occupancy in RPM. Genome-wide Set1A distribution in Figure 2.4A was determined by ChIP-seq and calculated by HOMER (*260*). For the evaluation of differential H3K4me3 occupancy levels shown in the MA plot in Figure 2.8A, BEDTools (*261*) was used to quantify the read counts within peaks, edgeR v3.0.8 (*262*) was used to evaluate statistical differences, and custom R scripts were used for data plotting. For the classification of broad and narrow H3K4me3 peaks in Figure 3.5 and Figure 3.8, we overlapped broad peaks called using SICER v1.1 (*263*) and narrow TSS-proximal peaks called using MACS, retaining the peaks called using SICER that overlapped with narrow TSS-proximal peaks. Peaks were mapped to a TSS using the GREAT algorithm (*161*), where TSS-proximal peaks were required to be located within 5kb from the identified TSS. To determine the H3K4me3 levels for breadth analyses shown in Figure 3.5D, Figure 3.8G, Figure 3.8H, Figure 3.6B, Figure 3.7J, and Figure 3.7H-I, the BEDTools map function was used to quantify the coverage over peak regions, and custom R scripts were used for data plotting. The differential H3K4me3 peaks shown in Figure 3.6B and Figure 3.7J for the respective mutant lines were defined as having a log2FC of H3K4me3 peak coverage of < -1 relative to WT peak coverage. To identify Set1-dependent H3K4me3 and Mll2-dependent H3K4me3 peaks in Figure 3.7H-I, H3K4me3 peaks were first called from at least two WT biological replicates using MACS, and overlapping regions determined using BEDtools intersect were chosen for subsequent differential binding analysis. To define sites with differential binding between WT and the indicated mutant line, BAM files of H3K4me3 ChIP-seq from WT and mutant samples were converted to a bed file using BEDTools

v.2.29.1. BEDtools coverage was then used to quantify read counts within the initially identified overlapping H3K4me3 regions in WT and mutant samples. Using the edgeR v3.12.1 (*262*) package from Bioconductor, the quantified counts were normalized to counts per million, and differential peaks were then identified by fitting a negative binomial generalized log linear model to the normalized counts data. Statistically significantly decreased peaks in the mutant cell lines were identified using Benjamini-Hochberg adjusted p-values <0.01 and log2FC < -1. Set1- or Mll2-dependent H3K4me3 peaks were determined by overlapping the respective decreased H3K4me3 regions with Set1A or Mll2 binding.

For RNA-seq: for differential gene expression, gene count tables from HTSeq were used as input for edgeR v3.0.8 in Chapter 2 and for edgeR v3.12.1 in Chapter 3, and genes with Benjamini-Hochberg adjusted p-values < 0.01 were regarded as differentially expressed and used for downstream GO functional analysis with Metascape (*264*) with default parameters. Custom perl and R scripts were used to generate MA (log ratio & mean average) plots in Chapter 2. The heatmap in Figure 2.6A featuring gene expression levels from differentially expressed genes identified by edgeR were normalized, converted into Z-scores, and the results were visualized using the pheatmap R package, where the genes and samples were subjected to unsupervised hierarchical clustering. Log2FC heatmaps of nearest gene expression relevant to the clustered peaks were determined using custom scripts and visualized using Java TreeView.

7 Appendix

7.1 CRISPR sgRNA oligo sequences used for mutant ESC generation

Intended gene mutation	Oligo Name	Sequence
Set1A$^{\Delta SET}$	Left	gccggagccgtatccatgag
	Right	atctcctcatccacgccgat
Set1BKO	Left	ccagatccacgcgaaaagcc
	Right	ctaccggtgccccttccctg
Mll2$^{\Delta SET}$	Left	ggtcagaaagggtcctaaag
	Right	tgccccttgggtggacagat
Set1BKO-Set1A$^{\Delta SET}$ -Mll2$^{\Delta SET}$	Left	ggtcagaaagggtcctaaag
	Right	gtaagtggcgtgaaggtttg
Mll1KO	Left	Previously reported (*98*)
	Right	Previously reported (*98*)

7.2 Primers used for ESC PCR genotyping

Gene	Oligo Name	Primer # (for mousework)	Sequence
Set1A	WT forward	P1	acatgagcctggaaaagtgg
	WT reverse	P2	catgtccgctaccatctgtg
	Set1A$^{\Delta SET}$ forward	P3	taatcgggtgctttctgagc
	Set1A$^{\Delta SET}$ reverse	P4	tacaggctggtaccccaggt
Set1BKO	WT forward	N/A	gaatctggaccaaaaacaag
	WT reverse	N/A	acttgcccagcagttaaaaa
	Set1BKO forward	N/A	tcagcgtctaataactcaagc
	Set1BKO reverse	N/A	cccaagatccactatcaatg
Mll2$^{\Delta SET}$	forward	N/A	cccaagatccactatcaatg
	reverse	N/A	ggaacatgtagcacccaatacc
Set1BKO-Set1A$^{\Delta SET}$ -Mll2$^{\Delta SET}$	forward	N/A	tccaccaggtgtgcagataa
	reverse	N/A	ccatggacaggaaggtagga

7.3 Primers used in qRT-PCR assays

Gene	Oligo Name	Sequence
Hoxa cluster	Forward	Previously reported (*98, 265, 266*)
	Reverse	Previously reported (*98, 265, 266*)
Sox2	Forward	GAACGCCTTCATGGTATGGT
	Reverse	TCTCGGTCTCGGACAAAAGT
Oct4	Forward	AATGCCGTGAAGTTGGAGAA
	Reverse	CCTTCTGCAGGGCTTTCAT
Nanog	Forward	TGCTTACAAGGGTCTGCTACTG
	Reverse	GAGGCAGGTCTTCAGAGGAA
Bmp5	Forward	CCACAGAACAATTTGGGCTTA
	Reverse	AGTACCTCGCTTGCCTTGAA
Epor	Forward	GTCCTCATCTCGCTGTTGCT
	Reverse	ATGCCAGGCCAGATCTTCT

7.4 Primers used for PCR genotyping Set1A$^{\Delta SET}$ mutant mice in established colony

Gene	Oligo Name	Primer #	Sequence
Set1A	WT forward	P5	GGCCGGAGCCGTATCCATGA
	Set1A$^{\Delta SET}$ forward	P6	GGCCGGAGCCGTATCCACA
	Reverse	P7	CCTTGGGGGAGAAGCTCTGT

Relevant first-author and co-author manuscripts during Ph.D. studies

Christie C. Sze, Patrick A. Ozark, Kaixiang Cao, Michal Ugarenko, Siddhartha Das, Lu Wang, Stacy A. Marshall, Emily J. Rendleman, Caila A. Ryan, Didi Zha, Delphine Douillet, Fei Xavier Chen, and Ali Shilatifard. Coordinated regulation of cellular identity-associated H3K4me3 breadth by the COMPASS family. *Sci Adv.* 2020. In press.

Delphine Douillet, **Christie C. Sze**, Caila Ryan, Andrea Piunti, Avani P. Shah, Michal Ugarenko, Stacy A. Marshall, Emily J. Rendleman, Didi Zha, Kathryn A. Helmin, Zibo Zhao, Kaixiang Cao, Marc A. Morgan, Benjamin D. Singer, Elizabeth Bartom, Edwin R. Smith, Ali Shilatifard. An epigenetic equilibrium among Polycomb, COMPASS and DNA methylation machineries regulates developmental gene expression. *Nat Genet.* 2020 May 11. doi: 10.1038/s41588-020-0618-1.

Damiano Fantini, Alexander P. Glaser, Kalen J. Rimar, Yiduo Wang, Matthew Schipma, Nobish Varghese, Alfred Rademaker, Amir Behdad, Aparna Yellapa, Yanni Yu, **Christie Ching-Lin Sze**, Lu Wang, Zibo Zhao, Susan E. Crawford, Deqing Hu, Jonathan D. Licht, Clayton K. Collings, Elizabeth Bartom, Dan Theodorescu, Ali Shilatifard & Joshua J. Meeks. A Carcinogen-induced mouse model recapitulates the molecular alterations of human muscle invasive bladder cancer. *Oncogene*. 2018 Apr; 37(14):1911-1925.

Lu Wang, Clayton K. Collings, Zibo Zhao, Kira A. Cozzolino, Quanhong Ma, Kaiwai Liang, Stacy A. Marshall, **Christie C. Sze**, Rintaro Hashizume, Jeffrey Nicholas Savas, Ali Shilatifard. A cytoplasmic COMPASS is necessary for cell survival and triple-negative breast cancer pathogenesis by regulating metabolism. *Genes Dev*. 2017, 31(20):2056-2066.

Marc A.J. Morgan, Ryan A. Rickels, Clayton K. Collings, Xiaolin He, Kaixiang Cao, Hans-Martin Herz, Kira A. Cozzolino, Nebiyu A. Abshiru, Stacy A. Marshall, Emily J. Rendleman, **Christie C. Sze**, Andrea Piunti, Neil L. Kelleher, Jeffrey N. Savas, and Ali Shilatifard. A cryptic Tudor domain links BRWD2/PHIP to COMPASS-mediated histone H3K4 methylation. *Genes Dev.* 2017, 31(19):2003-2014.

Ryan Rickels*, Hans-Martin Herz*, **Christie C. Sze**, Kaixiang Cao, Marc A. Morgan, Clayton K. Collings, Maria Gause, Yohhei Takahashi, Lu Wang, Emily J. Rendleman, Stacy A. Marshall, Annika Krueger, Elizabeth T. Bartom, Andrea Piunti, Edwin R. Smith, Nebiyu A. Abshiru, Neil L. Kelleher, Dale Dorsett, and Ali Shilatifard. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet*. 2017, 49(11):1647-1653.

Christie C. Sze, Kaixiang Cao, Clayton K. Collings, Stacy A. Marshall, Emily J. Rendleman, Patrick A. Ozark, Fei Xavier Chen, Marc A. Morgan, Lu Wang, and Ali Shilatifard. Histone H3K4 methylation-dependent and -independent functions of Set1A/COMPASS in embryonic stem cell self-renewal and differentiation. *Genes Dev*. 2017, 31(17):1732-1737

Kaixiang Cao, Clayton K. Collings, Stacy A. Marshall, Marc A. Morgan, Emily J. Rendleman, Lu Wang, **Christie C. Sze**, Tianjiao Sun, Elizabeth T. Bartom, and Ali Shilatifard. SET1A/COMPASS and shadow enhancers in the regulation of homeotic gene expression. *Genes Dev*. 2017, 31(8):787-801.

Christie C. Sze and Ali Shilatifard. MLL3/MLL4/COMPASS Family on Epigenetic Regulation of Enhancer Function and Cancer. *Cold Spring Harb Perspect Med*. 2016, doi: 10.1101/cshperspect.a026427.

**Co-first authors*

References

1. R. D. Kornberg, Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-871 (1974).

2. S. R. Bhaumik, E. Smith, A. Shilatifard, Covalent modifications of histones during development and disease pathogenesis. *Nat Struct Mol Biol* **14**, 1008-1016 (2007).

3. T. Kouzarides, Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).

4. J. C. Black, C. Van Rechem, J. R. Whetstine, Histone lysine methylation dynamics: establishment, regulation, and biological impact. *Mol Cell* **48**, 491-507 (2012).

5. P. W. Ingham, Whittle, R., Trithorax: a new homeotic mutation of Drosophila melanogaster causing transformations of abdominal and thoracic imaginal segments. *Mol. Gen. Genet.* **179**, 607-614 (1980).

6. A. Barski *et al.*, High-resolution profiling of histone methylations in the human genome. **129**, 823-837 (2007).

7. B. E. Bernstein *et al.*, Methylation of histone H3 Lys 4 in coding regions of active genes. **99**, 8695-8700 (2002).

8. H. Santos-Rosa *et al.*, Active genes are tri-methylated at K4 of histone H3. **419**, 407-411 (2002).

9. E. Smith, A. Shilatifard, The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes. **40**, 689-701 (2010).

10. N. D. Heintzman *et al.*, Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311-318 (2007).

11. N. D. Heintzman *et al.*, Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).

12. E. Smith, A. Shilatifard, Enhancer biology and enhanceropathies. *Nature structural & molecular biology* **21**, 210-219 (2014).

13. A. J. Bannister *et al.*, Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. **280**, 17732-17736 (2005).

14. J. Ernst *et al.*, Mapping and analysis of chromatin state dynamics in nine human cell types. **473**, 43-49 (2011).

15. A. H. Peters *et al.*, Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. **12**, 1577-1589 (2003).

16. B. E. Bernstein *et al.*, A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).

17. D. E. Sterner, S. L. Berger, Acetylation of histones and transcription-related factors. **64**, 435-459 (2000).

18. D. J. Steger, J. L. Workman, Remodeling chromatin structures for transcription: what happens to the histones? **18**, 875-884 (1996).

19. C. Tse, T. Sera, A. P. Wolffe, J. C. Hansen, Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. **18**, 4629-4638 (1998).

20. M. P. Creyghton *et al.*, Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936 (2010).

21. A. Rada-Iglesias *et al.*, A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).

22. G. E. Zentner, P. J. Tesar, P. C. Scacheri, Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research* **21**, 1273-1283 (2011).

23. J. Dover *et al.*, Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. **277**, 28368-28371 (2002).

24. A. Shilatifard, Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. **75**, 243-269 (2006).

25. Z. W. Sun, C. D. Allis, Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. **418**, 104-108 (2002).

26. K. Cao *et al.*, An Mll4/COMPASS-Lsd1 epigenetic axis governs enhancer function and pluripotency transition in embryonic stem cells. *Sci Adv* **4**, eaap8747 (2018).

27. A. Jambhekar, A. Dhall, Y. Shi, Roles and regulation of histone methylation in animal development. **20**, 625-641 (2019).

28. J. S. Lee, E. Smith, A. Shilatifard, The language of histone crosstalk. **142**, 682-685 (2010).

29. M. J. Evans, M. H. Kaufman, Establishment in culture of pluripotential cells from mouse embryos. **292**, 154-156 (1981).

30. A. Bradley, M. Evans, M. H. Kaufman, E. Robertson, Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. **309**, 255-256 (1984).

31. A. De Los Angeles *et al.*, Hallmarks of pluripotency. **525**, 469-478 (2015).

32. S. L. Mansour, K. R. Thomas, M. R. Capecchi, Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. **336**, 348-352 (1988).

33. J. Wu, J. C. Izpisua Belmonte, Dynamic Pluripotent Stem Cell States and Their Applications. **17**, 509-525 (2015).

34. K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. **126**, 663-676 (2006).

35. Y. Toyooka, D. Shimosato, K. Murakami, K. Takahashi, H. Niwa, Identification and characterization of subpopulations in undifferentiated ES cell culture. **135**, 909-918 (2008).

36. Q. L. Ying *et al.*, The ground state of embryonic stem cell self-renewal. **453**, 519-523 (2008).

37. A. Gaspar-Maia, A. Alajem, E. Meshorer, M. Ramalho-Santos, Open chromatin in pluripotency and reprogramming. **12**, 36-47 (2011).

38. S. H. Orkin, K. Hochedlinger, Chromatin connections to pluripotency and cellular reprogramming. **145**, 835-850 (2011).

39. S. Efroni *et al.*, Global transcription in pluripotent embryonic stem cells. **2**, 437-447 (2008).

40. A. Gaspar-Maia *et al.*, Chd1 regulates open chromatin and pluripotency of embryonic stem cells. **460**, 863-868 (2009).

41. K. Ahmed *et al.*, Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. **5**, e10531 (2010).

42. V. Azuara *et al.*, Chromatin signatures of pluripotent cell lines. **8**, 532-538 (2006).

43. B. Wen, H. Wu, Y. Shinkai, R. A. Irizarry, A. P. Feinberg, Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. **41**, 246-250 (2009).

44. R. D. Hawkins *et al.*, Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. **6**, 479-491 (2010).

45. E. Meshorer *et al.*, Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. **10**, 105-116 (2006).

46. G. Pan *et al.*, Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. **1**, 299-312 (2007).

47. C. L. Fisher, A. G. Fisher, Chromatin states in pluripotent, differentiated, and reprogrammed cells. **21**, 140-146 (2011).

48. S. Ziemin-van der Poel *et al.*, Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias. *Proc Natl Acad Sci U S A* **88**, 10735-10739 (1991).

49. M. Djabali *et al.*, A trithorax-like gene is interrupted by chromosome 11q23 translocations in acute leukaemias. *Nat Genet* **2**, 113-118 (1992).

50. Y. Gu *et al.*, The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to Drosophila trithorax, to the AF-4 gene. *Cell* **71**, 701-708 (1992).

51. D. C. Tkachuk, S. Kohler, M. L. Cleary, Involvement of a homolog of Drosophila trithorax by 11q23 chromosomal translocations in acute leukemias. *Cell* **71**, 691-700 (1992).

52. T. Miller *et al.*, COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc Natl Acad Sci U S A* **98**, 12902-12907 (2001).

53. A. Roguev *et al.*, The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *EMBO J* **20**, 7137-7148 (2001).

54. J. Schneider *et al.*, Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. *Mol Cell* **19**, 849-856 (2005).

55. J. C. Eissenberg, A. Shilatifard, Histone H3 lysine 4 (H3K4) methylation in development and differentiation. *Dev Biol* **339**, 240-249 (2010).

56. M. Mohan *et al.*, The COMPASS family of H3K4 methylases in Drosophila. *Mol Cell Biol* **31**, 4310-4318 (2011).

57. C. D. Allis *et al.*, New nomenclature for chromatin-modifying enzymes. *Cell* **131**, 633-636 (2007).

58. A. Shilatifard, The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu Rev Biochem* **81**, 65-95 (2012).

59. C. C. Sze, A. Shilatifard, MLL3/MLL4/COMPASS Family on Epigenetic Regulation of Enhancer Function and Cancer. *Cold Spring Harb Perspect Med* **6**, (2016).

60. R. van Nuland *et al.*, Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Molecular and cellular biology* **33**, 2067-2077 (2013).

61. M. J. Stassen, D. Bailey, S. Nelson, V. Chinwalla, P. J. Harte, 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).

62. B. Tschiersch *et al.*, The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J* **13**, 3822-3831 (1994).

63. H. M. Herz, A. Garruss, A. Shilatifard, SET for life: biochemical activities and biological functions of SET domain-containing proteins. *Trends Biochem Sci* **38**, 621-639 (2013).

64. M. B. Ardehali *et al.*, Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. *EMBO J* **30**, 2817-2828 (2011).

65. J. Schultz, F. Milpetz, P. Bork, C. P. Ponting, SMART, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 5857-5864 (1998).

66. I. Letunic, T. Doerks, P. Bork, SMART: recent updates, new developments and status in 2015. *Nucleic acids research* **43**, D257-260 (2015).

67. M. Wu *et al.*, Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. *Mol Cell Biol* **28**, 7337-7344 (2008).

68. G. Hallson *et al.*, dSet1 is the main H3K4 di- and tri-methyltransferase throughout Drosophila development. *Genetics* **190**, 91-100 (2012).

69. P. Wang *et al.*, Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Mol Cell Biol* **29**, 6074-6085 (2009).

70. D. Hu *et al.*, The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nat Struct Mol Biol* **20**, 1093-1097 (2013).

71. Y. Sedkov *et al.*, Molecular genetic analysis of the Drosophila trithorax-related gene which encodes a novel SET domain protein. *Mech Dev* **82**, 171-179 (1999).

72. H. M. Herz *et al.*, Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes & development* **26**, 2604-2620 (2012).

73. D. Hu *et al.*, The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. *Molecular and cellular biology* **33**, 4745-4754 (2013).

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

75. D. Hu *et al.*, Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification. *Mol Cell* **65**, 460-475 e466 (2017).

76. T. Cierpicki *et al.*, Structure of the MLL CXXC domain-DNA complex and its functional role in MLL-AF9 leukemia. *Nat Struct Mol Biol* **17**, 62-68 (2010).

77. C. Xu, C. Bian, R. Lam, A. Dong, J. Min, The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nat Commun* **2**, 227 (2011).

78. C. Xu *et al.*, DNA Sequence Recognition of Human CXXC Domains and Their Structural Determinants. *Structure* **26**, 85-95 e83 (2018).

79. S. D. Taverna, H. Li, A. J. Ruthenburg, C. D. Allis, D. J. Patel, How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. **14**, 1025-1040 (2007).

80. T. A. Milne *et al.*, Multiple interactions recruit MLL1 and MLL1 fusion proteins to the HOXA9 locus in leukemogenesis. **38**, 853-863 (2010).

81. S. S. Dhar *et al.*, Trans-tail regulation of MLL4-catalyzed H3K4 methylation by H4R3 symmetric dimethylation is mediated by a tandem PHD of MLL4. **26**, 2749-2762 (2012).

82. J. S. Lee *et al.*, Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell* **131**, 1084-1096 (2007).

83. N. J. Krogan *et al.*, The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Mol Cell* **11**, 721-729 (2003).

84. J. H. Lee, D. G. Skalnik, Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes. *Mol Cell Biol* **28**, 609-618 (2008).

85. H. H. Ng, F. Robert, R. A. Young, K. Struhl, Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* **11**, 709-719 (2003).

86. L. M. Soares *et al.*, Determinants of Histone H3K4 Methylation Patterns. *Mol Cell* **68**, 773-785 e776 (2017).

87. H. J. Bae *et al.*, The Set1 N-terminal domain and Swd2 interact with RNA polymerase II CTD to recruit COMPASS. **11**, 2181 (2020).

88. Y. S. Ang *et al.*, Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145**, 183-197 (2011).

89. L. Fang *et al.*, H3K4 Methyltransferase Set1a Is A Key Oct4 Coactivactor Essential for Generation of Oct4 Positive Inner Cell Mass. *Stem Cells*, (2016).

90. S. El Ashkar *et al.*, LEDGF/p75 is dispensable for hematopoiesis but essential for MLL-rearranged leukemogenesis. *Blood* **131**, 95-107 (2018).

91. C. M. Hughes *et al.*, Menin associates with a trithorax family histone methyltransferase complex and with the hoxc8 locus. *Mol Cell* **13**, 587-597 (2004).

92. T. A. Milne *et al.*, Menin and MLL cooperatively regulate expression of cyclin-dependent kinase inhibitors. *Proc Natl Acad Sci U S A* **102**, 749-754 (2005).

93. A. Yokoyama, M. L. Cleary, Menin critically links MLL proteins with LEDGF on cancer-associated target genes. *Cancer Cell* **14**, 36-46 (2008).

94. A. S. Bledau *et al.*, The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation. *Development* **141**, 1022-1035 (2014).

95. D. Brici *et al.*, Setd1b, encoding a histone 3 lysine 4 methyltransferase, is a maternal effect gene required for the oogenic gene expression program. **144**, 2606-2617 (2017).

96. C. Deng *et al.*, USF1 and hSET1A mediated epigenetic modifications regulate lineage differentiation and HoxB4 transcription. *PLoS Genet* **9**, e1003524 (2013).

97. B. K. Tusi *et al.*, Setd1a regulates progenitor B-cell-to-precursor B-cell development through histone H3 lysine 4 trimethylation and Ig heavy-chain rearrangement. *FASEB J* **29**, 1505-1515 (2015).

98. K. Cao *et al.*, SET1A/COMPASS and shadow enhancers in the regulation of homeotic gene expression. *Genes Dev* **31**, 787-801 (2017).

99. C. C. Sze *et al.*, Histone H3K4 methylation-dependent and -independent functions of Set1A/COMPASS in embryonic stem cell self-renewal and differentiation. *Genes Dev* **31**, 1732-1737 (2017).

100. T. Hoshii *et al.*, A Non-catalytic Function of SETD1A Regulates Cyclin K and the DNA Damage Response. *Cell* **172**, 1007-1021 e1017 (2018).

101. L. Wang *et al.*, A cytoplasmic COMPASS is necessary for cell survival and triple-negative breast cancer pathogenesis by regulating metabolism. *Genes Dev* **31**, 2056-2066 (2017).

102. R. Rickels *et al.*, An Evolutionary Conserved Epigenetic Mark of Polycomb Response Elements Implemented by Trx/MLL/COMPASS. *Mol Cell* **63**, 318-328 (2016).

103. B. D. Yu, J. L. Hess, S. E. Horning, G. A. Brown, S. J. Korsmeyer, Altered Hox expression and segmental identity in Mll-mutant mice. **378**, 505-508 (1995).

104. R. Terranova, H. Agherbi, A. Boned, S. Meresse, M. Djabali, Histone and DNA methylation defects at Hox genes in mice expressing a SET domain-truncated form of Mll. **103**, 6629-6634 (2006).

105. S. Denissov *et al.*, Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. *Development* **141**, 526-537 (2014).

106. G. Barbagiovanni *et al.*, KMT2B Is Selectively Required for Neuronal Transdifferentiation, and Its Loss Exposes Dystonia Candidate Genes. **25**, 988-1001 (2018).

107. S. Glaser *et al.*, Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. **133**, 1423-1432 (2006).

108. C. V. Andreu-Vieyra *et al.*, MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. **8**, (2010).

109. S. Glaser *et al.*, The histone 3 lysine 4 methyltransferase, Mll2, is only required briefly in development and spermatogenesis. **2**, 5 (2009).

110. H. M. Herz, D. Hu, A. Shilatifard, Enhancer malfunction in cancer. *Mol Cell* **53**, 859-866 (2014).

111. S. Y. Ang *et al.*, KMT2D regulates specific programs in heart development via histone H3 lysine 4 di-methylation. *Development* **143**, 810-821 (2016).

112. M. U. Kaikkonen *et al.*, Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular cell* **51**, 310-325 (2013).

113. A. Ortega-Molina *et al.*, The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nature medicine* **21**, 1199-1208 (2015).

114. Y. Jang, C. Wang, L. Zhuang, C. Liu, K. Ge, H3K4 Methyltransferase Activity Is Required for MLL4 Protein Stability. **429**, 2046-2054 (2017).

115. R. Rickels *et al.*, Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet* **49**, 1647-1653 (2017).

116. C. Kandoth *et al.*, Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013).

117. M. S. Lawrence *et al.*, Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).

118. T. Singh *et al.*, Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci* **19**, 571-577 (2016).

119. T. Hiraide *et al.*, De novo variants in SETD1B are associated with intellectual disability, epilepsy and autism. **137**, 95-104 (2018).

120. T. Salz *et al.*, Histone Methyltransferase hSETD1A Is a Novel Regulator of Metastasis in Breast Cancer. *Mol Cancer Res* **13**, 461-469 (2015).

121. T. Li *et al.*, SET1A Cooperates With CUDR to Promote Liver Cancer Growth and Hepatocyte-like Stem Cell Malignant Transformation Epigenetically. *Mol Ther*, (2015).

122. K. Tajima *et al.*, SETD1A modulates cell cycle progression through a miRNA network that regulates p53 target genes. *Nat Commun* **6**, 8257 (2015).

123. K. Arndt *et al.*, SETD1A protects HSCs from activation-induced functional decline in vivo. *Blood*, (2018).

124. M. R. Higgs *et al.*, Histone Methylation by SETD1A Protects Nascent DNA through the Nucleosome Chaperone Activity of FANCD2. **71**, 25-41.e26 (2018).

125. K. Tajima *et al.*, SETD1A protects from senescence through regulation of the mitotic gene expression program. **10**, 2854 (2019).

126. A. Takata, I. Ionita-Laza, J. A. Gogos, B. Xu, M. Karayiorgou, De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. *Neuron* **89**, 940-947 (2016).

127. A. Takata *et al.*, Loss-of-function variants in schizophrenia risk and SETD1A as a candidate susceptibility gene. *Neuron* **82**, 773-780 (2014).

128. Y. J. Choi *et al.*, Frameshift mutation of a histone methylation-related gene SETD1B and its regional heterogeneity in gastric and colorectal cancers with high microsatellite instability. *Hum Pathol* **45**, 1674-1681 (2014).

129. J. H. Lee, D. G. Skalnik, Rbm15-Mkl1 interacts with the Setd1b histone H3-Lys4 methyltransferase via a SPOC domain that is required for cytokine-independent proliferation. *PLoS One* **7**, e42965 (2012).

130. Y. Song *et al.*, Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91-95 (2014).

131. K. Den *et al.*, A novel de novo frameshift variant in SETD1B causes epilepsy. **64**, 821-827 (2019).

132. W. D. Jones *et al.*, De novo mutations in MLL cause Wiedemann-Steiner syndrome. **91**, 358-364 (2012).

133. E. Meyer *et al.*, Mutations in the histone methyltransferase gene KMT2B cause complex early-onset dystonia. *Nat Genet* **49**, 223-237 (2017).

134. M. Zech *et al.*, Haploinsufficiency of KMT2B, Encoding the Lysine-Specific Histone Methyltransferase 2B, Results in Early-Onset Generalized Dystonia. **99**, 1377-1387 (2016).

135. D. W. Parsons *et al.*, The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435-439 (2011).

136. H. Ashktorab *et al.*, Distinct genetic alterations in colorectal cancer. *PloS one* **5**, e8879 (2010).

137. B. Akhtar-Zaidi *et al.*, Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**, 736-739 (2012).

138. D. T. Jones *et al.*, Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100-105 (2012).

139. T. J. Pugh *et al.*, Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**, 106-110 (2012).

140. R. D. Morin *et al.*, Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298-303 (2011).

141. L. Pasqualucci *et al.*, Analysis of the coding genome of diffuse large B-cell lymphoma. *Nature genetics* **43**, 830-837 (2011).

142. M. J. Ellis *et al.*, Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353-360 (2012).

143. C. S. Grasso *et al.*, The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239-243 (2012).

144. A. C. da Silva Almeida *et al.*, The mutational landscape of cutaneous T cell lymphoma and Sezary syndrome. *Nature genetics* **47**, 1465-1470 (2015).

145. J. Tan *et al.*, Genomic landscapes of breast fibroepithelial tumors. *Nature genetics* **47**, 1341-1345 (2015).

146. Y. Gui *et al.*, Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nature genetics* **43**, 875-878 (2011).

147. M. Ruault, M. E. Brun, M. Ventura, G. Roizes, A. De Sario, MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently deleted in myeloid leukaemia. *Gene* **284**, 73-81 (2002).

148. Y. B. Gao *et al.*, Genetic landscape of esophageal squamous cell carcinoma. *Nature genetics* **46**, 1097-1102 (2014).

149. D. C. Lin *et al.*, Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nature genetics* **46**, 467-473 (2014).

150. J. Lee *et al.*, A tumor suppressive coactivator complex of p53 containing ASC-2 and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8513-8518 (2009).

151. C. Chen *et al.*, MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. *Cancer cell* **25**, 652-665 (2014).

152. M. Garg *et al.*, Profiling of somatic mutations of acute myeloid leukemia, FLT3-ITD subgroup at diagnosis and relapse. *Blood*, (2015).

153. J. Zhang *et al.*, Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nature medicine* **21**, 1190-1198 (2015).

154. L. Wang *et al.*, Resetting the epigenetic balance of Polycomb and COMPASS function at enhancers for cancer therapy. **24**, 758-769 (2018).

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

156. J. H. Kim *et al.*, UTX and MLL4 coordinately regulate transcriptional programs for cell proliferation and invasiveness in breast cancer cells. *Cancer research* **74**, 1705-1717 (2014).

157. S. B. Ng *et al.*, Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. **42**, 790-793 (2010).

158. D. Cocciadiferro *et al.*, Dissecting KMT2D missense mutations in Kabuki syndrome patients. **27**, 3651-3668 (2018).

159. T. S. Koemans *et al.*, Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder. **13**, e1006864 (2017).

160. K. M. Dorighi *et al.*, Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**, 568-576 e564 (2017).

161. C. Y. McLean *et al.*, GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

162. Q. L. Ying, M. Stavridis, D. Griffiths, M. Li, A. Smith, Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat Biotechnol* **21**, 183-186 (2003).

163. C. Galonska, M. J. Ziller, R. Karnik, A. Meissner, Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor Binding before Global Epigenetic Reprogramming. *Cell Stem Cell* **17**, 462-470 (2015).

164. A. Barski *et al.*, High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).

165. M. R. Higgs *et al.*, BOD1L Is Required to Suppress Deleterious Resection of Stressed Replication Forks. *Mol Cell* **59**, 462-477 (2015).

166. P. Ernst *et al.*, Definitive hematopoiesis requires the mixed-lineage leukemia gene. **6**, 437-443 (2004).

167. B. D. Yu, R. D. Hanson, J. L. Hess, S. E. Horning, S. J. Korsmeyer, MLL, a mammalian trithorax-group gene, functions as a transcriptional maintenance factor in morphogenesis. **95**, 10632-10636 (1998).

168. A. J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-395 (2011).

169. E. I. Campos, D. Reinberg, Histones: annotating chromatin. *Annu Rev Genet* **43**, 559-599 (2009).

170. R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**, 245-254 (2003).

171. J. E. Audia, R. M. Campbell, Histone Modifications and Cancer. *Cold Spring Harb Perspect Biol* **8**, a019521 (2016).

172. A. Piunti, A. Shilatifard, Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science* **352**, aad9780 (2016).

173. L. Wang, A. Shilatifard, UTX Mutations in Human Cancer. *Cancer Cell* **35**, 168-176 (2019).

174. B. E. Bernstein *et al.*, Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169-181 (2005).

175. H. Santos-Rosa *et al.*, Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407-411 (2002).

176. J. F. Flanagan *et al.*, Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature* **438**, 1181-1185 (2005).

177. R. J. Sims, 3rd *et al.*, Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *J Biol Chem* **280**, 41789-41792 (2005).

178. S. M. Lauberth *et al.*, H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell* **152**, 1021-1036 (2013).

179. M. Vermeulen *et al.*, Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58-69 (2007).

180. D. Faucher, R. J. Wellinger, Methylated H3K4, a transcription-associated histone modification, is involved in the DNA damage response pathway. *PLoS Genet* **6**, (2010).

181. P. V. Pena *et al.*, Histone H3K4me3 binding is required for the DNA repair and apoptotic activities of ING1 tumor suppressor. *J Mol Biol* **380**, 303-312 (2008).

182. B. A. Benayoun *et al.*, H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673-688 (2014).

183. K. Chen *et al.*, Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* **47**, 1149-1157 (2015).

184. N. J. Krogan *et al.*, COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression. *J Biol Chem* **277**, 10753-10755 (2002).

185. J. H. Lee, C. M. Tate, J. S. You, D. G. Skalnik, Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. *J Biol Chem* **282**, 13419-13428 (2007).

186. Z. Tang *et al.*, SET1 and p300 act synergistically, through coupled histone modifications, in transcriptional activation by p53. *Cell* **154**, 297-310 (2013).

187. P. Ernst, M. Mabon, A. J. Davidson, L. I. Zon, S. J. Korsmeyer, An Mll-dependent Hox program drives hematopoietic progenitor expansion. *Curr Biol* **14**, 2063-2069 (2004).

188. R. Rickels *et al.*, An Evolutionary Conserved Epigenetic Mark of Polycomb Response Elements Implemented by Trx/MLL/COMPASS. *Mol Cell* **63**, 318-328 (2016).

189. S. Glaser *et al.*, Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. *Development* **133**, 1423-1432 (2006).

190. H. Yagi *et al.*, Growth disturbance in fetal liver hematopoiesis of Mll-mutant mice. *Blood* **92**, 108-117 (1998).

191. B. D. Yu, J. L. Hess, S. E. Horning, G. A. Brown, S. J. Korsmeyer, Altered Hox expression and segmental identity in Mll-mutant mice. *Nature* **378**, 505-508 (1995).

192. D. A. Brown *et al.*, The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. *Cell Rep* **20**, 2313-2327 (2017).

193. T. Clouaire *et al.*, Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev* **26**, 1714-1728 (2012).

194. X. Liu *et al.*, Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature* **537**, 558-562 (2016).

195. J. A. Dahl *et al.*, Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* **537**, 548-552 (2016).

196. B. Zhang *et al.*, Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**, 553-557 (2016).

197. G. Mas *et al.*, Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat Genet* **50**, 1452-1462 (2018).

198. C. V. Andreu-Vieyra *et al.*, MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. *PLoS Biol* **8**, (2010).

199. C. W. Hanna *et al.*, MLL2 conveys transcription-independent H3K4 trimethylation in oocytes. *Nat Struct Mol Biol* **25**, 73-82 (2018).

200. M. J. Boland, K. L. Nazor, J. F. Loring, Epigenetic regulation of pluripotency and differentiation. **115**, 311-324 (2014).

201. M. E. Lalonde, X. Cheng, J. C*ot*e, Histone target selection within chromatin: an exemplary case of teamwork. **28**, 1029-1041 (2014).

202. M. E. Lalonde *et al.*, Exchange of associated factors directs a switch in HBO1 acetyltransferase histone tail specificity. **27**, 2009-2024 (2013).

203. S. K. Hota, B. G. Bruneau, ATP-dependent chromatin remodeling during mammalian development. **143**, 2882-2897 (2016).

204. L. Wang *et al.*, INO80 facilitates pluripotency gene activation in embryonic stem cell self-renewal, reprogramming, and blastocyst development. **14**, 575-591 (2014).

205. A. Dantas *et al.*, Biological Functions of the ING Proteins. **11**, (2019).

206. M. Shiseki *et al.*, p29ING4 and p28ING5 bind to p53 and p300, and enhance p53 activity. **63**, 2373-2378 (2003).

207. K. S. Champagne *et al.*, The crystal structure of the ING5 PHD finger in complex with an H3K4me3 histone peptide. **72**, 1371-1376 (2008).

208. F. Wang *et al.*, ING5 activity in self-renewal of glioblastoma stem cells via calcium and follicle stimulating hormone pathways. **37**, 286-301 (2018).

209. K. W. Mulder *et al.*, Diverse epigenetic strategies interact to control epidermal differentiation. **14**, 753-763 (2012).

210. M. S. Kim *et al.*, The histone acetyltransferase Myst2 regulates Nanog expression, and is involved in maintaining pluripotency and self-renewal of embryonic stem cells. **589**, 941-950 (2015).

211. L. MacPherson *et al.*, HBO1 is required for the maintenance of leukaemia stem cells. **577**, 266-270 (2020).

212. F. M. Perez-Campo, J. Borrow, V. Kouskoff, G. Lacaud, The histone acetyl transferase activity of monocytic leukemia zinc finger is critical for the proliferation of hematopoietic precursors. **113**, 4866-4874 (2009).

213. T. D. Merson *et al.*, The transcriptional coactivator Querkopf controls adult neurogenesis. **26**, 11359-11370 (2006).

214. N. Avvakumov, J. C*ot*e, The MYST family of histone acetyltransferases and their intimate links to cancer. **26**, 5395-5407 (2007).

215. B. N. Sheikh, A. Akhtar, The many lives of KATs - detectors, integrators and modulators of the cellular environment. **20**, 7-23 (2019).

216. B. J. Klein *et al.*, Histone H3K23-specific acetylation by MORF is coupled to H3K14 acylation. **10**, 4724 (2019).

217. A. K. Voss, C. Collin, M. P. Dixon, T. Thomas, Moz and retinoic acid coordinately regulate H3K9 acetylation, Hox gene expression, and segment identity. **17**, 674-686 (2009).

218. A. K. Voss *et al.*, MOZ regulates the Tbx1 locus, and Moz mutation partially phenocopies DiGeorge syndrome. **23**, 652-663 (2012).

219. Y. Doyon *et al.*, ING tumor suppressor proteins are critical regulators of chromatin acetylation required for genome expression and perpetuation. **21**, 51-64 (2006).

220. D. Lv *et al.*, Histone Acetyltransferase KAT6A Upregulates PI3K/AKT Signaling through TRIM24 Binding. **77**, 6190-6201 (2017).

221. L. Sim*o-Riudalbas *et al.*, KAT6B Is a Tumor Suppressor Histone H3 Lysine 23 Acetyltransferase Undergoing Genomic Loss in Small Cell Lung Cancer. **75**, 3936-3945 (2015).

222. Y. Qiu *et al.*, Combinatorial readout of unmodified H3R2 and acetylated H3K14 by the tandem PHD finger of MOZ reveals a regulatory mechanism for HOXA9 transcription. **26**, 1376-1391 (2012).

223. Y. Feng *et al.*, BRPF3-HBO1 regulates replication origin activation and histone H3K14 acetylation. **35**, 176-192 (2016).

224. A. J. Kueh, M. P. Dixon, A. K. Voss, T. Thomas, HBO1 is required for H3K14 acetylation and normal transcriptional activity during embryonic development. **31**, 845-860 (2011).

225. Y. Mishima *et al.*, The Hbo1-Brd1/Brpf2 complex is responsible for global acetylation of H3K14 and required for fetal liver erythropoiesis. **118**, 2443-2453 (2011).

226. D. M. Newman, A. K. Voss, T. Thomas, R. S. Allan, Essential role for the histone acetyltransferase KAT7 in T cell development, fitness, and survival. **101**, 887-892 (2017).

227. R. L. Foy *et al.*, Role of Jade-1 in the histone acetyltransferase (HAT) HBO1 complex. **283**, 28817-28826 (2008).

228. T. Thomas, A. K. Voss, K. Chowdhury, P. Gruss, Querkopf, a MYST family histone acetyltransferase, is required for normal cerebral cortex development. **127**, 2537-2548 (2000).

229. T. Katsumoto *et al.*, MOZ is essential for maintenance of hematopoietic stem cells. **20**, 1321-1330 (2006).

230. J. Su *et al.*, Genomic Integrity Safeguards Self-Renewal in Embryonic Stem Cells. **28**, 1400-1409.e1404 (2019).

231. I. Vitale, G. Manic, R. De Maria, G. Kroemer, L. Galluzzi, DNA Damage in Stem Cells. **66**, 306-319 (2017).

232. T. W. Burke, J. G. Cook, M. Asano, J. R. Nevins, Replication factors MCM2 and ORC1 interact with the histone acetyltransferase HBO1. **276**, 15397-15408 (2001).

233. M. Iizuka, B. Stillman, Histone acetyltransferase HBO1 interacts with the ORC1 subunit of the human initiator protein. **274**, 23027-23034 (1999).

234. N. Avvakumov *et al.*, Conserved molecular interactions within the HBO1 acetyltransferase complexes regulate cell proliferation. **32**, 689-703 (2012).

235. N. Liu *et al.*, ING5 is a Tip60 cofactor that acetylates p53 in response to DNA damage. **73**, 3749-3760 (2013).

236. U. Linzen *et al.*, ING5 is phosphorylated by CDK2 and controls cell proliferation independently of p53. **10**, e0123736 (2015).

237. J. Paggetti *et al.*, Crosstalk between leukemia-associated proteins MOZ and MLL regulates HOX gene expression in human cord blood CD34+ cells. **29**, 5019-5031 (2010).

238. N. Saksouk *et al.*, HBO1 HAT complexes target chromatin throughout gene coding regions via multiple PHD finger interactions with histone H3 tail. **33**, 257-265 (2009).

239. C. C. Sze *et al.*, Coordinated regulation of cellular identity-associated H3K4me3 breadth by the COMPASS family. *Sci Adv*, Manuscript submitted (2020).

240. D. Douillet *et al.*, Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation. (2020).

241. D. L. Carlone *et al.*, Reduced genomic cytosine methylation and defective cellular differentiation in embryonic stem cells lacking CpG binding protein. **25**, 4881-4891 (2005).

242. Y. Bi *et al.*, WDR82, a key epigenetics-related factor, plays a crucial role in normal early embryonic development in mice. **84**, 756-764 (2011).

243. K. Nishimura, T. Fukagawa, H. Takisawa, T. Kakimoto, M. Kanemaki, An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nat Methods* **6**, 917-922 (2009).

244. E. P. Nora *et al.*, Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).

245. F. A. Ran *et al.*, Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-2308 (2013).

246. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for CRISPR screening. **11**, 783-784 (2014).

247. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

248. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. **25**, 1754-1760 (2009).

249. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. **25**, 2078-2079 (2009).

250. G. E. Truett *et al.*, Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). **29**, 52, 54 (2000).

251. J. G. Doench *et al.*, Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. **34**, 184-191 (2016).

252. J. Joung *et al.*, Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. **12**, 828-863 (2017).

253. U. Parekh *et al.*, Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. **7**, 548-555.e548 (2018).

254. T. I. Lee, S. E. Johnstone, R. A. Young, Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc* **1**, 729-748 (2006).

255. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

256. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

257. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

258. L. Shen, N. Shao, X. Liu, E. Nestler, ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. **15**, 284 (2014).

259. F. Ramirez *et al.*, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).

260. S. Heinz *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).

261. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

262. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

263. C. Zang *et al.*, A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952-1958 (2009).

264. S. Tripathi *et al.*, Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell Host Microbe* **18**, 723-735 (2015).

265. C. Lin *et al.*, Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC). *Genes Dev* **25**, 1486-1498 (2011).

266. Y. Zhang, Z. Liu, M. Medrzycki, K. Cao, Y. Fan, Reduction of Hox gene expression by histone H1 depletion. *PLoS One* **7**, e38829 (2012).