


RESEARCH PAPER

 OPEN ACCESS 

Loss of KMT2C reprograms the epigenomic landscape in hPSCs resulting in NODAL overexpression and a failure of hemogenic endothelium specification

Shailendra Maurya^a, Wei Yang^b, Minori Tamai^a, Qiang Zhang^c, Petra Erdmann-Gilmore^c, Amelia Bystry^a, Fernanda Martins Rodrigues^c, Mark C. Valentine^a, Wing H Wong^a, Reid Townsend^c, and Todd E. Druley ^a

^aDepartment of Pediatrics, Division of Pediatric Hematology and Oncology, Washington University in St Louis School of Medicine, St. Louis, Missouri, United States; ^bMcDonnell Genome Institute, Genome Technology Access Center, Washington University in St Louis School of Medicine, St. Louis, Missouri, United States; ^cDepartment of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri, USA

ABSTRACT

Germline or somatic variation in the family of KMT2 lysine methyltransferases have been associated with a variety of congenital disorders and cancers. Notably, *KMT2A*-fusions are prevalent in 70% of infant leukaemias but fail to phenocopy short latency leukaemogenesis in mammalian models, suggesting additional factors are necessary for transformation. Given the lack of additional somatic mutation, the role of epigenetic regulation in cell specification, and our prior results of germline *KMT2C* variation in infant leukaemia patients, we hypothesized that germline dysfunction of *KMT2C* altered haematopoietic specification. In isogenic *KMT2C* KO hPSCs, we found genome-wide differences in histone modifications at active and poised enhancers, leading to gene expression profiles akin to mesendoderm rather than mesoderm highlighted by a significant increase in *NODAL* expression and WNT inhibition, ultimately resulting in a lack of *in vitro* hemogenic endothelium specification. These unbiased multi-omic results provide new evidence for germline mechanisms increasing risk of early leukaemogenesis.

ARTICLE HISTORY

Received 04 January 2021

Revised 02 July 2021

Accepted 09 July 2021

KEYWORDS



Histone methyltransferases; development; gene expression; chromatin remodelling; hemogenic endothelium; pluripotency; mesoderm


Introduction:

Paediatric cancers typically harbour relatively few somatic mutations and frequently demonstrate developmentally immature phenotypes, suggesting a contribution from germline variation that might result in aberrant tissue development [1]. Our group previously found an enrichment of heterozygous germline missense mutations in *KMT2C* in infants with leukaemia, compared to healthy controls [2]. This enrichment was independent of the presence of *KMT2A* fusions, which are the hallmark somatic mutation in infant leukaemia (>75% of cases) and occur *in utero* [3]. In mammals, somatic mutations of *KMT2C* and *KMT2D* are associated with various malignancies [4], with clear evidence for tumour suppressor roles [5,6]. Given the enrichment of *KMT2C* germline mutations in infant leukaemia and the genome-wide epigenetic changes mediated by *KMT2C*, we

hypothesized that germline *KMT2C* dysfunction may adversely impact early developmental stages of haematopoiesis, or perhaps mesoderm more broadly.

The human COMPASS complexes are comprised of highly conserved proteins from yeast to humans that regulate gene expression through histone modifications [7,8]. Six different lysine methyltransferases (KMT) anchor COMPASS complexes in higher eukaryotes and are categorized into three subgroups based on homologies in amino acid sequence and subunit composition: 1] *SET1A* (NM_014712), *SET1B* (NM_015048); 2] *MLL1/KMT2A* (NM_05933), *MLL2/KMT2B* (NM_014727); 3] *MLL3/KMT2C*, (NM_170606), *MLL4/KMT2D* (NM_003482) [9]. While incompletely understood, the literature suggests that paralogs exert non-overlapping and highly specialized functions by regulating the transcription of discrete subsets of genes [9–11].

CONTACT Todd E. Druley  druley_t@wustl.edu  Department of Pediatrics, Division of Pediatric Hematology and Oncology, Washington University in St Louis School of Medicine, United States

 Supplemental data for this article can be accessed [here](#).

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

However, KMT2C and KMT2D are partially redundant in function [7,12], as both proteins play an essential role in mediating monomethylation at histone 3, lysine 4 (H3K4me1), primarily at enhancers [13]. In contrast, recent studies have highlighted other non-redundant and broad-ranging functions [14,15], such as a role for KMT2C-specific transcriptional regulation that is independent of its H3K4me1 activity on enhancers [16]. Other reports describe KMT2C-mediated histone trimethylation (H3K4me3) at promoters [17,18]. With respect to early development, *KMT2C* knockout (KO) mice die around birth with no apparent morphological abnormalities, while *KMT2DKO* mice showed early embryonic lethality around E9.5 [19]. Loss of *KMT2C* in mice also leads to aberrant myelopoiesis, causing myeloid infiltration into lymphoid organs; however, the loss of *KMT2C* alone was insufficient to drive leukaemia [20]. The role of *KMT2C* has also been characterized in nuclear receptor functioning [17,21,22], metabolism [23] and circadian rhythms [13,21].

While all SET and KMT2 proteins are epigenetic modifiers, each histone modification is associated with particular regulatory elements and mediates specific functions, enabling complex control over gene transcription [24]. KMT2C and KMT2D are associated with H3K4me1, which is a highly dynamic histone modification and correlates with cell type-specific gene expression profiles, whereas H3K4me3 marks ‘active’ promoters and is more invariant across cell types [reviewed by 19, 24]. H3K4me1, along with H3K27ac, mark ‘active’ enhancers, while the combination of H3K4me1 with H3K27me3 (mediated via polycomb proteins) is a repressive mark associated with ‘poised’ enhancers [25,26].

Currently, no comparative study exists describing the role of KMT2C and its epigenetic regulation in human pluripotent stem cells (hPSC). Pluripotent/precursor cells have multilineage potential, at the precursor stage, cells have pre-marked genomic regions, which cooperate in terminal transcriptional programmes for fate determination [27–30] and may vary in a quantifiable manner upon KMT2C dysfunction.

We found that *KMT2CKO* hPSCs have a highly variable epigenetic landscape compared to their isogenic controls and are unable to complete the endothelial to haematopoietic transition *in vitro*. To interrogate this mechanism, we have performed a multi-omics analysis revealing that *KMT2CKO* human pluripotent cells have a transcriptional profile closer to mesendoderm with a significant upregulation of NODAL/TGF β signalling.

Results

KMT2CKO hiPSCs retain pluripotency

Given our observation that infant leukaemia is enriched in heterozygous germline missense mutations in *KMT2C* [2], we interrogated the role of KMT2C in blood development by creating hPSC models (hiPSC and hESC) with isogenic *KMT2C* knockouts (Supp Figure 1a) amenable for directed haematopoietic differentiation *in vitro*. Changes in RNA and protein expression were specific to KMT2C loss (Supp Figure 1b–d). We next asked if the loss of KMT2C altered pluripotency. We observed no morphological differences between wild type and *KMT2CKO* cells (Supp Figure 2a) as well as comparable immunostaining for pluripotency markers Oct4, Sox2, and Nanog (Supp Figure 2b,c). In addition, teratoma assays were performed and the *KMT2CKO* line generated a teratoma demonstrating all three germ layers (Supp Figure 3), suggesting that loss of KMT2C does not overtly alter the hiPSCs pluripotent state.

KMT2CKO human pluripotent cells fail to specify hemogenic endothelium *in vitro*

To identify potential haematopoietic phenotypes due to *KMT2CKO*, we next differentiated our hiPSCs to mesoderm and haematopoietic progenitors using published protocols for haematopoietic specification by Keller [31], which activates the Wnt pathway via exogenous application of the GSK3 inhibitor, CHIR99021, to specify definitive haematopoietic progenitors or the Wnt inhibitor, IWP2, to enable NODAL/Activin signalling and the

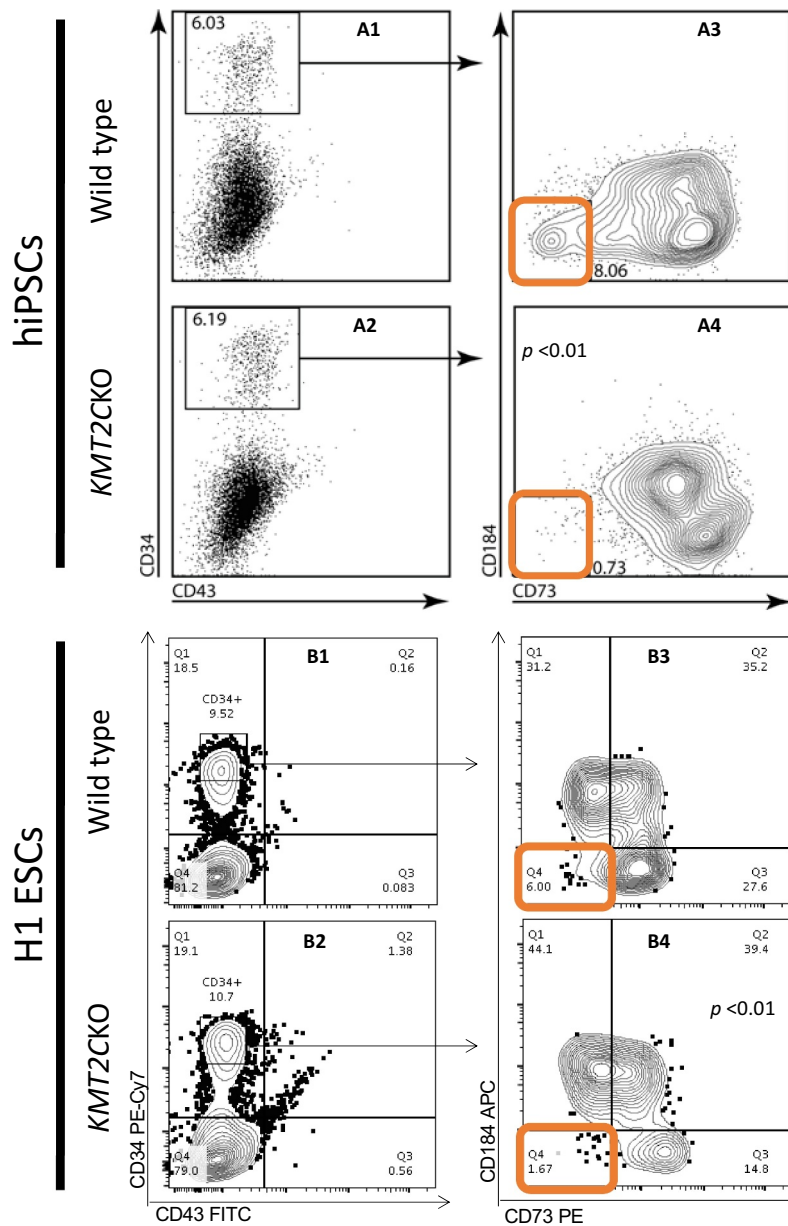


Figure 1. *KMT2CKO* pluripotent cells fail to specify hemogenic endothelium. hiPSCs (A panels) and H1 hESCs (B panels) with and without *KMT2C* were directed through haematopoietic differentiation according to the protocol by Sturgeon et al. (Sturgeon CM, *Nat Biotech* 2014). In all four cell lines, comparable amounts of CD34⁺/CD43⁻ cells were specified (A1, A2, B1, B2). To differentiate between arterial, venous and hemogenic endothelium, the CD34⁺ CD43⁻ cells were further subsorted via CD73 and CD184. HE is CD73⁻CD184⁻ (boxes). In both *KMT2CKO* pluripotent lines, there is a failure to specify hemogenic endothelium at levels equivalent to WT. Chi-square analyses found the decrease in hemogenic endothelium to be significant with *p*-values ≤ 0.01 for both human iPSCs and ESCs as listed in Experimental Procedures under 'Directed hematopoietic differentiation.'

specification of primitive haematopoiesis. As shown in Figure 1.A1 and 1.A2, both WT and *KMT2CKO* hiPSCs generate comparable numbers of CD34⁺ CD43⁻ progenitors (6.03% and 6.19%, respectively). However, progenitors of the arterial, venous, and haematopoietic system are all CD34⁺ CD43⁻.

To differentiate these subpopulations, these cells are then subsorted with CD73 and CD184. Hemogenic endothelium (HE) is CD73⁻CD184⁻, while venous endothelium is CD73⁺ CD184⁻ and arterial endothelium is CD73^{mid}CD184⁺ [31]. As shown Figure 1.A3 and 1.A4, the *KMT2CKO* hiPSCs failed

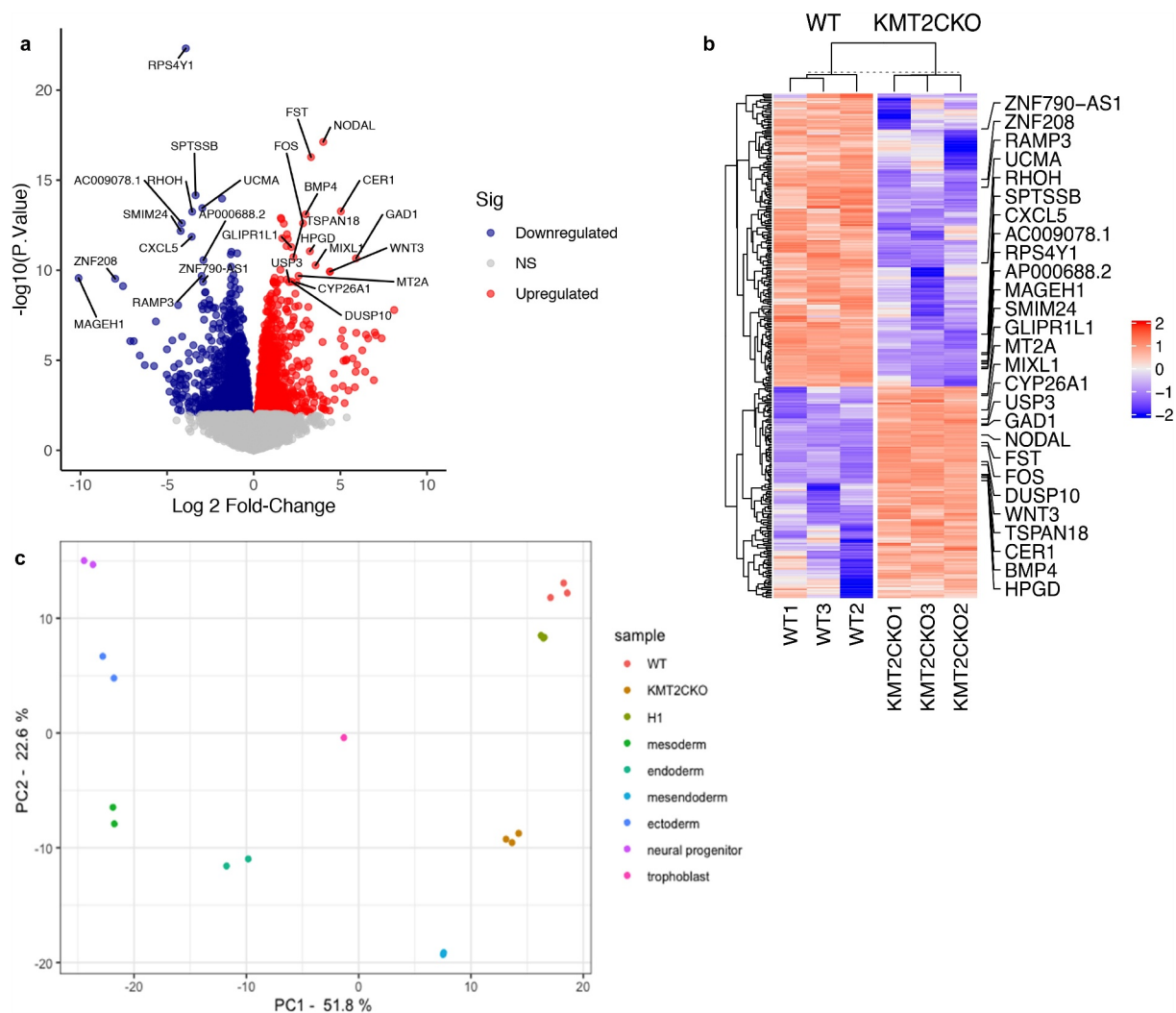


Figure 2. RNA-seq reveals 319 differentially expressed genes in *KMT2CKO* hiPSCs. (a) Volcano plot for fold change in expression (Y-axis) against the log of the fold change (X-axis). (b) Triplicates of RNA-seq from each cell line show consistent differences in gene expression across 319 genes. (c) PCA analysis reveals that *KMT2CKO* cells more closely resemble mesendoderm than any of the three germ layers.

to specify CD73-CD184- HE compared to WT. To validate this observation, we established the same *KMT2CKO* in H1 hESCs and took these cells through the same directed differentiation (Figure 1.B1-B4). H1 WT and *KMT2CKO* hESCs exhibited the same morphology from pluripotency through mesoderm (consistent with our hiPSC teratoma assay results) and embryoid body formation (Supp Figure 4). We observed the same failure of hemogenic endothelium specification (Figure 1, B4, box). This failure was specific to *KMT2C*, as H1 hESCs transduced with a scrambled gDNA vector did give rise to HE (Supp Fig 5). Furthermore, *KMT2CKO* H1 ESCs showed a significant lack of colony-forming capacity for all primitive

haematopoietic progenitors (Suppl Fig 6). In contrast, the subpopulations of venous and arterial endothelium were equivalent between WT and *KMT2CKO* hiPSCs and hESCs. Given this clear blood-specific phenotype due to *KMT2C* loss, we sought to identify a mechanism.

RNAseq analysis identifies gene expression as similar to mesendoderm

We performed transcriptome analysis to identify gene expression differences and compare against known cell types. Pairwise comparisons identified 319 differentially expressed genes upon *KMT2CKO* (133 downregulated and 186

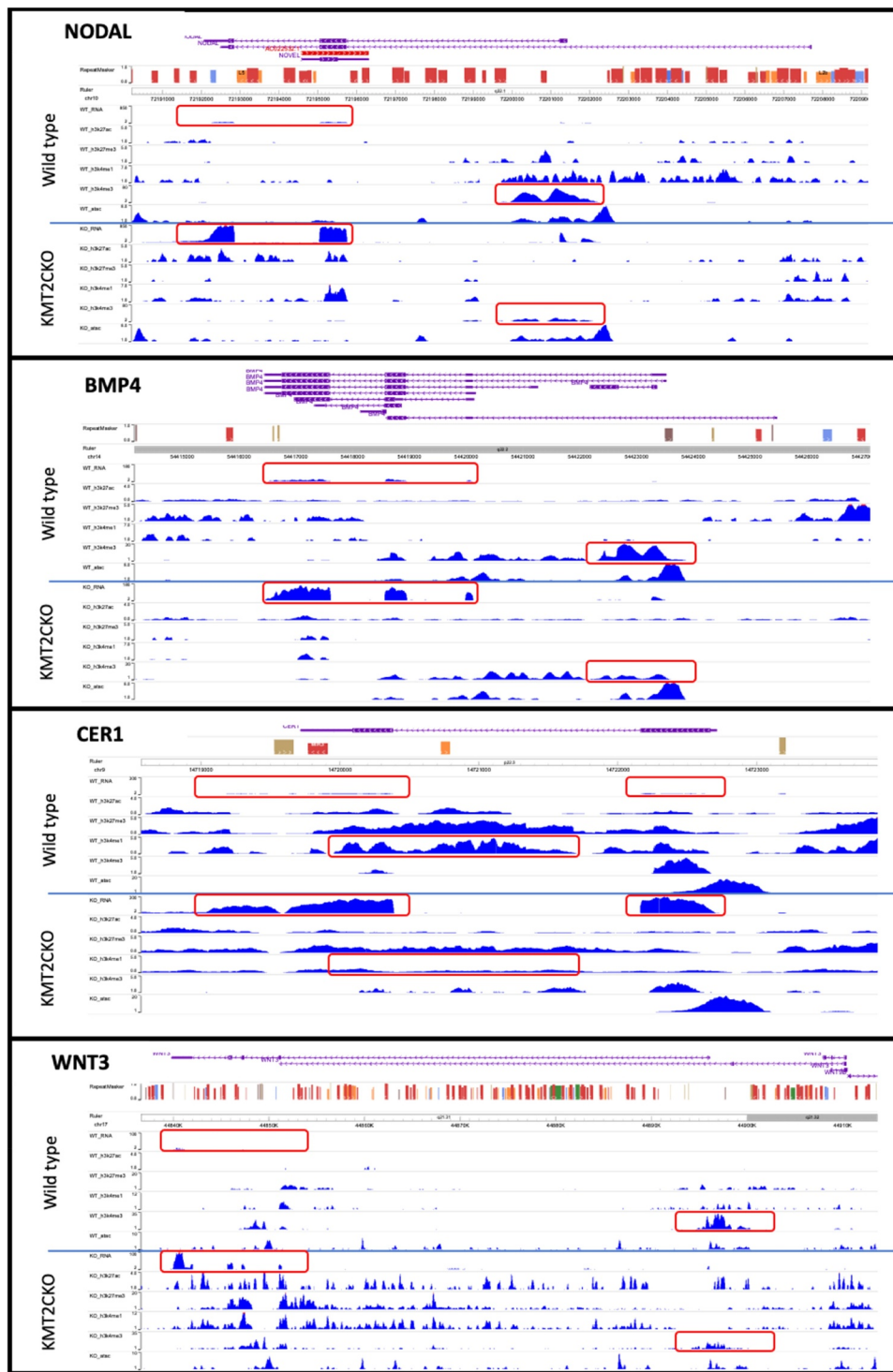


Figure 3. Epigenome tracks showing histone modifications, ATAC-seq peaks and relative RNA expression for *NODAL*, its regulator *CER1* and two of its ligands, *BMP4* and *WNT3* in WT and *KMT2CKO* hiPSCs. Each gene demonstrates higher expression in the *KMT2CKO* compared to WT (boxes in the RNA tracks) while *NODAL*, *BMP4* and *WNT3* show the expected decrease in H3K4me3 in the *KMT2CKO* (boxes in the H3K4me3 tracks).

upregulated; Supp Table 1A,B) with a false discovery rate (FDR) <0.05 and log of fold change >2 (Figure 2a,b). Genes up/downregulated >10-fold in *KMT2CKO* compared to WT are listed in Table 1.

In *KMT2CKO*, we observed the highest upregulated expression of *NODAL*, its ligands (*BMP4*, *WNT3*) and its regulators (*FST*, *CER1*, *MIXL1*, *LEFTY1*). Prior studies on human pluripotent

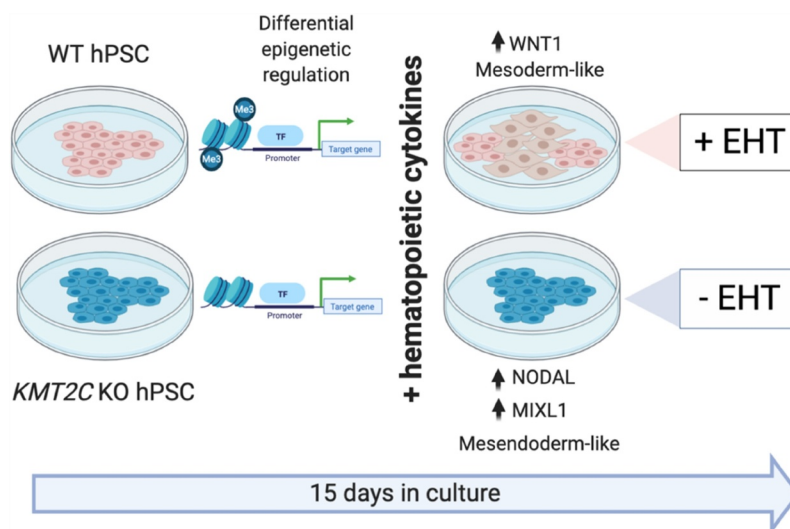


Figure 4. Schematic overview of how the lack of *KMT2C*-mediated histone modifications in hPSCs alters cell fate specification *in vitro*.

Table 1. Genes whose expression is increased or decreased >10-fold in *KMT2C* KO.

Down- or up-regulated	Rank	Gene	Fold change (FC)	Log(FC)
Down-regulated	1	RPS4Y1	-38.274	-3.920
	2	SPTSSB	-16.984	-3.355
	3	UCMA	-15.773	-2.970
	4	RHOH	-15.440	-3.552
	5	AC009078.1	-14.436	-4.149
	6	SMIM24	-13.782	-4.218
	7	CXCL5	-13.317	-3.577
	8	AP000688.2	-11.542	-2.907
	9	RAMP3	-10.443	-3.008
	10	MAGEH1	-10.297	-10.110
	11	ZNF208	-10.253	-2.930
	12	ZNF790-AS	-10.048	
Up-regulated	1	NODAL	22.946	4.010
	2	FST	21.084	3.306
	3	CER1	15.480	5.027
	4	BMP4	15.207	2.998
	5	FOS	14.426	2.847
	6	GLIPR1L1	12.508	2.164
	7	HPGD	12.191	3.240
	8	TSPAN18	11.758	2.290
	9	GAD1	11.664	5.901
	10	MIXL1	11.172	3.566
	11	WNT3	10.734	4.397
	12	MT2A	10.447	2.590
	13	DUSP10	10.062	2.192
	14	USP3	10.038	2.050
	15	CYP26A1	10.012	2.461

cells have demonstrated that NODAL/TGF β contributes to the maintenance of pluripotency [32,33] and is regulated via OCT4(POU5F1)/SOX2 TF binding and blocks differentiation [34].

Collectively, these differentially expressed genes are part of one or more tightly integrated gene regulatory networks (Supp Fig 7). With these transcriptome data, RNA-seq read counts from human

embryonic stem cell lines; HUES64 or H1 were obtained from ENCODE (<https://www.encodeproject.org/>) for each of the cellular phenotypes surveyed in the PCA plot shown in Figure 2c. From our transcriptome data for the *KMT2CKO* and WT lines, the top 100 differentially expressed genes (Supp Table 2) according to adjusted p values (not fold change) were compared against the ENCODE data. From this analysis, we conclude that the gene expression profile of *KMT2CKO* cells is closest to that of mesendoderm, suggesting that the lack of *KMT2C* prevents the

Table 2. Active enhancer transcription factor (TF) binding motifs that are significantly enriched in either wild type (WT) or *KMT2C* KO.

Enriched in	Rank	Motif for which TF binding site	HOMER P-value	TF family subtype	
Enriched in WT	1	OCT4-SOX2-TCF-NANOG	1E-34		
	2	OCT4	1E-34	Homeobox	
	Enriched in <i>KMT2C</i> KO	1	TAL1/SCL	1E-44	bHLH
		2	AR-halfsite	1E-21	NR
		3	ZFX	1E-21	ZF
		4	REST-NRSF	1E-20	ZF
		5	ASCL1	1E-20	bHLH
		6	ATOH1	1E-18	bHLH
		7	TCF12	1E-18	bHLH
		8	TCF21	1E-15	bHLH
9		EBF1	1E-14	EBF	
10		NF1-halfsite	1E-14	CTF	
11		FOX-Ebox	1E-14	Forkhead	
12		SMAD4	1E-14	MAD	
13		SMAD2	1E-14	MAD	
14		FOXA1 (GSE26831)	1E-14	Forkhead	
15		FOXA1 (GSE27824)	1E-14	Forkhead	
16		MYOG	1E-14	bHLH	
17		FOXA2	1E-13	Forkhead	
18		REPIN1/AP4	1E-13	bHLH	

cells from gene expression necessary to fully commit to either mesoderm or endoderm.

KMT2CKO alters chromatin accessibility

We next performed ATAC-seq and compared regions of differential chromatin landscape between *KMT2CKO* and WT hiPSC lines. The absence of KMT2C resulted in a substantial decrease in ATAC-seq peaks at promoter regions (a 19.83% decrease at promoters up to 3 kb of a TSS), consistent with closed chromatin and inaccessible TF binding sites. Further, this decrease was followed by a commensurate increase of 23.04% in ATAC-seq peaks at introns, downstream sequences, and distal intergenic regions (Supp Fig 8A,B) in the *KMT2CKO* line, suggesting that KMT2C's known activity at enhancers and distal regulatory elements is essential to first open specific promoters for TF binding and subsequent gene expression. The lack of KMT2C kept enhancers open for TF binding rather than allowing promoter binding sites to open.

To identify which genes may be regulated via this mechanism, we next performed TF motif enrichment within these differentially accessible ATAC regions. Open chromatin in WT was significantly enriched for 87 different motifs (Supp Table 3). Of these, binding sites for CTCF and CTCFL were most significantly enriched followed by binding sites for several homeobox (OCT4/POU5F1, OCT6/POU3F1) and high mobility group (SOX2, SOX3, SOX6, SOX10, SOX15) TFs, which includes two of the Yamanaka factors and presumably localized to promoter regions lost upon *KMT2CKO* (Supp Fig 8A). In contrast, the absence of KMT2C resulted in rearrangement of available TF binding motifs resulting in maintenance of OCT4/POU5F1 and SOX2 binding sites, but a significant increase in binding sites for the Zinc finger proteins of the cerebellum (ZIC) and their reverse complement ('Unknown ESC element') along with several ETS TF family (ERG, FLI1, ETV1, ETV2, ETS1) binding sites (Supp Table 4). Lim et al. previously established that Zic proteins maintain pluripotency in murine ESCs under the regulation of Oct4/Pou5f1, Nanog and Sox2 [35], consistent with

Table 3. Poised enhancer transcription factor (TF) binding motifs that are significantly enriched in either wild type (WT) or *KMT2C* KO.

Enriched in WT or KO	Rank	Motif for which TF binding site	HOMER P-value	TF family subtype
Enriched in WT	1	OCT4-SOX2-TCF-NANOG	1E-25	
Enriched in <i>KMT2C</i> KO	1	TAL1/SCL	1E-59	bHLH
	2	ATOH1	1E-37	bHLH
	3	FOXL2	1E-32	Forkhead
	4	FOXA2	1E-29	Forkhead
	5	REST-NRSF	1E-27	ZF
	6	TCF12	1E-24	bHLH
	7	FOXA1 (GSE26831)	1E-24	Forkhead
	8	ASCL1	1E-23	bHLH
	9	TCF21	1E-21	bHLH
	10	FOX-Ebox	1E-21	Forkhead
	11	REPIN1/AP4	1E-21	bHLH
	12	MYOG	1E-19	bHLH
	13	'Unknown ESC element' (ZIC complementary sequence)	1E-19	ZF
	14	FOXA1 (GSE27824)	1E-19	Forkhead
	15	FOXP1	1E-18	Forkhead
	16	SMAD2	1E-17	MAD
	17	ZIC	1E-17	ZF
	18	OLIG2	1E-16	bHLH
	19	EBF1	1E-16	EBF
	20	ZFX	1E-16	ZF
	21	AR-halfsite	1E-15	NR
	22	NFY	1E-15	NTF
	23	LHX1	1E-15	Homeobox
	24	NeuroD1	1E-14	bHLH
	25	SOX3	1E-13	HMG
	26	NF1-halfsite	1E-13	CTF

KMT2CKO cells retaining a pluripotent phenotype.

The histone modification landscape in hPSCs

ATAC-seq or DNA hypersensitivity mapping does not distinguish between different types of regulatory elements (active enhancers, poised enhancers, and bivalent promoters), is biased, and provides little information on domain level features [36]. Therefore, to characterize these regulatory regions, we performed ChIPmentation for four histone modifications in WT and *KMT2CKO* hPSCs. The histone modifications defining 'primed,' 'active' and 'poised' enhancers as well as 'bivalent' versus 'active' promoters are listed in Experimental Procedures and Supplementary Table 5. 'Active' enhancers (AE) correlate with tissue-specific gene expression, while 'poised' enhancers (PE) correlate with potential gene expression at subsequent developmental stages [25,37,38].

As shown in Supp Fig 9A, the proportion of each histone modification was similarly distributed across the genome of WT and *KMT2CKO*. However, the absence of KMT2C resulted in a variable distribution of histone modification between cell lines. Given that AE are the primary targets of KMT2C, differences were primarily observed for H3K4me1 marks (80% difference) and H3K27ac marks (87% difference) (Supp Fig 9B,C). In contrast, H3K27me3 marks showed a lesser difference of 56% while H3K4me3 marks at promoters showed the least difference with only 19% of peaks different between wild type and *KMT2CKO* (Supp Fig 9D,E, respectively).

(i) Comparison of the active enhancer landscape: The active enhancer landscape is mainly shaped by the cooperative binding of ubiquitous and cell-type-specific TFs [39]. As KMT2C is known to mediate the H3K4me1 at enhancers, we first compared the active enhancer landscape between wild type and *KMT2CKO* cells. As shown in the Venn diagram of Supp Fig 10A, there were a total of 29,161 active enhancer peaks called. Of these, only 2,231 (7.7%) were independent of *KMT2CKO* and shared between both lines. In contrast, 19,311 active enhancer peaks were specific to the *KMT2CKO* line, while 7,619 were specific to wild type. Supp Fig 10B,C shows the results of GO term analyses for the AE specific to WT versus *KMT2CKO*, respectively. In general, these results suggest movement away from cellular differentiation (WT) towards more functions associated with GTPase, RAS activity along with cellular junction organization and function (*KMT2CKO*).

To identify the putative functions of these *KMT2CKO*-specific active enhancer subgroups, we performed TF binding motif analysis (Table 2). In wild type, consistent with the pluripotent status of the cells and the available TF binding motif analysis from ATAC-seq, the only significantly enriched active enhancer motifs were for the cooperative binding site for OCT4/POU5F1-SOX2-TCF-NANOG and OCT4/POU5F1 alone. In contrast, the absence of KMT2C demonstrated a loss of open OCT4 TF binding sites and a significant enrichment of 18 different active enhancer TF binding motifs.

(ii) Comparison of poised enhancers: Poised enhancers (PE), marked by H3K4me1 and H3K27me3, are thought to be incapable of driving gene expression when cells are in a pluripotent state [26]. However, the loss of H3K27me3, coupled with the acquisition of H3K27 acetylation (H3K27ac), endows these enhancers with gene regulatory functions and converts PE to AE. Given this background, we identified overlapping H3K4me1 and H3K27me3 peaks to identify and compare PE marks between wild type and *KMT2CKO* pluripotent cells. In wild type hiPSCs, we identified 4,134 unique overlapping H3K4me1 and H3K27me3 marks compared to a nearly 4.5-fold increase of 18,593 unique overlapping marks in *KMT2CKO* hiPSCs. The absence of KMT2C demonstrated the expected increase in poised enhancers (Supp Fig 11A) with considerable differences in genomic loci. With respect to putative functional differences, GO terms associated with the PE marks specific to each hiPSC line generally changed from developmental functions in wild type to ion transport functions in *KMT2CKO* (Supp Fig 11B,C), supporting the lack of ion transport function by KMT2C and its broader role in guiding development.

On the global level, the substantial increase in PE following *KMT2C* knockout suggested alterations in gene regulatory networks involved in differentiation and cell specification. These genome signatures are binding sites for pioneer transcription factors and lineage determining transcription factors [40–43]. To identify TF binding sites within these differential regions, we applied motif analyses for wild type-specific and *KMT2CKO*-specific PE signature (Table 3).

In wild type, consistent with our results from ATAC-seq and AE ChIPmentation, the only significantly enriched PE motif was the cooperative binding site for OCT4/POU5F1-SOX2-TCF-NANOG (Table 3). In contrast, the absence of KMT2C demonstrated a loss of PE binding sites for OCT4 and a significant enrichment of 26 different PE TF binding motifs. Specifically, we noted that these motifs consisted primarily of binding sites for regulators of classic WNT1/b-catenin signalling (TAL1/SCL, ATOH1, TCF12, ASCL1) along with multiple forkhead proteins (FOXL2, FOXA2, FOXA1, FOX:Ebox, FOXP1). As pioneer

TFs at PE are known to unmask chromatin domains during development [44], this finding is consistent with those of Wang et al., who found FOX TFs bound to PE during specification of hESC-derived endodermal lineage intermediates [45], suggesting that knocking out KMT2C may result in human pluripotent cells being primed for endodermal fate specification.

(iii) Characterizing bivalent promoters. Pluripotent cells are enriched for promoters harbouring the activating H3K4me3 mark as well as the repressive H3K27me3 mark, a state called ‘bivalency.’ While bivalent promoters are not unique to pluripotent cells, they are enriched in these cell types, mainly marking developmental and lineage-specific genes, which are generally static but can be rapidly activated or repressed. While KMT2C is not known to have a direct role in establishing bivalency, a few studies have observed that KMT2C maintains H3K4me3 marks [17,18]. We identified a total of 5,568 bivalent promoters in WT and *KMT2CKO* cells (Supp Fig 12) without significant differences in TF binding motifs, which is consistent with KMT2C not having a clear role in establishing bivalency but more of an impact at distal regulatory elements.

Enhancers’ act synergistically to promote gene expression:

We next sought to investigate the relationship between enhancer status and gene expression. We assigned each identified enhancer to the nearest promoter, allowing a maximal distance of 500 kb between enhancer and target promoter. As expected, genes associated with AE show significantly higher average expression levels, followed by those associated with primed enhancers, then poised enhancers, and finally genes not associated with a marked enhancer (Supp Fig 13A,B; Supp Tables 6,7), regardless of WT or *KMT2CKO*. Additionally, the more AE associated with a given gene, that gene demonstrated significantly higher expression levels (Supp Fig 13 C,D; Supp Tables 8,9). These results support our interpretation that differences in active and poised enhancers are correlated with gene expression differences and that more active enhancers

associated with a given gene results in significantly higher overall expression.

This is further visualized in Figure 3 showing aligned ChIPmentation, ATAC-seq and RNA-seq peaks for *NODAL*, its regulator *CER1* and two ligands, *BMP4* and *WNT3* compared between WT and knockout. All four genes have significant mRNA overexpression in *KMT2CKO* compared to WT along with a decrease in trimethylation of H3K4 and, to a lesser extent, H3K27. Supp Fig 7 shows the interconnected gene regulatory network(s) that include these four genes along with 30 others. Similar epigenome browser tracks for the remaining genes are shown in Supp Fig 14.

Proteome and phospho-proteome suggest heterogeneity between WT and *KMT2CKO*

Since proteins are the ultimate functional effectors of activity in biological systems, we sought to correlate our epigenetic and expression results with an unbiased survey of global proteomic and phospho-proteomic expression. *KMT2CKO* displayed consistent changes in the basal proteomic and phosphorylation status of proteins (Supp Fig 15A-D). Among 678 differentially expressed proteins, 331 proteins were upregulated, while 347 proteins were downregulated (Supp Fig 15B, Supp Table 10A,B). Only 299 phospho-proteins showed differential expression – 134 phosphoproteins were upregulated, and 165 proteins were down-regulated (Supp Fig 15D, Supp Table 11A, B). As expected, KMT2C was one of the most underexpressed proteins in both samples, and in accordance with RNA-seq downregulation, one of the most downregulated proteins was RPS4Y1 (logFC(-2.02); P-value 7.2×10^{-9}) while, curiously, the most upregulated proteins were several metallothioneins (MT1A/B/E/F/G/H/M/X) along with LEFTY1 (logFC(0.93); P-value 5.0×10^{-5}), another *NODAL* regulator. To further compare to our existing datasets (transcriptome and proteome), we correlated the normalized log10-transformed transcriptome and proteome expression values to each other. This showed a significant ($p < 2.2e-16$), but relatively weak, Pearson-correlation coefficient of 0.18 (Supp Fig 15E). This weak correlation between transcript and protein levels is consistent with our

observation that *KMT2CKO* does not impact the cells' pluripotent status but impacts their ability to differentiate. Overall, the types of over- and under-represented gene set terms were similar between transcriptome, proteome and phospho-proteome (Supp Figure 9A,B) and overlapped with RNAseq GO analysis, suggesting that the transcriptome/proteome co-processing of our samples did not induce any significant biases in terms of functional complexity. To test for possible large-scale systematic compositional biases caused by *KMT2C* deletion, we performed gene set enrichment analysis via GAGE methodology from our differentially expressed proteome data with KEGG and gene ontology biological processes pathway data for WNT and NODAL pathways (Supp Table 12). Across three biological processes and the WNT pathway as a whole, the *p*-value for each analysis was ≤ 0.05 , suggesting a pathway-specific enrichment for a change of function. In sum, these results suggest that deletion of *KMT2C* reprogrammes the cis-regulatory elements that may change the actual binding position of some master regulator (e.g., TAL1) at these cis-regulatory elements, thereby impacting terminal differentiation.

Discussion

The canonical WNT/ β -catenin pathway is essential for multiple developmental milestones including haematopoietic specification [31,46] and aberrant Wnt signalling has been associated with subtypes of leukaemogenesis [reviewed in 47]. Physiologic Wnt signalling can also be regulated by *KMT2A* [48], but in *KMT2A*-rearranged leukaemias, which comprise more than 70% of infant leukaemia cases [49], Wnt signalling is fully dependent upon *KMT2A* [50]. Despite decades of model organism research on *KMT2A*-rearranged leukaemias, these fusions alone, when expressed at physiologic levels without other mutations, very rarely (if ever) induce a neo/perinatal leukaemia in murine models that phenocopies human infant leukaemia [51–53], suggesting additional factors were required for infant leukaemogenesis. To that end, we previously examined germline exomes from infant leukaemia patients and found a significant enrichment of missense germline variants in multiple

COMPASS complex members, particularly *KMT2C* [2].

Against this context, we hypothesized that the missense germline *KMT2C* mutations in infant leukaemia skews normal blood development from the very start of mesoderm differentiation such that the resulting haematopoietic progenitors are more easily transformed with the addition of a somatic driver, such as a *KMT2A* fusion. To explore the role of *KMT2C* in pluripotent cells, we focused on a multi-omic, proteomic and functional study in hPSCs and found that the absence of *KMT2C* does not impair the pluripotent phenotype, but does result in a heavily altered epigenetic landscape leading to altered gene and protein expression and ultimately, a failure of hemogenic endothelium or primitive haematopoietic specification *in vitro*, leaving the resulting cells with a transcriptional profile closer to mesendoderm than mesoderm (Figures 2,4). Our ATAC-seq results for *KMT2CKO* demonstrated a global reduction in open chromatin at promoters that bind chromatin topology regulators CTCF and CTCFL/BORIS. Enhancer-promoter interactions are mediated by architectural proteins, such as CTCF, MEDIATOR, and COHESIN, which regulate the organization of topologically associated domains (TADs) in a cell- and gene-specific manner during development [54–56]. More specifically, CTCF is required for proper expression of *Hox* gene clusters during differentiation [55]. CTCF deletion alters chromatin structure and subsequent transcription of myeloid-specific factors [57] ultimately driving aberrant *HOX* gene transcription in AML [58].

Without *KMT2C*, open chromatin shifts to accessibility for binding sites associated with the zinc finger of the cerebellum (ZIC) family of C2H2 zinc finger TFs. This is consistent with a prior report showing that *KMT2C/D* loss leads to a global reduction of chromatin interactions at enhancers in the ES cells [59]. The fact that we do not find a commensurate increase of ZIC or decrease in CTCF/CTCRL mRNA/protein expression suggests that *KMT2C* (potentially as part of its COMPASS complex) mediates histone modifications that alter the chromatin landscape but does not directly regulate transcription or translation of either gene family.

In contrast, murine models of Zic proteins have demonstrated that these transcription factors are essential for maintaining pluripotency of ES cells [60], but can also inhibit canonical Wnt/ β -catenin signalling *in vitro* and *in vivo* [61]. Consistent with our results, Zic2 was previously found to be enriched at AE and PE in ES cells and is essential for chromatin accessibility and regulation of transcriptional programmes during development [60,62]. Furthermore, ZIC2 was shown to interact with SMAD2/SMAD3 and cause early developmental NODAL-dependent transcriptional alterations at FOXA2 targets [63]. Deregulation of ZIC proteins has been associated with at least 20 different cancer types [reviewed in 64]. In some cases, ZIC family members are overexpressed while in others, DNA methylation results in a lack of ZIC protein expression. The end result is disruption of either the canonical Wnt/ β -catenin, TGF β , or sonic hedgehog pathways in different cell types at different developmental stages, thereby contributing to transformation.

We next annotated cis-regulatory elements modified via KMT2C by using histone modification patterns for AE and PE, which endow cells with the ability to interpret environmental cues correctly [37,65]. Transcription factor binding at active enhancers is a key determinant of tissue-specific gene expression [66–68] an essential step to execute developmental decisions for proper temporal and spatial control which is critical for embryonic development and correct fate decisions. We reasoned that uncovering the functionally relevant TFs associated with developmentally dynamic enhancers would identify lineage-specific regulators in controlling haematopoietic specification. Motif analysis of the KMT2C-dependent changes in AE and PE further complemented the shift towards open chromatin at ZIC binding sites, as we noted a shift away from OCT4/POU5F1 enhancers towards enhancers associated with NODAL and TGF β signalling. Of the KMT2CKO-specific enriched AE and PE (Tables 3B, 4B), nearly all are associated with NODAL/TGF β , but specifically ATOH1, ASCL1, TCF12, TAL1/SCL, SMAD2, multiple FOX genes, along with the same ZIC and its complementary TF binding

motif observed in our ATAC-seq results. This strongly implies that the lack of KMT2C has resulted in these pluripotent cells turning off WNT/ β -catenin in favour of NODAL/TGF β signalling.

This interpretation was further supported by transcriptome sequencing where the loss of KMT2C resulted in the largest fold expression increase in NODAL, itself, along with additional effectors: FST, BMP4, CER1, GAD1, MIXL1, and ligands WNT3 and BMP4. NODAL has been shown to be necessary for maintaining pluripotency in hESCs [69] as well as inhibiting mesoderm differentiation [70] and promoting endoderm differentiation [71]. Indeed, the KMT2CKO cells demonstrated a pluripotent phenotype, behaved identically to their isogenic WT counterparts *in vitro* from day 0 to day 3, and then failed to specify not only definitive hemogenic endothelium, all consistent with the increased NODAL expression, but also primitive haematopoietic progenitors which require NODAL/Activin signalling, suggesting that additional effectors remain inactive in the KMT2CKO cells.

In summary, somatic mutations in KMT2C have been implicated in various cancers and germline, missense mutations have been associated with infant leukaemia, which trace transformation to *in utero* development [3]. Given the relationship between paediatric cancers and aberrant developmental mechanisms, we sought to interrogate the role of KMT2C starting at pluripotency rather than focusing on transformation of terminally differentiated haematopoietic cells. While this multi-omic and functional assessment of KMT2C in human pluripotent cells is unique and expansive, translational interpretation to human cancer phenotypes should be cautiously interpreted as germline or somatic KMT2C mutations are heterozygous and almost always missense, suggesting a hypomorphic, rather than null, impact on protein function that likely alters cellular behaviours in more subtle ways. With respect to germline variability, we postulate that such variation does not drive transformation, but merely creates a more easily transformed cell type such that when a stochastic driver mutation (e.g., KMT2A-fusion) is present at a critical developmental stage, transformation occurs. This model

is consistent with other studies of aberrant development and early oncogenesis [1]. Future work will further explore this hypothesis with additional functional studies *in vitro* and *in vivo*.

Experimental procedures

Wild type and isogenic *KMT2CKO* hPSCs

Reprogrammed human inducible pluripotent stem cells (hiPSC) were generated from white blood cells collected from a healthy human male by the Washington University Genome Engineering and iPSC Core (GEiC). From this control line, the GEiC generated an isogenic, bi-allelic *KMT2CKO* line via CRISPR-guided non-homologous end-joining using a guide RNA targeting exon 3, resulting in truncation of the remaining 56 exons (Supp Figure 1a). The same process was also used for human embryonic stem cells (H1) (Wisconsin Stem Cell Bank). Supp Figure 1b-d demonstrates that the knockout of *KMT2C* was specific, compared to its paralogs, in hPSCs.

Teratoma assays

Teratoma assays were performed by the Washington University Mouse Genetics Core in the Division of Comparative Medicine (DCM) using the protocol published by Nelakanti [72]. Briefly, 1×10^6 cells diluted in 50 mL of Matrigel™ was injected bilaterally into the gastrocnemius of two NOD-SCID IL2Rgamma^{null} (NSG) mice. After eight weeks, the mice were sacrificed, and the muscles harvested for tumours. Tumours only grew in one of the two mice, which were evaluated independently by veterinary pathologists at the DCM.

Directed haematopoietic differentiation

Directed haematopoietic differentiation of human pluripotent stem cells was performed as published by Sturgeon and colleagues [31]. Statistical analyses of the flow sorted cells shown in Figure 1 were performed by Chi-Square analysis as documented in the table below.

RNA sequencing

Cells were cultured to 70% confluency and then washed once with PBS, trypsinized and pelleted by centrifugation at 500 g for 10 min at 4°C. Cell pellets were transferred to the Genome Technology Access Center (GTAC) at Washington University for mRNA selection, sequencing library preparation, and sequencing on the Illumina NextSeq500 platform.

RNA-seq analysis

RNA-seq reads were aligned to the Ensembl release 72 primary assemblies with STAR version 2.5.1a [73]. Gene counts were derived from the number of uniquely aligned unambiguous reads by Subread: feature count version 1.4.6-p5 [74]. All gene counts were then imported into the R/Bioconductor package EdgeR [75], and TMM normalization size factors were calculated to adjust for samples for differences in library size. Ribosomal genes and genes not expressed in the smallest group size minus one sample greater than one count-per-million were excluded from further analysis. The TMM size factors and the matrix of counts were then imported into the R/Bioconductor package Limma [76]. Weighted likelihoods based on the observed mean-variance relationship of every gene and sample were then calculated for all samples with the voomWithQualityWeights [77]. The performance of all genes was assessed with plots of the residual standard deviation of every gene to their average log-count with a robustly fitted trend line of the residuals. Differential expression analysis was then performed to analyse for differences between conditions, and the results were filtered for only those genes with Benjamini-Hochberg FDR adjusted p-values ≤ 0.05 .

ChIPmentation

ChIPmentation was carried out as previously described [78] with minor modifications. Briefly, cells were washed once with PBS followed by fixation using 1% formaldehyde in up to 1 ml PBS for 10 min at room temperature. Glycine was used to stop the reaction. Cells were collected at 500 g for

10 min at 4°C (subsequent work was performed in a 4°C cold room and used ice-cold buffers unless otherwise specified) and washed once with 150 µl ice-cold PBS supplemented with protease inhibitors (Thermo Scientific #A32955). After that, fixed cells were either stored at -80°C for future experiments or lysed in sonication buffer supplemented with a protease inhibitor, as described, and then sonicated in a Covaris microtube (AFA fibre crimp-cap) with a Covaris E220 sonicator using the following settings: Peak incident power: 200; Duty factor: 10%; Cycles per burst: 200; Treatment time: 150 seconds (or until the DNA fragments' size is in the range of 250–700 bp). Following sonication, equilibration buffer was added into the lysate. Lysates were centrifuged at 14,000 RPM at 4°C for 10 minutes. Supernatant containing the sonicated chromatin was transferred into a 1.5 ml DNA LoBind Eppendorf tube for immunoprecipitation. For each immunoprecipitation, 20 µl magnetic DynabeadTM Protein A (Life Technologies) were washed twice and resuspended in 2X PBS supplemented with 0.1% BSA. For each immunoprecipitation, 1 µg of the appropriate antibody (described below) was added and bound to beads by rotating at least 6 hours at 4°C. Blocked antibody and conjugated beads were then placed on a DYNAL Invitrogen magnetic bead separator, supernatant was aspirated, and the sonicated lysate was added to the beads followed by overnight incubation at 4°C on a rotator. Beads were washed as described in original protocol at 4°C (in a cold room) with various buffers as provided in the protocol. Beads were then resuspended in 25 µl tagmentation mix (19 µl tagmentation buffer + 1 µl Tagment DNA enzyme supplemented with 5 µl nuclease free water) from the Nextera DNA Sample Prep kit (Illumina) and incubated at 37°C for 10 minutes in a thermocycler. The beads were washed with appropriate buffer (150 µl) per the protocol and then transferred into a 1.5 mL microfuge tube. Supernatant was immediately aspirated, leaving beads attached to the wall of the tube while in place on the magnetic separator. Bead pellets were then resuspended with 45 µl elution buffer supplemented with proteinase K (NEB) and incubated for 1 hour at 55°C and then 8–10 hours at 65°C to revert formaldehyde cross-linking. After placing on the DYNAL

Invitrogen magnetic bead separator, the supernatant was transferred to a clean microfuge tube, and the beads were discarded. Finally, DNA was purified via MinElute kit (Qiagen). From this purified DNA, qPCR was performed as described in the protocol to estimate the optimum number of enrichment cycles. The final enrichment of the libraries was then performed according to protocol and subsequently purified using AMPure XP beads followed by a size selection to recover libraries with a fragment length of 250–400 bp prior to sequencing.

Antibodies used in ChIPmentation

ChIP antibodies were purchased from Diagenode: H3K4me3 (#C15410003), H3K4me1 (#C15410037), H3K27ac (#C15410174), H3K27me3 (#C15410069), Rabbit IgG (#C15410206).

ChIPmentation analysis

Biological replicates were prepared for each histone modification – H3K4me1, H3K4me3, H3K27ac, and H3K27me3 – in both WT and *KMT2CKO* hiPSC along with two replicates of rabbit IgG as a negative control. Raw sequence reads were processed using the ENCODE Transcription factor and Histone ChIP-Seq processing pipeline (<http://github.com/ENCODE-DCC/chip-seq-pipeline2>), accessed 27 February 2019). The pipeline filtered and mapped the reads to hg19, validated the quality of the data, and generated fold change signal tracks over the control samples using MACS2. Peaks were further called using epic2 [79] using a false discovery rate (FDR) of 0.05, enabling both broad and narrow histone mark peaks to be efficiently identified. Motif search around enhancer signal and bivalent promoters signal were conducted using homer v4.8.3 (<http://homer.ucsd.edu/homer/index.html>). Identifying an enrichment of differential peak-associated genes as called by Gene Ontology (<http://geneontology.org>) was performed using Bioconductor R package clusterProfiler v3.12.0 [80].

ATAC-seq library preparation, sequencing, and analysis

To map chromatin accessibility, we used the Assay for Transposase Accessible Chromatin (ATAC-seq) protocol optimized by Semenkovich [81]. Sequence reads were demultiplexed and mapped using bowtie (<http://bowtie-bio.sourceforge.net/index.shtml>) to hg19. Peaks were identified, and signal tracks were generated with MACS2 using the ENCODE ATAC-seq pipeline (<http://github.com/ENCODE-DCC/atac-seq-pipeline>), assessed on 13 May 2019. Consistency among replicates was assessed based on Irreproducible Discovery Rates (IDR). Differential binding peaks between *KMT2CKO* and wild type were identified with the R package DiffBind using an FDR <0.05. Signal tracks of fold enrichment were visualized with the WashU Epigenome browser (<https://epigenomegateway.wustl.edu>).

Definition of enhancers and promoters

As listed in Supp Table 5, promoters were defined as non-overlapping -1kb and +1kb intervals around transcription start sites (TSS). Enhancers were defined by H3K4me1 peaks and were assigned to their closest promoter, allowing for a maximum distance of 500 kb. Active enhancers were those overlapped with H3K27ac peaks. Poised and primed enhancers were assigned to promoters after excluding those associated with any active enhancers. Poised enhancers overlapped with H3K27me3, whereas primed enhancers did not. Promoters were defined by H3K4me3 peaks within 1kb of TSS. Bivalent promoters were defined by the overlapping peak of H3K4me3 and H3K27me3.

Peptide preparation, isobaric labelling, and off-line fractionation for LC-MS

The frozen cell pellets (~10 million cells) were solubilized [82] in 0.5 mL of 8 M urea buffer (8 M urea, 75 mM NaCl, 50 mM Tris (pH 8.0), 1 mM EDTA, 2 µg/mL aprotinin, 10 µg/mL leupeptin, 1 mM PMSF, 1:100 vol/vol Phosphatase Inhibitor Cocktail 2, 1:100 vol/vol Phosphatase Inhibitor Cocktail 3, 10 mM NaF) with

ultrasonication using a Covaris S220X sonicator (Peak Incident Power: 150 W, Duty Factor: 10%, cycles/burst: 500, time: 8 min, temp: 4°C). The protein content was determined by the bicinchoninic acid (BCA) method as shown in Supp Table 13. For the reference pool, 60 µg from each sample was combined and 2 × 250 µg was processed with the samples. A protein aliquot (250 µg) was digested with trypsin after reduction and alkylation of disulphide bonds. Peptides were prepared and labelled with tandem mass tag reagents prior to off-line fractionation using high-pH reversed phase chromatography [82]. Aliquots of the twenty-five fractions (~0.5 µg) were analysed using LC-MS. The 25 fractions were further combined to 13 fractions for phosphopeptide enrichment as previously described [82] and analysed by LC-MS.

Nano-LC-MS

The samples in 1% (vol/vol) aqueous FA were loaded (2.5 µL) onto a 75 µm i.d. × 50 cm Acclaim® PepMap 100 C18 RSLC column (Thermo-Fisher Scientific) on an EASY nano-LC (Thermo Fisher Scientific). The column was equilibrated using constant pressure (700 bar) with 11 µL of solvent A (1% (vol/vol) aqueous FA). The peptides were eluted using the following gradient programme with a flow rate of 300nL/min and using solvents A and B (1% (vol/vol) FA/MeCN): solvent A containing 5% B for 5 min, increased to 23% B over 105 min, to 35% B over 20 min, to 95% B over 1 min and constant 95% B for 19 min. The data were acquired in data-dependent acquisition (DDA) mode. The MS1 scans were acquired with the Orbitrap™ mass analyser over $m/z = 350$ to 1500 and resolution set to 70,000. Twelve data-dependent high-energy collisional dissociation spectra (MS2) were acquired from each MS1 scan with a mass resolving power set to 35,000, a range of $m/z = 100$ –2000, an isolation width of 1.2 m/z , and a normalized collision energy setting of 32%. The maximum injection time was 60 ms for parent-ion analysis and 120 ms for product-ion analysis. The ions that were selected for MS2 were dynamically excluded for 40 sec. The automatic gain control (AGC) was set at a target value of 3e6 ions for MS1 scans and

1e5 ions for MS2. Peptide ions with charge states of one or ≥ 7 were excluded for HCD acquisitions.

Protein identification

The unprocessed MS data from the mass spectrometer were converted to peak lists using Proteome Discoverer (version 2.1.0.81, Thermo-Fisher Scientific) with the integration of reporter-ion intensities of TMT 10-plex at a mass tolerance of ± 3.15 mDa. The MS2 spectra with charges +2, +3 and +4 were analysed using Mascot software [83] (Matrix Science, London, UK; version 2.5.1). Mascot was set up to search against a SwissProt database of human (version June 2016, 20,237 entries) and common contaminant proteins (cRAP, version 1.0 Jan. 1st, 2012, 116 entries), assuming the digestion enzyme was trypsin/P with a maximum of 4 missed cleavages allowed. The searches were performed with a fragment ion mass tolerance of 0.02 Da and a parent ion tolerance of 20 ppm. Carbamidomethylation of cysteine was specified in Mascot as a fixed modification. Deamidation of asparagine, formation of pyro-glutamic acid from N-terminal glutamine, acetylation of protein N-terminus, oxidation of methionine, and pyro-carbamidomethylation of N-terminal cysteine were specified as variable modifications. Peptide spectrum matches (PSM) were filtered at 1% false-discovery rate (FDR) by searching against a reversed database and the ascribed peptide identities were accepted. The uniqueness of peptide sequences among the database entries was determined using the principle of parsimony. Protein identities were inferred using a greedy set cover algorithm and the identities containing ≥ 2 Occam's razor peptides were accepted [84].

Protein relative quantification

The processing, quality assurance, and analysis of TMT data were performed with proteoQ (version 1.0.0.0, <https://github.com/qzhang503/proteoQ>), a tool developed with the tidyverse approach [85,86] under the free software environment for statistical computing and graphics, R (R Core Team (2019). R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>) and RStudio (RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>). Briefly, reporter-ion intensities under 10-plex TMT channels were first obtained from Mascot, followed by the removal of PSM entries from shared peptides or with intensity values lower than $1E3$. Intensity of PSMs was converted to logarithmic ratios at base two, in relative to the average intensity of reference samples within a 10-plex TMT. Under each TMT channel, Dixon's outlier removals were carried out recursively for peptides with greater than two identifying PSMs. The median of the ratios of PSM that can be assigned to the same peptide was first taken to represent the ratios of the incumbent peptide. The median of the ratios of peptides was then taken to represent the ratios of the incumbent protein.

To align protein ratios under different TMT channels, likelihood functions were first estimated for the log-ratios of proteins using finite mixture modelling, assuming two-component Gaussian mixtures (R package: mixtools: normalmixEM) [87]. The ratio distributions were then aligned in that the maximum likelihood of the log-ratios is centred at zero for each sample. Scaling normalization was performed to standardize the log-ratios of proteins across samples. To discount the influence of outliers from either log-ratios or reporter-ion intensities, the values between the 5th and 95th percentile of log-ratios and 5th and 95th percentile of intensity were used in the calculations of the standard deviations.

Informatic and statistical analysis

Metric multidimensional scaling (MDS) and Principal component analysis (PCA) of protein log₂-ratios were performed with the base R function `stats:cmdscale` and `stats:prcomp`, respectively. Heat-map visualization of protein log₂-ratios was performed with `pheatmap` (<https://rdr.io/cran/pheatmap/>). Linear modelling was performed using the contrast fit approach in Limma [76], to assess the statistical significance in protein abundance differences between indicated groups of contrasts. Adjustments of p-values for

multiple comparisons were performed with Benjamini-Hochberg (BH) correction.

Highlights

- *KMT2C* KO in hPSCs causes epigenetic differences at active and poised enhancers.
- These differences result in increased NODAL and decreased WNT signaling.
- *KMT2C* KO hPSCs expression profiling resembles mesendoderm rather than mesoderm.
- *KMT2C* KO hPSCs fail to specify hemogenic endothelium *in vitro*.

Acknowledgments

The authors would like to thank Samantha Morris, Ph.D.; Christopher Sturgeon, Ph.D.; Thor Theunissen, Ph.D. and Maggie Ferris, M.D., Ph.D., with helpful discussions surrounding development of this manuscript. The expert technical assistance of Yiling Mi and Rose Connors is gratefully acknowledged. The deep-scale proteomic experiments were performed at the Washington University Proteomics Shared Resource (WU-PSR), R. Reid Townsend, Director. The WU-PSR is supported in part by the WU Institute of Clinical and Translational Sciences (NCATS UL1 TR000448), the Mass Spectrometry Research Resource (NIGMS P41 GM103422) and the Siteman Comprehensive Cancer Center Support Grant (NCI P30 CA091842). We would also like to thank Suellen Greco, D.V.M., Ph.D., and Washington University's Division of Comparative Medicine as well as Jessica Hoisington-Lopez and the Edison Family Center of Genome Sciences and Systems Biology's Sequencing and Innovation Laboratory for assistance with data generation. Support for this project was provided by the Eli Seth Matthews Leukemia Foundation and the Kellsie's Hope Foundation to TED.

Data access and availability

All data are publicly available at the NCBI Gene Expression Omnibus under accession number GSE159003.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Cancer Institute [P30 CA091842]; National Institute of General Medical Sciences [P41 GM103422]; Eli Seth Matthews Leukemia Foundation; Kellsie's Hope Foundation; Institute for Clinical and Translational Research, Washington University School of Medicine (US) [NCATS UL1 TR000448].

Author contributions

Conceptualization, S.S.M, R.T, and T.E.D.; Methodology, S.S. M., M.T. and M.C.V., Investigation, S.S.M, M.T., P. E-G., A. B., and M.C.V.; Formal Analysis, W.Y., M.T., Q.Z., F.M.R., and W.H.W. Writing – Original Draft, S.S.M.; Writing – Review & Editing, S.S.M., P. E-G., and T.E.D. Resources, T. E.D.; Supervision, T.E.D.; Funding Acquisition, T.E.D.

Nomenclature

MeCN, acetonitrile
 FA, formic acid
 HCD, higher-energy collision-induced dissociation
 MS1, mass spectra of peptide precursors
 MS2, fragmentation mass spectra of peptides selected in narrow mass range (2 Da) from MS1 scan
nano-LC-MS, capillary liquid chromatography interfaced to a mass spectrometer
 TFA, trifluoroacetic acid
 PMSF-Phenylmethylsulfonyl fluoride
 NaF-Sodium fluoride
 EDTA-Ethylenediaminetetraacetic acid
 ZF: Zinc finger
 bHLH: Basic Helix loop Helix
 HMG: High Mobility group transcription Factor
 bZIP: Basic Leucine Zipper
 ETS: E-twenty-six
 NR: Nuclear Receptor
 HTH: Helix turn Helix
 CTF: CCAAT box binding transcription factor
 hPSC: human pluripotent stem cell
 hiPSC: human induced pluripotent stem cell
 hESC: human embryonic stem cell
 HE: hemogenic endothelium

ORCID

Todd E. Druley  <http://orcid.org/0000-0002-3245-7561>

References

- [1] Federico S, Brennan R, Dyer MA. Childhood cancer and developmental biology: a crucial partnership. *Curr Top Dev Biol.* 2011;94:1–13. doi: 10.1016/B978-0-12-380916-2.00001-2.

- [2] Valentine MC, Linabery AM, Chasnoff S, et al. Excess congenital non-synonymous variation in leukemia-associated genes in MLL-infant leukemia: a Children's Oncology Group report. *Leukemia*. 2014;28(6):1235–1241. .
- [3] Ford AM, Ridge SA, Cabrera ME, et al. In utero rearrangements in the trithorax-related oncogene in infant leukaemias. *Nature*. 1993;363(6427):358–360.
- [4] Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013;502(7471):333–339. .
- [5] Chen C, Liu Y, Rappaport AR, et al. MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. *Cancer Cell*. 2014;25(5):652–665. .
- [6] Lee J, Kim D-H, Lee S, et al. A tumor suppressive coactivator complex of p53 containing ASC-2 and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proc Natl Acad Sci U S A*. 2009a;106(21):8513–8518.
- [7] Hu D, Gao X, Morgan MA, et al. The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. *Mol Cell Biol*. 2013;33(23):4745–4754.
- [8] Sze CC, Shilatifard A. MLL3/MLL4/COMPASS family on epigenetic regulation of enhancer function and cancer. *Cold Spring Harb Perspect Med*. 2016;6(11):a026427. doi: 10.1101/cshperspect.a026427.
- [9] Shilatifard A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu Rev Biochem*. 2012;81(1):65–95.
- [10] Kouzarides T. Chromatin modifications and their function. *Cell*. 2007;128(4):693–705.
- [11] Li B, Carey M, Workman JL. The role of chromatin during transcription. *Cell*. 2007;128(4):707–719.
- [12] Herz HM, Mohan M, Garruss AS, et al. Enhancer-associated H3K4 monomethylation by trithorax-related, the drosophila homolog of mammalian MLL3/MLL4. *Genes Dev*. 2012;26(23):2604–2620.
- [13] Valekunja UK, Edgar RS, Oklejewicz M, et al. Histone methyltransferase MLL3 contributes to genome-scale circadian transcription. *Proc Natl Acad Sci U S A*. 2013;110(4):1554–1559.
- [14] Jozwik KM, Chernukhin I, Serandour AA, et al. FOXA1 directs H3K4 monomethylation at enhancers via recruitment of the methyltransferase MLL3. *Cell Rep*. 2016;17(10):2715–2723.
- [15] Zhang J, Dominguez-Sola D, Hussein S, et al. Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nat Med*. 2015;21(10):1190–1198. .
- [16] Dorigi KM, Swigut T, Henriques T, et al. Mll3 and Mll4 facilitate enhancer RNA synthesis and transcription from promoters independently of H3K4 monomethylation. *Mol Cell*. 2017;66(4):568–576.e4.
- [17] Ananthanarayanan M, Li Y, Surapureddi S, et al. Histone H3K4 trimethylation by MLL3 as part of ASCOM complex is critical for NR activation of bile acid transporter genes and is downregulated in cholestasis. *Am J Physiol Gastrointest Liver Physiol*. 2011;300(5):G771–81. doi: 10.1152/ajpgi.00499.2010. Epub 2011 Feb 17.
- [18] Rampias T, Karagiannis D, Aygeris M, et al. The lysine-specific methyltransferase KMT 2C/ MLL 3 regulates DNA repair components in cancer. *EMBO Rep*. 2019;20(3):e46821. doi: 10.15252/embr.201846821. Epub2019Jan21.
- [19] Lee JE, Wang C, Xu S, et al. H3K4 mono- And di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *Elife*. 2013;201324;2:e01503. doi: 10.7554/eLife.01503.
- [20] Arcipowski KM, Bulic M, Gurbuxani S, et al. Loss of Mll3 catalytic function promotes aberrant myelopoiesis. *PLoS One*. 2016;11(9):e0162515.
- [21] Kim DH, Rhee JC, Yeo S, et al. Crucial roles of mixed-lineage leukemia 3 and 4 as epigenetic switches of the hepatic circadian clock controlling bile acid homeostasis in mice. *Hepatology*. 2015;61(3):1012–1023.
- [22] Lee S, Kim DH, Goo YH, et al. Crucial roles for interactions between MLL3/4 and INI1 in nuclear receptor transactivation. *Mol Endocrinol*. 2009b;23(5):610–619.
- [23] Kim DH, Lee J, Lee B, et al. ASCOM controls farnesoid X receptor transactivation through its associated histone H3 lysine 4 methyltransferase activity. *Mol Endocrinol*. 2009;23(10):1556–1562.
- [24] Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? *Mol Cell*. 2013;49(5):825–837.
- [25] Creighton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 2010;107(50):21931–21936. .
- [26] Rada-Iglesias A, Bajpai R, Swigut T, et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*. 2011;470(7333):279–285.
- [27] González AJ, Setty M, Leslie CS. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat Genet*. 2015;47(11):1249–1259.
- [28] Rubin AJ, Barajas BC, Furlan-Magaril M, et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet*. 2017;49(10):1522–1528. .
- [29] Samstein RM, Arvey A, Josefowicz SZ, et al. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell*. 2012;151(1):153–166. .
- [30] Xu CR, Cole PA, Meyers DJ, et al. Chromatin “pre-pattern” and histone modifiers in a fate choice for liver and pancreas. *Science*. 2011;332(6032):963–966.
- [31] Sturgeon CM, Ditadi A, Awong G, et al. Wnt signaling controls the specification of definitive and primitive

- hematopoiesis from human pluripotent stem cells. *Nat Biotechnol.* **2014**;32(6):554–561.
- [32] Blakeley P, Fogarty NME, Del Valle I, et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Dev.* **2015**;142:3151–3165.
- [33] Vallier L, Mendjan S, Brown S, et al. Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. *Development.* **2009**;136(8):1339–1349. .
- [34] Pan G, Thomson JA. Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Res.* **2007**;17(1):42–49.
- [35] Lim LS, Loh YH, Zhang W, et al. Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Mol Biol Cell.* **2007**;18(4):1348–1358.
- [36] Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, et al. Extensive variation in chromatin states across humans. *Science.* **2013**;342(6159):750–752. .
- [37] Rada-Iglesias A, Wysocka J. Epigenomics of human embryonic stem cells and induced pluripotent stem cells: insights into pluripotency and implications for disease. *Genome Med.* **2011**;3(6):36. doi: [10.1186/gm252](https://doi.org/10.1186/gm252).
- [38] Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **2011**;21(8):1273–1283.
- [39] Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* **2010**;38(4):576–589.
- [40] Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet.* **2013**;14(6):390–403.
- [41] Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell.* **2014**;14(6):762–775.
- [42] Lee DY, Hayes JJ, Pruss D, et al. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell.* **1993**;72(1):73–84.
- [43] Levine M. Transcriptional enhancers in animal development and evolution. *Curr Biol.* **2010**;20(17):R754–63. doi: [10.1016/j.cub.2010.06.070](https://doi.org/10.1016/j.cub.2010.06.070).
- [44] Mayran A, Drouin J. Pioneer transcription factors shape the epigenetic landscape. *J Biol Chem.* **2018**;293(36):13795–13804.
- [45] Wang A, Yue F, Li Y, et al. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell.* **2015**;16(4):386–399. .
- [46] Luis TC, Ichii M, Brugman MH, et al. Wnt signaling strength regulates normal hematopoiesis and its deregulation is involved in leukemia development. *Leukemia.* **2012**;26(3):414–421.
- [47] Staal FJT, Famili F, Perez LG, et al. Aberrant Wnt signaling in leukemia. *Cancers (Basel).* **2016**;8(9):78. doi: [10.3390/cancers8090078](https://doi.org/10.3390/cancers8090078).
- [48] Wang P, Lin C, Smith ER, 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.* **2009**;29:6074–6085.
- [49] Eguchi M, Eguchi-Ishimae M, Greaves M. The role of the MLL gene in infant leukemia. *Int J Hematol.* **2003**;78(5):390–401.
- [50] Wang Y, Krivtsov AV, Sinha AU, et al. The wnt/ β -catenin pathway is required for the development of leukemia stem cells in AML. *Science.* **2010**;327(80):1650–1653.
- [51] Bueno C, Ayllón V, Montes R, et al. FLT3 activation cooperates with MLL-AF4 fusion protein to abrogate the hematopoietic specification of human ESCs. *Blood.* **2013**;121(19):3867–3878. S1-3. .
- [52] Bursen A, Schwabe K, Ruster B, et al. The AF4.MLL fusion protein is capable of inducing ALL in mice without requirement of MLL.AF4. *Blood.* **2010**;115(17):3570–3579.
- [53] Montes R, Ayllón V, Gutierrez-Aranda I, et al. Enforced expression of MLL-AF4 fusion in cord blood CD34+ cells enhances the hematopoietic repopulating cell function and clonogenic potential but is not sufficient to initiate leukemia. *Blood.* **2011**;117(18):4746–4758. .
- [54] Kagey MH, Newman JJ, Bilodeau S, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature.* **2010**;467(7314):430–435. .
- [55] Narendra V, Rocha PP, An D, et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science.* **2015**;347(6225):1017–1021.
- [56] Phillips-Cremens JE, Sauria MEG, Sanyal A, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell.* **2013**;153(6):1281–1295. .
- [57] Ouboussad L, Kreuz S, Lefevre PF. CTCF depletion alters chromatin structure and transcription of myeloid-specific factors. *J Mol Cell Biol.* **2013**;5(5):308–322.
- [58] Luo H, Wang F, Zha J, et al. CTCF boundary remodels chromatin domain and drives aberrant HOX gene transcription in acute myeloid leukemia. *Blood.* **2018**;132(8):837–848. .
- [59] Yan J, Chen SAA, Local A, et al. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res.* **2018**;28:204–220.
- [60] Luo Z, Gao X, Lin C, et al. Zic2 is an enhancer-binding factor required for embryonic stem cell specification. *Mol Cell.* **2015**;57(4):685–694.
- [61] Fujimi TJ, Hatayama M, Aruga J. Xenopus Zic3 controls notochord and organizer development through suppression of the Wnt/ β -catenin signaling pathway. *Dev Biol.* **2012**;361(2):220–231.
- [62] Frank CL, Liu F, Wijayatunge R, et al. Regulation of chromatin accessibility and Zic binding at enhancers in

- the developing cerebellum. *Nat Neurosci.* 2015;18(5):647–656. .
- [63] Houtmeyers R, Gainkam OT, Glanville-Jones HA, et al. Zic2 mutation causes holoprosencephaly via disruption of NODAL signalling. *Hum Mol Genet.* 2016;25(18):3946–3959. .
- [64] Houtmeyers R, Souopgui J, Tejpar S Deregulation of ZIC family members in oncogenesis. In: Aruga J. (eds) Zic family. *Advances in Experimental Medicine and Biology*, vol 1046. Singapore: Springer. https://doi.org/10.1007/978-981-10-7311-3_16
- [65] Bernstein BE, Mikkelsen TS, Xie X, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell.* 2006;125(2):315–326. .
- [66] Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature.* 2009;459(7243):108–112. .
- [67] Visel A, Blow MJ, Li Z, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature.* 2009;457(7231):854–858. .
- [68] Xu J, Shao Z, Glass K, et al. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell.* 2012;23(4):796–811. .
- [69] Vallier L, Reynolds D, Pedersen RA. Nodal inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway. *Dev Biol.* 2004;275(2):403–421. .
- [70] Kubo A, Shinozaki K, Shannon JM, et al. Development of definitive endoderm from embryonic stem cells in culture. *Development.* 2004;131(7):1651–1662. .
- [71] Jones CM, Kuehn MR, Hogan BL, et al. Nodal-related signals induce axial mesoderm and dorsalize mesoderm during gastrulation. *Development.* 1995;121(11):3651–3662. .
- [72] Nelakanti RV, Kooreman NG, Wu JC. Teratoma formation: a tool for monitoring pluripotency in stem cell research. *Curr Protoc Stem Cell Biol.* 2015;2015:4a.8.1–4a.8.17. .
- [73] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. .
- [74] Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30(7):923–930. .
- [75] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2009;26(1):139–140. .
- [76] Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47. .
- [77] Liu R, Holik AZ, Su S, et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res.* 2015;43(15):e97. doi: 10.1093/nar/gkv412. Epub2015Apr29. .
- [78] Schmidl C, Rendeiro AF, Sheffield NC, et al. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods.* 2015;12(10):963–965. .
- [79] Stovner EB, Sætrum P, Hancock J, et al. Epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics.* 2019;35(21):4392–4393. .
- [80] Yu G, Wang LG, Han Y, et al. ClusterProfiler: an R package for comparing biological themes among gene clusters. *Omi A J Integr Biol.* 2012;16(5):284–287. .
- [81] Semenkovich NP, Planer JD, Ahern PP, et al. Impact of the gut microbiota on enhancer accessibility in gut intraepithelial lymphocytes. *Proc Natl Acad Sci U S A.* 2016;113(51):14805–14810. .
- [82] Mertins P, Tang LC, Krug K, et al. Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. *Nat Protoc.* 2018;13(7):1632–1661. .
- [83] Perkins DN, Pappin DJC, Creasy DM, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. In: *Electrophoresis.* 1999;20(18):3551–67. doi: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2. PMID: 10612281. .
- [84] Koskinen VR, Emery PA, Creasy DM, et al. Hierarchical clustering of shotgun proteomics data. *Mol Cell Proteomics.* 2011;10(6):M110.003822. doi: 10.1074/mcp.M110.003822. .
- [85] Sauter RM. *Advanced R.* In: *Technometrics.* 2nd ed. Vol. 62. 2020. p. 417. Taylor & Francis. .
- [86] Wickham H (2017). Easily install and load the “Tidyverse.” .
- [87] Benaglia T, Chauveau D, Hunter DR, et al. mixtools: an R Package for analyzing finite mixture models. *J Stat Softw.* 2009;32(6):1–29. .