Zinc finger protein SALL4 functions through an AT-rich motif to regulate heterochromatin formation

Nikki R. Kong[1,2], Mahmoud A. Bassal[3], Hong Kee Tan[3,4], Jesse V. Kurland[5], Kol Jia Yong[3], John J. Young[6], Yang Yang[7], Fudong Li[7], Jonathan Lee[8], Yue Liu[1,2], Chan-Shuo Wu[3], Alicia Stein[1], Hongbo Luo[9], Leslie E. Silberstein[9], Martha L. Bulyk[1,5], Daniel G. Tenen[2,3,]*, Li Chai[1,2,]*

[1]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, U.S.A.;

[2]Harvard Stem Cell Institute, Boston, MA 02115, U.S.A.;

[3]Cancer Science Institute, National University of Singapore, Singapore 117599, Singapore;

[4]Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore, 117599, Singapore;

[5]Division of Genetics, Department of Medicine, Brigham and Women's Hospital/ Harvard Medical School, Boston, MA 02115, U.S.A.;

[6]Department of Biology, Simmons University, Boston, MA 02115, U.S.A.;

[7]Hefei National Laboratory for Physical Sciences at Microscale and School of Life; Sciences, University of Science and Technology of China, Hefei, Anhui, 230026, China;

[8]Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA, 02115, U.S.A.;

[9]Joint Program in Transfusion Medicine, Department of Laboratory Medicne, Children's Hospital Boston; Boston, MA, 02115 U.S.A.

*Co-corresponding authors: lchai@bwh.harvard.edu and daniel.tenen@nus.edu.sg

Abstract

The zinc finger transcription factor SALL4 is highly expressed in embryonic stem cells, down-regulated in most adult tissues, but reactivated in many aggressive cancers. This

unique expression pattern makes SALL4 an attractive target for designing therapeutic strategies. However, whether SALL4 binds DNA directly to regulate gene expression is unclear and many of its targets in cancer cells remain elusive. Here, through an unbiased screen and Cleavage Under Targets and Release Using Nuclease (CUT&RUN) experiments, we identified and validated the DNA binding domain of SALL4 and its consensus binding sequence. Combined with RNA-seq analyses after SALL4 knockdown, we discovered hundreds of SALL4 target genes that it directly regulates in aggressive liver cancer cells. Our genomic studies uncovered a novel epigenetic regulatory mechanism by which SALL4 promotes heterochromatin formation in cancer cells via repressing genes that encode a family of histone 3 lysine 9-specific demethylases (KDMs). Taken together, these results elucidated the mechanism of SALL4 DNA binding and suggest a previously unknown mechanism for its regulation of chromatin landscape of cancer cells, thus present novel pathways and molecules to target in SALL4-dependent tumors.

Keywords: SALL4, transcription, Protein Binding Microarray, KDM3A, KDM4C, heterochromatin, liver cancer, CUT&RUN, ATAC-seq.

Introduction

Spalt-like (SALL) 4 is a transcription factor that is a member of the SALL gene family, the homologs of the *Drosophila* homeotic gene *Spalt*, which are conserved throughout eukaryotes [1, 2]. *SALL4* knockout mice die at embryonic E6.5 shortly after implantation [3] and the protein plays an important role in maintaining the stemness of murine embryonic stem cells (ESCs) [3-5]. Along with Nanog, Lin28, and Esrrb, Sall4 can promote the reprogramming of mouse embryonic fibroblasts to pluripotent stem cells [6]. Furthermore, Sall4 interacts with Oct4 and Nanog directly to regulate the expression of the ESC-specific gene signature [5, 7, 8]. Its expression in early human embryos is also

crucial; heterozygous mutations in *SALL4* are associated with Okihiro Syndrome, which is typically presented with forearm malformation and eye movement defects [1, 9, 10]. Recently, SALL4 has been reported to be a neo-substrate of Cereblon E3 ligase complex in thalidomide-mediated teratogenicity [11, 12]. Normally, SALL4 is down-regulated in most adult tissues except germ cells [13-15] and hematopoietic stem cells [16]. However, it is dysregulated in hematopoietic pre-leukemias and leukemias, including high-grade myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) [17, 18]. SALL4 is also reactivated in a significant fraction of almost all solid tumors, including lung cancer, endometrial cancer, germ cell tumors, and hepatocellular carcinomas [19-22]. This unique expression pattern demonstrates that SALL4 can be a potential link between pluripotency and cancer, and thus targeted therapeutically with limited side effects in adult normal tissues. Accordingly, SALL4-positive liver cancers share a similar gene expression signature to that of fetal liver tissues, and are associated with a more aggressive cancer phenotype, drug resistance, and worse patient survival [19, 23].

Despite its important roles in pluripotency and association with certain types of cancers, it is still unclear how SALL4 functions as a transcription factor. SALL4 has four zinc finger clusters (ZFC), three of which contain either a pair or a trio of C2H2-type zinc fingers, which are thought to confer nucleic acid-binding activity [1]. However, these clusters are scattered throughout the linear polypeptide sequence and it is not known which ZFC of SALL4 is responsible for DNA binding. Demonstrating their functional importance, SALL4 ZFC with either missense or frameshift mutations are frequently found in patients with Okihiro Syndrome [10, 24-26], which is proposed to be a result of impaired zinc finger function. Furthermore, thalidomide-mediated SALL4 degradation depends on its zinc finger amino acid sequences that show species-specific selectivity [11, 12]. Despite

evidence of their functional importance, it is not known whether SALL4 binds DNA through its ZFCs directly or if so, which ZFC is responsible for binding, and what consensus sequence it prefers. Finally, while SALL4 has been shown to function as a transcriptional repressor by recruiting the histone deacetylases (HDAC)-containing Nucleosome remodeling and histone deacetylase complex (NuRD) [18, 19, 27], many of its target genes and downstream pathways have not been elucidated. Its association with the NuRD complex has led to hypothesis that SALL4 plays a role in global chromatin regulation. However, the evidence for its direct involvement with heterochromatin or euchromatin has been unclear [28-30].

Here, we identified a SALL4-dependent heterochromatin-regulating pathway, the histone 3 lysine 9-specific demethylases (KDM3/4), that may be important for its oncogenic functions. Additionally, we have used an unbiased screen to discover SALL4's DNA binding domain and its consensus binding sequence, which was found in the *KDM* genes. These results were further confirmed by a novel method of targeted in-situ genome wide profiling of SALL4 binding sites in liver cancer cells. Understanding its mechanism as a transcription factor can thus provide new insight of how SALL4-dependent pathways can be targeted in therapeutic approaches.

Results

SALL4 binds an AT-rich motif

In order to identify the SALL4 consensus binding sequence(s), we used the universal protein binding microarray (PBM) technology [31] to conduct an unbiased analysis of all possible DNA sequences to which purified FLAG-tagged SALL4 proteins bind. Using this method, we discovered that SALL4 prefers to bind an AT-rich sequence with little degeneracy: AA[A/T]TAT[T/G][A/G][T/A] (Figure 1A), where the WTATB in the

center of the motif represents the core sequence. In addition, this sequence is highly specific compared to other AT-rich sequences on the array (Supplementary Figure 1A) and FLAG peptide alone does not bind this sequence (Supplementary Figure 1B). Along its linear polypeptide sequence, SALL4 has three clusters of C2H2-type zinc finger clusters (ZFC 2-4) either in pairs or in a triplet, as well as one C2HC-type zinc finger. Zinc finger motifs are frequently associated with nucleic acid binding [32], however, it is not known which zinc fingers of SALL4 are responsible for its DNA binding activity. Therefore, we deleted two of the clusters individually and generated SALL4A mutants that lack either their ZFC2 or ZFC4 domains, hereafter referred to as AΔZFC2 and AΔZFC4, respectively, and repeated the PBM experiments. Interestingly, AΔZFC2 mutant was unaffected compared to WT SALL4A in DNA binding specificity, but the AΔZFC4 mutant was unable to bind the AT-rich consensus motif (Figure 1B, C), suggesting that ZFC4 is responsible for sequence recognition by the protein. SALL4 also has a shorter isoform resulting from alternative splicing, SALL4B, which shares only the ZFC4 domain with SALL4A. Supporting our finding that ZFC4 is the DNA sequence recognition domain of SALL4, we confirmed that WT SALL4B also binds the specific AT-rich motif but SALL4BΔZFC4 mutant lacking this domain cannot (Supplementary Figure 1C, D). To validate our PBM results, we performed electrophoretic mobility shift assays (EMSA) with two randomly picked oligonucleotides on the PBM chip, both containing the AT-rich consensus binding site. These assays demonstrated that SALL4 could shift biotinylated oligonucleotides containing the WT AT-rich sequence but not those with the consensus sequence randomly scrambled (Figure 1D, Supplementary Figure 2A, and Supplementary Table 1). Furthermore, anti-FLAG or anti-SALL4 antibodies were able to diminish or super-shift the signal, the latter through binding to and slowing down the electrophoretic mobility of the SALL4-DNA

complex, while mouse IgG isotype control could not (Figure 1E, and Figure 1F lanes 4-7 compared to lane 2). This demonstrated that the binding event was highly specific to SALL4, and could not be attributed to any other proteins co-purified with SALL4. Next, we performed EMSA experiments with AΔZFC2 and AΔZFC4 mutants described above along with a SALL4A mutant lacking ZFC3 (refers to as AΔZFC3). While AΔZFC3 can still bind WT oligonucleotides, suggesting ZFC3 is not involved in DNA binding, neither AΔZFC2 nor AΔZFC4 mutant were able to bind the oligos (Figure 1F, lanes 8-10 compared to lane 2), suggesting that deleting either ZFC2 or ZFC4 of SALL4 can impair its DNA binding ability. These results were consistent with our observation that the smaller B isoform of SALL4 lacking ZFC2 and ZFC3 bind DNA more weakly than the longer A isoform (Supplementary Figure 2B, lanes 6-9 compared to lanes 2-5).

To further validate the SALL4 DNA binding domain, we performed isothermal titration calorimetry (ITC) experiments with purified ZFC2-4 domains of SALL4 with either the WT or mutated probes used in EMSA experiments (Supplementary Table 1). ITC experiments showed that while SALL4's ZFC4 domain can bind the WT probe, neither ZFC2 nor ZFC3 domain can bind the WT oligos strongly (Supplementary Figure 3A, top panel). In comparison, none of the three ZFCs binds DNA probes with mutated AT-rich binding motif (Supplementary Figure 3B, bottom panel). Results from these experiments orthogonally confirmed our PBM and EMSA findings of a specific SALL4 motif that is AT-rich, as well as the importance of SALL's ZFC4 domain in DNA sequence recognition and binding.

The SALL4 motif is enriched in CUT&RUN binding experiments

Given the *in vitro* results demonstrating that SALL4 binds to an AT-rich DNA motif, we sought to determine if it binds a similar motif in cells. Previous SALL4 ChIP-seq

experiments were largely performed in murine cells. These experiments in human cells were challenging because of use of sub-optimal antibodies and interrogating soluble nuclear extracts via ChIP experiments in human cells may not be ideal, in that SALL4 is located in the chromatin fraction (Supplementary Figure 4A). Here, we took advantage of the availability of a highly specific antibody against human SALL4 (Cell Signaling Technology, clone D16H12, lot 2) and performed the Cleavage under targets and release using nuclease (CUT&RUN) assay [33], which eliminates the cross-linking step while generating reads with low background.

We chose SNU398 liver cancer cells to identify endogenous SALL4 binding sites genome-wide because i) these cells have high SALL4 expression compared to other cancer cells (Supplementary Figure 4B) and regular SALL4 chromatin immunoprecipitation data are available in these cells [34]; ii) SALL4 is required for the viability of a large fraction of hepatocellular carcinoma cells [19, 35], and iii) SALL4 serves as a biomarker for worse prognosis in liver cancer [23, 36, 37].

Three separate CUT&RUN experiments using SNU398 liver cancer cell nuclei revealed SALL4 binds over 11,200 common peaks genome-wide at least two-fold above isotype control IgG peak enrichment (Supplementary Table 2). Furthermore, SALL4 peaks were distributed 33.8% intergenic, 30% intronic, and 21% within the proximal promoter (Figure 2A) and can be annotated near 4,364 genes. On average, the reads are about 80-100bp long, allowing for better identification of SALL4 binding motif. Consequently, when *de novo* motif search was performed on SALL4 peaks, we observed a significant enrichment of the AT-rich motif with the core WTATB motif that were independently identified from PBM experiments (Figure 2B).

We previously published SALL4 ChIP-seq in SNU398 liver cancer cells using a specific assay that first enriched for chromatin fractions in the cells where the majority of SALL4 protein is located, which we termed CRUD-ChIP [34, 38]. Even though SALL4 CRUD-ChIP data did not reveal a strong SALL4 motif, possibly due to the longer length of the reads, they demonstrated that SALL4 is localization at many chromatin regulator gene loci including *KDM3A* [39] (Supplementary Figure 4C) and *KDM4C* [40, 41] (Supplementary Table 2), which was further confirmed by our CUT&RUN data (Figure 2C). Interestingly, when we categorized all genes encoding lysine demethylases that had SALL4 CUT&RUN peaks, those that specifically encode demethylases of di- or tri-methylated histone 3 lysine 9 were particularly enriched, compared to those encoding demethylases of tri-methylated histone 3 lysine 4. We validated one of these targets, *KDM3A*, via CRUD-ChIP-qPCR with primers targeting its CUT&RUN and ChIP binding sites in liver cancer cells (Supplementary Figure 4D). Then we used the *KDM3A* peak to design probes for EMSA experiments. We found that SALL4A binds double-stranded oligonucleotides containing the SALL4 motif, but not when this motif was mutated (Figure 2D, Supplementary Table 1).

Taken together, these genome-wide binding assays have confirmed the bona fide SALL4 motif we found *in vitro*. In addition, peaks that are common to ChIP and CUT&RUN, such as those found at the *KDM3A* and *KDM4C* for SALL4, may represent the most robust binding.

Discovery of new SALL4 gene targets in liver cancer cells

SALL4 is a transcription factor, though its target genes in liver cancer that are associated with transcriptional and chromatin regulation are not well understood. Furthermore, it has been shown that SALL4 can act as a transcriptional activator and

repressor dependent on the cellular context [5, 27, 42, 43]. To understand how SALL4 binding affects the expression of its downstream target genes, we performed RNA-seq at 72 hours after SALL4 knock-down (KD) in biological triplicates. Representative western blot showing SALL4 expression after KD with two different shRNAs (sequences found in Supplementary Table 3) compared to scrambled control shRNA (SCR) are shown in Figure 3A (RNA-seq was performed with shRNA-2). Among differentially expressed genes with p-value <0.05 (red dots in Figure 3B), we found that several chromatin modifiers were up-regulated after SALL4 KD, such as *KDM3A* and *RCOR3* [44], and both were SALL4 targets from our ChIP and CUT&RUN experiments. When we expanded this analysis to include genes with fold changes regardless of p-value, and focused on genes with SALL4-bound peaks in both CRUD-ChIP and CUT&RUN data in liver cancer cells, we found that 355 SALL4-bound genes were up-regulated upon its KD, and 259 SALL4-bound genes were down-regulated (Figure 3C). Ingenuity Pathway Analysis (IPA, Qiagen Bioinformatics) showed that many genes involved in metabolism were down-regulated after SALL4 KD in liver cancer cells, such as *ALDH1A1* [45, 46], and *DUSP5* [47] (Figure 3C, Supplementary Table 4). Intriguingly, many genes encoding transcription factors and chromatin modifiers were up-regulated after SALL4 KD (Figure 3C), among them, *KDM3A*, *KDM4C*, *RCOR3*, and JARID2 [48, 49] (complete list can be found in Supplementary Table 4). Since SALL4 can interact with the HDAC-containing nucleosome remodeling and deacetylase (NuRD) complex [27] (Supplementary Figure 5), and we have previously shown that blocking SALL4's transcriptional repressive function by interrupting its interaction with NuRD was an effective therapeutic approach in liver cancer cells [18, 19, 34], we focused on the up-regulated genes after SALL4 KD and performed qPCR analysis to validate the RNA-seq results (Figure 3D).

In order to confirm the importance of SALL4 binding on *KDM3A* gene expression, we further used CRISPR/Cas9 to disrupt the SALL4 motif in its binding site on the *KDM3A* promoter (guide RNA sequence can be found in Supplementary Table 3). After sorting cells that had successful CRISPR targeting through GFP expression on the lentiviral vector, we collected mRNA and performed qPCR. We observed a two-fold increase in expression of *KDM3A* in cells with deleted SALL4 binding sites compared to mock-transfected control cells (Supplementary Figure 6).

Our combined analyses of RNA-seq, CUT&RUN, and CRUD-ChIP data demonstrated that SALL4 directly regulates a subset of chromatin modifying genes in cancer cells, raising the possibility that SALL4 could regulate the global chromatin landscape of cells.

SALL4 regulates heterochromatin

It has been shown previously that SALL4 KD in liver cancer cells led to cell death [19] and we were able to confirm that cells could not survive after prolonged loss of SALL4 expression after transducing SNU398 cells with lentiviruses harboring two different short hairpin (sh) RNAs (Figure 4A, B, Supplementary Figure 7). However, at an earlier time point (40 hours post-viral transduction), SALL4 KD was highly efficient (Figure 4A, and detected by GFP expression from the shRNA lentiviral plasmid) but no significant cell death was observed, as GFP and DAPI double positive cells did not increase until 72 hours post-transduction (Figure 4B), allowing better assessment of true SALL4-dependent cellular functions and fewer potential secondary effects.

Because of SALL4's known association with the NuRD complex, heterochromatin regulation [28, 30], and its regulation of *KDM3/4* pathway (Figure 2 and 3), we sought to ascertain whether there were any changes in the global chromatin landscape. We

collected protein from liver cancer cells at 40 hours after SALL4 KD and performed western blotting. We first confirmed our RNA-seq data by showing that KDM3A protein level was up-regulated upon SALL4 KD at the early time point of 40hr after transduction (Figure 4C). Then, using an antibody that has specificity for H3K9me2/3, we found that while total histone 3 levels were unchanged, there was a marked reduction of global H3K9me2/3 levels after SALL4 KD through two different shRNAs (Figure 4C). We further confirmed SALL4's importance in heterochromatin regulation by performing immunofluoresence for heterochromatin protein 1α (HP1) protein after SALL4 KD. HP1 binds H3K9me2/3 and is a hallmark of heterochromatin [50]. Again, using two shRNAs targeting SALL4, we observed significant reduction of HP1 staining in the nuclei of KD cells compared to SCR control shRNA- targeted cells (Figure 4D) along with reduction in SALL4 staining (Supplementary 8).

Decreased SALL4 expression correlates with open chromatin

Loss of H3K9me2/3 and HP1 expression typically correspond to eventual open chromatin formation. To determine whether SALL4 KD had an effect on genome-wide chromatin accessibility, we performed ATAC-seq at 72 hours after KD. Analyzing results from two biological duplicates, we found that there were similar numbers of total intervals that were found in control SCR shRNA and shSALL4-1 KD cells, 371,604 and 278,455 respectively (Figure 5A). However, the majority of intervals in the SALL4 KD cells are much wider than those in SCR cells (Figure 5B). Finally, the unique intervals in the SALL4 KD cells were mostly localized in gene promoters compared to those in control KD cells (54.4% versus 28.4%, Figure 5C), suggesting that SALL4 down-regulation impacted the chromatin accessibility of many genes, and the unique promoter localization of these open chromatin regions may have direct effect on gene expression in liver cancer cells.

Discussion

Although SALL4 has been referred to as a transcriptional regulator, evidence of direct DNA binding has been lacking. Its zinc finger clusters have been individually tested in EMSA assays and shown to preferentially bind 5-hydroxymethylcytosine, and thus SALL4 was proposed to stabilize Tet2-DNA interaction during DNA demethylation in murine ESCs [51]. However, the role of SALL4 DNA binding to regulate specific gene expression was still unclear. Here, we discovered an AT-rich SALL4 binding motif via PBM assays and confirmed this finding via EMSA and ITC assays. This motif is further supported by the predictions of SALL4 zinc finger binding by zf.princeton.edu [52, 53], which identified a 4-bp ATTT motif for every zinc finger in the 2nd cluster of SALL4 and a 5-bp ATTTG motif for each finger in the 4th cluster. In addition, in a bacteria-1-hybrid (B1H) survey of over 47,000 natural C2H2 zinc fingers across many species that include the first finger of SALL4 ZFC2 and the last finger of SALL4 ZFC4 domains [54], both fingers were predicted to bind an ATTG motif. Taken together, these results further support our finding of a specific DNA motif for SALL4 binding. In addition, our novel motif was supported by CUT&RUN assays, which resulted in much smaller DNA fragments than ChIP-seq experiments, allowing for better *de novo* motif discovery. In all, we utilized three separate biochemical and two cellular assays to demonstrate SALL4 binds a unique AT-rich motif through its 4th zinc finger cluster.

Our CUT&RUN and ChIP-seq results in cancer cells, coupled with RNA-seq data from SALL4 KD in the same cells, identified hundreds of genes that SALL4 directly regulates. Of note, many of these targets are involved in transcriptional regulation or chromatin modification. We found that SALL4 can regulate heterochromatin formation through binding and repression of the *KDM* family of genes that effect changes in the

methylation status of H3K9. Our identification of the link between SALL4 KD, decreased H3K9me2/3, and increased open chromatin regions presents a novel mechanism by which SALL4 acts as a key regulator of global chromatin landscape. Interestingly, KDM3A has previously been shown to be highly expressed in post-meiotic male germ cells to control transcription of protamine in spermatids [55, 56]. Interestingly, one of the few adult tissues where SALL4 is expressed is in the spermatogonial progenitor cells, where it antagonizes PLZF transcription factor function to drive differentiation [13]. It remains to be seen whether SALL4 directly represses KDM3A in the testes, which would suggest that the SALL4-KDM3A pathway discovered here is applicable to other progenitor cell tissues. KDM3A has also been shown to be important in maintaining self-renewal property of ESCs [57]. Therefore, the SALL4/KDM pathway in embryonic development, cancer, and in spermatogenesis should be examined more closely, wherein SALL4 expression is potentially important to prevent premature over-expression of KDM3A. An unbiased shRNA screen found that KDM3A promotes a form of epithelial cell apoptosis, anoikis, through activating pro-apoptotic *BNIP3* genes [58].Therefore, SALL4's regulation of the KDM pathway may contribute to its regulation of both heterochromatin and cell death. Future studies are needed in multiple cancer cell types where SALL4 expression is known to be crucial to test its function in broadly regulating the global chromatin landscape in cancer.

Conclusions

Taken together, these results show that SALL4 regulates the chromatin landscape, and thus gene expression, in two ways: i) by directly recruiting HDAC-containing NuRD complex to deacetylate histones; and ii) by indirectly affecting heterochromatin levels through transcriptionally repressing demethylases of H3K9me2/3 (model in Figure 6).

Therefore, SALL4 is one of the key factors at the apex of a transcriptional regulatory network in liver cancer cells, where it is required for both cell survival and maintaining the heterochromatin status at certain important target genes. These findings contribute to further understanding of SALL4 function in cancer cells and the underlying molecular mechanisms, thereby uncovering novel therapeutic targets in SALL4-positive cancers.

Methods

Protein binding microarray (PBM)

SALL4 proteins were purified using M2 FLAG agarose beads (Sigma) from nuclear extracts of HEK293T cells that were transfected with FLAG-tagged SALL4. “Universal” all 10-mer double-stranded oligonucleotide arrays in 8 x 60K, GSE format (Agilent Technologies; AMADID #030236) were used to perform PBM experiments following previously described experimental protocols [31, 59]. Each SALL4A WT or mutant protein was assayed in PBM at 600nM. The PBM scan images were obtained using a GenePix 4000A Microarray Scanner (Molecular Devices). The resulting image data were processed using GenePix Pro v7.2 to obtain signal intensity data for each spot. The data were then further processed by using Masliner software (v1.02) [60] to combine scans from different intensity settings, increasing the effective dynamic range of the signal intensity values. If a dataset had any negative background-subtracted intensity (BSI) values (which can occur if the region surrounding a spot is brighter than the spot itself), consistent pseudocounts were added to all BSI values such that they all became nonnegative. All BSI values were normalized using the software for spatial de-trending providing in the Universal PBM Analysis Suite[59]. Motifs were derived using the Seed-and-Wobble algorithm, and Enologos was used to generate logos from PWMs, as previously described [31, 59].

Electrophoretic mobility shift assay

Biotinylated probes were designed based on PBM data and obtained from Integrated DNA Technology. Oligonucleotide annealing was performed by heating mixed oligonucleotides to 95 degrees for 5 minutes, and slowly cooled in a water bath (initially 70 degrees) overnight. SALL4 proteins (purified as described for PBM) were premixed with unlabeled probes or appropriate antibodies for 20 minutes at 4 degrees, then mixed with labeled probes and incubated for 20 minutes at room temperature. Binding buffers B1 and B2 containing poly dIdC blocking DNA, binding buffer C1, and stabilization buffer D (Active Motif) were used. Free DNA and protein-DNA complexes were run for 2 hours in the cold room in 6% polyacrylamide gels in tris-borate-EDTA, then transferred onto nylon membranes (0.45um pore), and visualized via streptavidin-HRP according to manufacturer's instructions (Thermo Fisher). 1ug of antibody against SALL4 (Santa Cruz EE-30) or isotype control IgG (Santa Cruz) was pre-incubated with the proteins, before incubation with the DNA probes. All EMSA probe sequences can be found in Supplementary Table 1.

Isothermal titration calorimetry

The DNA fragments encoding SALL4 ZFC2 (residues 381-436), ZFC3 (residues 566-648) and ZFC4 (residues 864-929) cloned into modified pGEX-4T1 vector (GE Healthcare) with a tobacco etch virus (TEV) cleavage site after the GST tag. All the proteins were expressed in *Escherichia coli* BL21 (DE3) cells (Novagen), purified using glutathione sepharose (GE healthcare), and cleaved by TEV protease overnight at 4 degrees to remove the GST tag. The cleaved protein was further purified by size-exclusion chromatography on a HiLoad 16/60 Superdex 75 column (GE healthcare), dialyzed with buffer C (20 mM Tris-HCl (pH 7.5), 150 mM NaCl), and concentrated for subsequent

isothermal titration calorimetry experiments (ITC). ITC assays were carried out on a Microcal PEAQ-ITC instrument (Malvern) at 25 degrees. The titration protocol consisted of a single initial injection of 1 µl, followed by 19 injections of 2 µl of the protein (concentration: 0.5-1 mM) into the sample cell containing double stranded DNA oligos (concentration: 20 µM). Data obtained from ITC assays were fitted to one-site binding model via the MicroCal PEAQ-ITC analysis software provided by the manufacturer and the oligonucleotide sequences can be found in Supplementary Table 1.

Cleavage under targets and release using nucleases (CUT&RUN)

A detailed protocol can be found on protocol.io from the Henikoff lab (dx.doi.org/10.17504/protocols.io.mgjc3un) [33]. Briefly, 2 million cell nuclei were immobilized on Concanavalin A beads after washing. SALL4 (CST D16H12) or H3K9me2/3 (Cell Signaling D4W1U) antibodies, or normal rabbit IgG (Cell Signaling DA1E) were incubated with the nuclei overnight in the presence of 0.02% digitonin at 4 degrees. The next day, 700ng/mL of proteinA-micrococcal nuclease (pA-Mnase purified in house with vector from Addgene 86973, protocol from Schmid et al.) [61] were incubated with the nuclei at 4 degrees for an hour. After washing, the tubes were placed in heat blocks on ice set to 0 degrees, CaCl2 (1mM) was added and incubated for 30 minutes before 2x Stop buffer containing EDTA was added. DNA was eluted by heat and high-speed spin, then phenol-chloroform extracted. Qubit was used to quantify purified DNA and Bioanalyzer (2100) traces were run to determine the size of the cleaved products. NEBNext Ultra II DNA library prep kit (NEB E7645) was used to make the libraries according to Liu et al.'s protocol, outlined on protocols.io (dx.doi.org/10.17504/protocols.io.wvgfe3w) [62]. Pair-end (42bp) Illumina sequencing was performed on the barcoded and amplified libraries. Detailed data analysis combining

Henikoff [33] and Orkin [62] labs' pipelines can be found on github (https://github.com/mbassalbioinformatics/CnRAP). Briefly, raw fastq files were trimmed with Trimmomatic v0.36 [63] in pair-end mode. Next, the kseq trimmer developed by the Orkin lab was run on each fastq file. BWA (v0.7.17-r1188)[64] was first run in "aln" mode on a masked hg38 genome downloaded form UCSC to create *.sai files; then BWA was run in "sampe" mode with the flag "-n 20" on the *.sai files. Afterwards, Stampy (v1.0.32) [65] was in "—sensitive" mode. Next, using SAMtools (v1.5)[66], bam files were sorted ("sort - | 0 –O bam"), had read pair mates fixed ("fixmate"), and indexed ("index"). Bam coverage maps were generated using bamCoverage from the deepTools suite (v2.5.7)[67]. The same procedure was run to align fastq files to a masked *Saccharomyces Cerevisiae* v3 (sacCer3) genome for spike-in control DNA, also downloaded form UCSC. A normalization factor was determined for each hg38 aligned replicate based on the corresponding number of proper-pairs aligned to the sacCer3 genome, as recommended in the Henikoff pipeline, this was calculated as follow: normalization factor = 10,000,000#"proper_pairs'2. Next, from the hg38 aligned bam files, "proper-paired" reads were extracted using SAMtools with the output piped into Bedtools, producing BED files of reads that have been normalized to the number of reds aligned to the sacCer3 genome. BedGraphs of these files were generated as intermediary fiels to facilitate generation of BigWig coverage maps using the bedGraphToBigWig too from UCSC (v4) [68]. For peak calling, the recently developed SEACR (v.1.1) [69]was utilized and run in "relaxed' mode to produce peak files as the BED files used were already normalized to the number of yeast spike-in reads. Subsequent peak file columns were re-arranged to facilitate motif discovery using HOMER (v4.10) [70]. Peaks were annotate using the R package ChIPSeeeker (v.1.20.0)[71]. Overlapping peak subsets within 3kb of each other were generated using

mergePeaks.py from the HOMER suite. Peak positions for those that are common to all three replicates and at least 2-fold above IgG control can be found in Supplementary Table 2. Heatmap of SALL4-bound genes encoding lysine demethylases was generated using the R package pheatmap (https://CRAN.R-project.org/package=pheatmap, Raivo Kolde. Pheatmap under R 3.6.2.

ChIP-seq ("CRUD-ChIP")

A detailed protocol was described previously in Liu *et al*. [34] based on a previous study [38]. Briefly, 10 million cells were crosslinked with 1% formaldehyde, then lysed in buffer containing 1% SDS, and spun at 20,000 RPM to isolate the chromatin. Sonication was performed in lysis buffer containing 0.1% SDS. Dynabeads Protein A (Thermo Fisher) were conjugated with SALL4 antibody (Cell Signaling, D16H12, lot 2) and incubated with sonicated chromatin overnight at 4 degrees. Washes (0.1% SDS lysis buffer, LiCl wash buffer, and tris-EDTA) were performed and chromatin-protein complex was reverse crosslinked by heating at 65 degrees with proteinase K (8ug/mL). The chromatin was incubated with RNaseA (8ug/mL) before it was extracted purified with phenol chloroform (pH 8) and precipitated by ethanol. SALL4 ChIP-qPCR was performed with primers found in Supplementary Table 5. Libraries were prepared according to manufacturer's instructions (Illumina) and bioinformatics analyses can be found in Liu et al[34].

Lentivirus-mediated gene expression knockdown and cell culture

Two shRNAs targeting SALL4 were previously described [16]**:** E5 and 507 (Supplementary Table 4); both target exon 2 of SALL4 mRNA. A pLL.3 vector containing the shRNA and GFP were transfected into HEK293T cells using TransIT-Lenti (Mirus). Viruses were collected at 48 hours and 72 hours post-transfection. After cell debris was filtered out with 45micron syringe filters, viral supernatants were spun at 20,000 RPM at 4

degrees for 2 hours, and re-suspended in RPMI media (Gibco). The viral titer was calculated by serial dilution and transduction of HeLa cells. SNU398 hepatocellular carcinoma cells were cultured in RPMI media with fetal bovine serum (Gibco). MOI of 2 were used for these cells. Transduction was performed with polybrene (8ug/mL) and spinning at 70g at room temperature for an hour. SNU398 cells were >90% GFP positive starting at 40 hours post-transduction. At either 40 or 72 hours after transduction, cells were either counted with trypan blue exclusion method or collected and stained with DAPI nuclear stain. GFP and DAPI+ dead cells were counted using a Canto II flow cytometer (BD Biosciences). Western blotting was performed by running the collected cell lysates in a 4-20% gradient tris-glycine SDS-PAGE gel, transferred onto methylcellulose, and blotted with antibodies raised against SALL4 (Santa Cruz, clone EE30), Actin (Sigma, clone AC-74), total histone 3 (Cell Signaling Technology, clone D1H2), di/tri-methyl histone 3 lysine 9 (CST, D4W1U), tri-methyl histone 3 lysine 9 (EMD/Millipore, catalog 07-442), or KDM3A (Abcam, catalog ab80598).

RNA-seq

RNA was extracted by Trizol 72 hours after SALL4 KD and libraries were made following manufacturer's instructions (Illumina). The raw fastq reads were trimmed using TrimGalore (v.0.4.5) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and mapped to hg38 genome downloaded from UCSC using STAR[72]. The read counts table was generated with the featureCounts [73]. Differential gene expression analysis was performed to compare scrambled control samples with SALL4 KD samples using DESeq2 [74] and the fold change was plotted as a volcano plot using ggplot2 (https://ggplot2.tidyverse.org) with p-value cut-off of 0.05. qPCR for selected targets were performed with primer sequences found in Supplementary Table 5.

Co-immunoprecipitation

Cells were lysed with RIPA lysis buffer and sonicated with a microtip sonicator at 90% duty, 15 bursts. The lysates were incubated with SALL4 antibody (same as ChIP) over night at 4 degrees, followed by 6 hour incubation with protein A/G beads at 4 degrees. After washing, the beads were boiled in 2x SDS sample buffer containing beta-mercaptoethanol and the supernatant was separated in Tris-glycine gels. Western blotting was performed with SALL4 antibody (EE30, Santa Cruz Biotechnology) and HDAC1/2 antibodies (Cell Signaling 8349).

Immunofluorescence staining

40 hours after SALL4 KD with shRNA 1 and 2, SNU398 liver cancer cells were fixed with 4% paraformaldehyde in PBS for 15 minutes. Fixed cells were permeabilized with 0.1% Triton-x in PBS, washed with PBS containing 0.1% Tween-20, and blocked with 3% bovine serum albumin (BSA). Primary antibodies against HP1 (Abcam ab77256) or SALL4 (Abcam ab57577) were incubated with cells overnight at 4 degrees, in PBS containing 0.3% BSA and 0.1% Tween-20. The next day, cells were washed with Tween-20/PBS, incubated with secondary anti-goat or anti-rabbit antibodies conjugated to Alexa Fluorophore 594 at room temperature for an hour. Cells were washed and stained with DAPI DNA stain for 5 minutes and mounted with Vectashield mounting medium. Images were taken with a confocal microscope (Zeiss LSM710) with the same settings for all samples.

ATAC-seq

OMNI-ATAC-seq [75] was performed with cells transduced with either scrambled shRNA (SCR) or shSALL4-1 for 72 hours. Two biological duplicates from each sample were made into libraries using the NEBNext Ultra II DNA library prep kit (NEB E7645)

according to manufacturer's instructions. Libraries were sequenced with miSeq. Raw fastq files had optical duplicates remoed using clumpify from BBMap (https://sourceforge.net/projects/bbmap). Next adaptor trimming was performed using BBDuk (from BBMap) and read were trimmed using trimmomatic. BWA was run as described above for CUT&RUN experiments. Intervals were called using Genrich and annotated using the R package ChIPSeeker (v.1.20.0) [71]. Overlapping interval subsets were generated using mergePeaks.py from the HOMER suite, and histogram of interval widths was plotted in Excel.

Declarations

All authors have consented to publication and declare that they have no conflict of interest.

Author contributions are: N.R.K., L.C., D.G.T., M.L.B., F.L., and S.P. conceived and designed the experiments; H.L and L.E.S. helped with the design of the experiments; N.R.K., H.K.T., J.V.K., K.J.Y., Y.Y., J.L., Y.L., and A.S. performed the experiments; N.R.K., M.A.B., J.J.Y., and C.S.W. performed the analyses; N.R.K., H.K.T., L.C., and D.G.T. prepared the manuscript.

Data availabilities: all CUT&RUN, ATAC-seq, and RNA-seq data have been deposted into Gene Expression Omnibus, with accession number GSE136332.

Funding: this work was supported by a Pathology Research Microgrant from the Brigham and Women's Hospital Department of Pathology. In addition, this work was supported by the National Institutes of Health (NIH) [grant number T32 HL066987 to N.R.K.]; National Institutes of Health [grant number P01 CA66996 and HL131477 to D.G.T., and R01 HG003985 to MLB,]. This work was further supported by the National Cancer Institute [grant number R35 CA197697 to D.G.T.]; the National Heart, Lung, and

Blood Institute [grant number P01 HL095489 to L.C.]; and the Leukemia and Lymphoma Society [grant number P-TRP-5855-15 to L.C.]. This work was also supported by the Singapore Ministry of Health's National Medical Research Council under its Singapore Translational Research (STaR) Investigator Award, and by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence Initiative.

We thank Bee Hui Liu for sharing data, Steve Gisselbrecht for assistance with PBM data analysis, and Yanzhou Zhang for discussion of the paper.

References

1. Al-Baradie R, Yamada K, St Hilaire C, Chan W-M, Andrews C, McIntosh N, Nakano M, Martonyi EJ, Raymond WR, Okumura S, Okihiro MM, Engle EC: **Duane radial ray syndrome (Okihiro syndrome) maps to 20q13 and results from mutations in SALL4, a new member of the SAL family.** *Am J Hum Genet* 2002, **71**:1195–1199.

2. Kohlhase J, Schuh R, Dowe G, Kühnlein RP, Jäckle H, Schroeder B, Schulz-Schaeffer W, Kretzschmar HA, Köhler A, Muller U, Raab-Vetter M, Burkhardt E, Engel W, Stick R: **Isolation, characterization, and organ-specific expression of two novel human zinc finger genes related to the Drosophila gene spalt.** *Genomics* 1996, **38**:291–298.

3. Sakaki-Yumoto M, Kobayashi C, Sato A, Fujimura S, Matsumoto Y, Takasato M, Kodama T, Aburatani H, Asashima M, Yoshida N, Nishinakamura R: **The murine homolog of SALL4, a causative gene in Okihiro syndrome, is essential for embryonic stem cell proliferation, and cooperates with Sall1 in anorectal, heart, brain and kidney development.** *Development* 2006, **133**:3005–3013.

4. Elling U, Klasen C, Eisenberger T, Anlag K, Treier M: **Murine inner cell mass-derived lineages depend on Sall4 function.** *Proc Natl Acad Sci USA* 2006, **103**:16319–16324.

5. Zhang J, Tam W-L, Tong GQ, Wu Q, Chan H-Y, Soh B-S, Lou Y, Yang J, Ma Y, Chai L, Ng H-H, Lufkin T, Robson P, Lim B: **Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1.** *Nat Cell Biol* 2006, **8**:1114–1123.

6. Buganim Y, Markoulaki S, van Wietmarschen N, Hoke H, Wu T, Ganz K, Akhtar-Zaidi B, He Y, Abraham BJ, Porubsky D, Kulenkampff E, Faddah DA, Shi L, Gao Q, Sarkar S, Cohen M, Goldmann J, Nery JR, Schultz MD, Ecker JR, Xiao A, Young RA, Lansdorp PM, Jaenisch R: **The developmental potential of iPSCs is greatly influenced by reprogramming factor selection.** *Cell Stem Cell* 2014, **15**:295–309.

7. Wu Q, Chen X, Zhang J, Loh Y-H, Low T-Y, Zhang W, Zhang W, Sze S-K, Lim B, Ng H-H: **Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells.** *J Biol Chem* 2006, **281**:24090–24094.

8. Lim CY, Tam W-L, Zhang J, Ang HS, Jia H, Lipovich L, Ng H-H, Wei C-L, Sung WK, Robson P, Yang H, Lim B: **Sall4 regulates distinct transcription circuitries in different blastocyst-derived stem cell lineages**. *Cell Stem Cell* 2008, **3**:543–554.

9. Kohlhase J, Heinrich M, Schubert L, Liebers M, Kispert A, Laccone F, Turnpenny P, Winter RM, Reardon W: **Okihiro syndrome is caused by SALL4 mutations.** *Hum Mol Genet* 2002, **11**:2979–2987.

10. Borozdin W, Wright MJ, Hennekam RCM, Hannibal MC, Crow YJ, Neumann TE, Kohlhase J: **Novel mutations in the gene SALL4 provide further evidence for acro-renal-ocular and Okihiro syndromes being allelic entities, and extend the phenotypic spectrum.** *J Med Genet* 2004, **41**:e102.

11. Donovan KA, An J, Nowak RP, Yuan JC, Fink EC, Berry BC, Ebert BL, Fischer ES: **Thalidomide promotes degradation of SALL4, a transcription factor implicated in Duane Radial Ray syndrome.** *Elife* 2018, **7**.

12. Matyskiela ME, Couto S, Zheng X, Lu G, Hui J, Stamp K, Drew C, Ren Y, Wang M, Carpenter A, Lee C-W, Clayton T, Fang W, Lu C-C, Riley M, Abdubek P, Blease K, Hartke J, Kumar G, Vessey R, Rolfe M, Hamann LG, Chamberlain PP: **SALL4 mediates teratogenicity as a thalidomide-dependent cereblon substrate.** *Nat Chem Biol* 2018, **14**:981–987.

13. Hobbs RM, Fagoonee S, Papa A, Webster K, Altruda F, Nishinakamura R, Chai L, Pandolfi PP: **Functional antagonism between Sall4 and Plzf defines germline progenitors.** *Cell Stem Cell* 2012, **10**:284–298.

14. Yamaguchi YL, Tanaka SS, Kumagai M, Fujimoto Y, Terabayashi T, Matsui Y, Nishinakamura R: **Sall4 is essential for mouse primordial germ cell specification by suppressing somatic cell program genes.** *Stem Cells* 2015, **33**:289–300.

15. Chan A-L, La HM, Legrand JMD, Mäkelä J-A, Eichenlaub M, De Seram M, Ramialison M, Hobbs RM: **Germline Stem Cell Activity Is Sustained by SALL4-Dependent Silencing of Distinct Tumor Suppressor Genes.** *Stem Cell Reports* 2017, **9**:956–971.

16. Gao C, Kong NR, Li A, Tatetu H, Ueno S, Yang Y, He J, Yang J, Ma Y, Kao GS, Tenen DG, Chai L: **SALL4 is a key transcription regulator in normal human hematopoiesis.** *Transfusion* 2013, **53**:1037–1049.

17. Ma Y, Cui W, Yang J, Qu J, Di C, Amin HM, Lai R, Ritz J, Krause DS, Chai L: **SALL4, a novel oncogene, is constitutively expressed in human acute myeloid leukemia (AML) and induces AML in transgenic mice**. *Blood* 2006, **108**:2726–2735.

18. Gao C, Dimitrov T, Yong KJ, Tatetsu H, Jeong H-W, Luo HR, Bradner JE, Tenen DG, Chai L: **Targeting transcription factor SALL4 in acute myeloid leukemia by**

interrupting its interaction with an epigenetic complex. *Blood* 2013, **121**:1413–1421.

19. Yong KJ, Gao C, Lim JSJ, Yan B, Yang H, Dimitrov T, Kawasaki A, Ong CW, Wong K-F, Lee S, Ravikumar S, Srivastava S, Tian X, Poon RT, Fan ST, Luk JM, Dan YY, Salto-Tellez M, Chai L, Tenen DG: **Oncofetal gene SALL4 in aggressive hepatocellular carcinoma.** *N Engl J Med* 2013, **368**:2266–2276.

20. Li A, Jiao Y, Yong KJ, Wang F, Gao C, Yan B, Srivastava S, Lim GSD, Tang P, Yang H, Tenen DG, Chai L: **SALL4 is a new target in endometrial cancer.** *Oncogene* 2013.

21. Miettinen M, Wang Z, McCue PA, Sarlomo-Rikala M, Rys J, Biernat W, Lasota J, Lee Y-S: **SALL4 expression in germ cell and non-germ cell tumors: a systematic immunohistochemical study of 3215 cases.** *Am J Surg Pathol* 2014, **38**:410–420.

22. Yong KJ, Li A, Ou W-B, Hong CKY, Zhao W, Wang F, Tatetsu H, Yan B, Qi L, Fletcher JA, Yang H, Soo R, Tenen DG, Chai L: **Targeting SALL4 by entinostat in lung cancer.** *Oncotarget* 2016, **7**:75425–75440.

23. Oikawa T, Kamiya A, Zeniya M, Chikada H, Hyuck AD, Yamazaki Y, Wauthier E, Tajiri H, Miller LD, Wang XW, Reid LM, Nakauchi H: **Sal-like protein 4 (SALL4), a stem cell biomarker in liver cancers.** *Hepatology* 2013, **57**:1469–1483.

24. Kohlhase J, Schubert L, Liebers M, Rauch A, Becker K, Mohammed SN, Newbury-Ecob R, Reardon W: **Mutations at the SALL4 locus on chromosome 20 result in a range of clinically overlapping phenotypes, including Okihiro syndrome, Holt-Oram syndrome, acro-renal-ocular syndrome, and patients previously reported to represent thalidomide embryopathy.** *J Med Genet* 2003, **40**:473–478.

25. Miertus J, Borozdin W, Frecer V, Tonini G, Bertok S, Amoroso A, Miertus S, Kohlhase J: **A SALL4 zinc finger missense mutation predicted to result in increased DNA binding affinity is associated with cranial midline defects and mild features of Okihiro syndrome.** *Hum Genet* 2006, **119**:154–161.

26. Terhal P, Rösler B, Kohlhase J: **A family with features overlapping Okihiro syndrome, hemifacial microsomia and isolated Duane anomaly caused by a novel SALL4 mutation.** *Am J Med Genet A* 2006, **140**:222–226.

27. Lu J, Jeong H-W, Jeong H, Kong N, Yang Y, Carroll J, Luo HR, Silberstein LE, Yupoma, Chai L: **Stem cell factor SALL4 represses the transcriptions of PTEN and SALL1 through an epigenetic repressor complex.** *PLoS ONE* 2009, **4**:e5577.

28. Böhm J, Kaiser FJ, Borozdin W, Depping R, Kohlhase J: **Synergistic cooperation of Sall4 and Cyclin D1 in transcriptional repression.** *Biochemical and Biophysical Research Communications* 2007, **356**:773–779.

29. Sathyan KM, Shen Z, Tripathi V, Prasanth KV, Prasanth SG: **A BEN-domain-containing protein associates with heterochromatin and represses transcription.** *J Cell Sci* 2011, **124**:3149–3163.

30. Kim J, Xu S, Xiong L, Yu L, Fu X, Xu Y: **SALL4 promotes glycolysis and chromatin remodeling via modulating HP1α-Glut1 pathway.** *Oncogene* 2017, **36**:6472–6479.

31. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities.** *Nat Biotechnol* 2006, **24**:1429–1435.

32. Struhl K: **Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins.** *Trends Biochem Sci* 1989, **14**:137–140.

33. Skene PJ, Henikoff JG, Henikoff S: **Targeted in situ genome-wide profiling with high efficiency for low cell numbers.** *Nat Protoc* 2018, **13**:1006–1019.

34. Liu BH, Jobichen C, Chia CSB, Chan THM, Tang JP, Chung TXY, Li J, Poulsen A, Hung AW, Koh-Stenta X, Tan YS, Verma CS, Tan HK, Wu C-S, Li F, Hill J, Joy J, Yang H, Chai L, Sivaraman J, Tenen DG: **Targeting cancer addiction for SALL4 by shifting its transcriptome with a pharmacologic peptide.** *Proceedings of the National Academy of Sciences* 2018, **115**:E7119–E7128.

35. Zhou W, Zou B, Liu L, Cui K, Gao J, Yuan S, Cong N: **MicroRNA-98 acts as a tumor suppressor in hepatocellular carcinoma via targeting SALL4.** *Oncotarget* 2016, **7**:74059–74073.

36. Tanaka Y, Aishima S, Kohashi K, Okumura Y, Wang H, Hida T, Kotoh K, Shirabe K, Maehara Y, Takayanagi R, Oda Y: **Spalt-like transcription factor 4 immunopositivity is associated with epithelial cell adhesion molecule expression in combined hepatocellular carcinoma and cholangiocarcinoma.** *Histopathology* 2016, **68**:693–701.

37. Wang H, Kohashi K, Yoshizumi T, Okumura Y, Tanaka Y, Shimokawa M, Iwasaki T, Aishima S, Maehara Y, Oda Y: **Coexpression of SALL4 with HDAC1 and/or HDAC2 is associated with underexpression of PTEN and poor prognosis in patients with hepatocellular carcinoma.** *Hum Pathol* 2017, **64**:69–75.

38. Yang BX, Farran El CA, Guo HC, Yu T, Fang HT, Wang HF, Schlesinger S, Seah YFS, Goh GYL, Neo SP, Li Y, Lorincz MC, Tergaonkar V, Lim T-M, Chen L, Gunaratne J, Collins JJ, Goff SP, Daley GQ, Li H, Bard FA, Loh Y-H: **Systematic identification of factors for provirus silencing in embryonic stem cells.** *Cell* 2015, **163**:230–245.

39. Gray SG, Iglesias AH, Lizcano F, Villanueva R, Camelo S, Jingu H, Teh BT, Koibuchi N, Chin WW, Kokkotou E, Dangond F: **Functional characterization of JMJD2A, a histone deacetylase- and retinoblastoma-binding protein.** *J Biol Chem* 2005, **280**:28507–28518.

40. Cloos PAC, Christensen J, Agger K, Maiolica A, Rappsilber J, Antal T, Hansen KH, Helin K: **The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3.** *Nature* 2006, **442**:307–311.

41. Ponnaluri VKC, Vavilala DT, Mukherji M: **Studies on substrate specificity of Jmjd2a-c histone demethylases.** *Biochemical and Biophysical Research*

Communications 2011, **405**:588–592.

42. Li A, Yang Y, Gao C, Lu J, Jeong H-W, Liu BH, Tang P, Yao X, Neuberg D, Huang G, Tenen DG, Chai L: **A SALL4/MLL/HOXA9 pathway in murine and human myeloid leukemogenesis.** *J Clin Invest* 2013, **123**:4195–4207.

43. Young JJ, Kjolby RAS, Kong NR, Monica SD, Harland RM: **Spalt-like 4 promotes posterior neural fates via repression of pou5f3 family members in Xenopus.** *Development* 2014, **141**:1683–1693.

44. Barrios ÁP, Gómez AV, Sáez JE, Ciossani G, Toffolo E, Battaglioli E, Mattevi A, Andrés ME: **Differential properties of transcriptional complexes formed by the CoREST family.** *Molecular and Cellular Biology* 2014, **34**:2760–2770.

45. Kiefer FW, Orasanu G, Nallamshetty S, Brown JD, Wang H, Luger P, Qi NR, Burant CF, Duester G, Plutzky J: **Retinaldehyde dehydrogenase 1 coordinates hepatic gluconeogenesis and lipid metabolism.** *Endocrinology* 2012, **153**:3089–3099.

46. Yang C-K, Wang X-K, Liao X-W, Han C-Y, Yu T-D, Qin W, Zhu G-Z, Su H, Yu L, Liu X-G, Lu S-C, Chen Z-W, Liu Z, Huang K-T, Liu Z-T, Liang Y, Huang J-L, Xiao K-Y, Peng M-H, Winkle CA, O'Brien SJ, Peng T: **Aldehyde dehydrogenase 1 (ALDH1) isoform expression and potential clinical implications in hepatocellular carcinoma.** *PLoS ONE* 2017, **12**:e0182208.

47. Kovanen PE, Rosenwald A, Fu J, Hurt EM, Lam LT, Giltnane JM, Wright G, Staudt LM, Leonard WJ: **Analysis of gamma c-family cytokine target genes. Identification of dual-specificity phosphatase 5 (DUSP5) as a regulator of mitogen-activated protein kinase activity in interleukin-2 signaling.** *J Biol Chem* 2003, **278**:5205–5213.

48. Shen X, Kim W, Fujiwara Y, Simon MD, Liu Y, Mysliwiec MR, Yuan G-C, Lee Y, Orkin SH: **Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells.** *Cell* 2009, **139**:1303–1314.

49. Li G, Margueron R, Ku M, Chambon P, Bernstein BE, Reinberg D: **Jarid2 and PRC2, partners in regulating gene expression.** *Genes Dev* 2010, **24**:368–380.

50. Zeng W, Ball AR, Yokomori K: **HP1: heterochromatin binding proteins working the genome.** *Epigenetics* 2010, **5**:287–292.

51. Xiong J, Zhang Z, Chen J, Huang H, Xu Y, Ding X, Zheng Y, Nishinakamura R, Xu G-L, Wang H, Chen S, Gao S, Zhu B: **Cooperative Action between SALL4A and TET Proteins in Stepwise Oxidation of 5-Methylcytosine.** *Molecular Cell* 2016, **64**:913–925.

52. Persikov AV, Osada R, Singh M: **Predicting DNA recognition by Cys2His2 zinc finger proteins.** *Bioinformatics* 2009, **25**:22–29.

53. Persikov AV, Singh M: **De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins.** *Nucleic Acids Res* 2014, **42**:97–108.

54. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM, Greenblatt J, Frey BJ, Hughes TR: **C2H2 zinc finger proteins greatly expand the human regulatory lexicon**. *Nat Biotechnol* 2015, **33**:555–562.

55. Yamanaka S: **Patient-specific pluripotent stem cells become even more accessible**. *Cell Stem Cell* 2010, **7**:1–2.

56. Okada Y, Scott G, Ray MK, Mishina Y, Zhang Y: **Histone demethylase JHDM2A is critical for Tnp1 and Prm1 transcription and spermatogenesis.** *Nature* 2007, **450**:119–123.

57. Loh Y-H, Zhang W, Chen X, George J, Ng H-H: **Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells.** *Genes Dev* 2007, **21**:2545–2557.

58. Pedanou VE, Gobeil S, Tabariès S, Simone TM, Zhu LJ, Siegel PM, Green MR: **The histone H3K9 demethylase KDM3A promotes anoikis by transcriptionally activating pro-apoptotic genes BNIP3 and BNIP3L.** *Elife* 2016, **5**.

59. Berger MF, Bulyk ML: **Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors.** *Nat Protoc* 2009, **4**:393–411.

60. Dudley AM, Aach J, Steffen MA, Church GM: **Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range.** *Proc Natl Acad Sci USA* 2002, **99**:7554–7559.

61. Schmid M, Durussel T, Laemmli UK: **ChIC and ChEC; genomic mapping of chromatin proteins.** *Molecular Cell* 2004, **16**:147–157.

62. Liu N, Hargreaves VV, Zhu Q, Kurland JV, Hong J, Kim W, Sher F, Macias-Trevino C, Rogers JM, Kurita R, Nakamura Y, Yuan G-C, Bauer DE, Xu J, Bulyk ML, Orkin SH: **Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch.** *Cell* 2018, **173**:430–442.e17.

63. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30**:2114–2120.

64. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–1760.

65. Lunter G, Goodson M: **Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.** *Genome Res* 2011, **21**:936–939.

66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.

67. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T: **deepTools: a flexible platform for exploring deep-sequencing data.** *Nucleic Acids Res* 2014, **42**(Web Server issue):W187–91.

68. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics* 2010, **26**:2204–2207.

69. Meers MP, Tenenbaum D, Henikoff S: **Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling.** *Epigenetics Chromatin* 2019, **12**:42.

70. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities**. *Molecular Cell* 2010, **38**:576–589.

71. Yu G, Wang L-G, He Q-Y: **ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization.** *Bioinformatics* 2015, **31**:2382–2383.

72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29**:15–21.

73. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30**:923–930.

74. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15**:550.

75. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY: **An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues.** *Nat Meth* 2017, **14**:959–962.

Figure 1. Discovery of a novel SALL4 DNA binding motif. **A)** Specific DNA sequence motif bound by SALL4A in universal PBM assays [31].The color bar above the position weighted matrix (PWM) indicates the linear structure of SALL4 and the 4 zinc finger clusters are denoted by black ovals. **B)** DNA sequences bound by SALL4A mutant lacking zinc finger cluster 2 (AΔZFC2). **C)** DNA sequences bound by SALL4A mutant lacking zinc finger cluster 4 (AΔZFC4). **D)** EMSA showing SALL4A shifts the AT-rich motif-containing oligos (lanes 1-3) but not when the motif is scrambled (lanes 4-6, gel cut for clarity) **E)** SALL4-DNA complex is super-shifted by a SALL4-specific monoclonal antibody (lane 4) but not by mouse IgG isotype control (lane 5). **F)** EMSA showing that SALL4-DNA complex was reduced or super-shifted in the presence of FLAG or SALL4 antibodies (1ug pre-incubated with proteins before the addition of DNA probes). Lanes 8-10 in the right panel show that the AΔZFC3 mutant binds to the same WT sequence, whereas binding by AΔZFC2 and AΔZFC4 is abrogated. Sequences for the oligonucleotides used for EMSA can be found in Supplementary Table 1.

Figure 2. SALL4 CUT&RUN showing its binding genome-wide in liver cancer cells. **A)** The genomic distribution of SALL4 CUT&RUN peaks from three biological triplicates. **B)** The top HOMER motifs from analysis of CUT&RUN peaks. The table on the right shows p-values and % of peaks containing the particular motif. **C)** Heatmap showing SALL4 CUT&RUN peaks that are enriched above IgG isotype control at genes that encode histone lysine demethylases, which were categorized as H3K9-specific, H3K4-specific, and H3K4/9-specific. **D)** EMSA showing that SALL4 binds *KDM3A* promoter region containing the AT-rich motif, but not the mutated sequence (Supplementary Table 1).

Figure 3. RNA-seq data revealed direct SALL4 target genes. **A)** Representative western blot showing SALL4 knockdown (KD) efficiency in liver cancer cells using two different

shRNAs and scrambled (SCR) as control; beta-Actin levels served as loading control. **B)** Volcano plot showing genes that are down- or up-regulated significantly after SALL4 KD with log2 fold change represented on the x-axis; red dots denote differentially expressed genes with p-values <0.05; experiment performed in biological triplicates with shSALL4-2 compared to SCR 72 hours after transduction. **C)** Ingenuity Pathway Analysis of SALL4-bound peaks near genes that were either down- or up-regulated after SALL4 KD (expanded analysis from (B). **D)** Quantitative real-time PCR (qPCR) analysis of three SALL4 direct targets 40 hours after SALL4 KD, results are summarized from either three or four independent experiments.

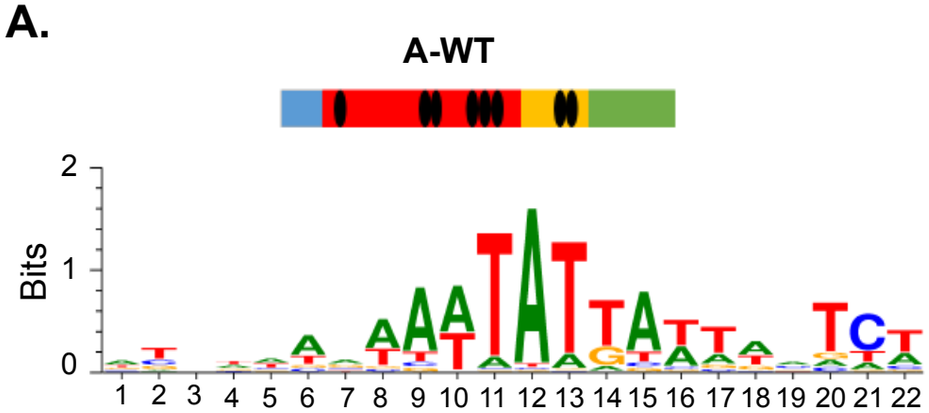
Figure 4. SALL4 KD led to cell death and decreased heterochromatin. **A)** SALL4 levels at 40hr, 72hr, and 120hr after KD, with beta-Actin protein levels as loading control, molecular weight marker is indicated on the right. **B)** Percentages of DAPI+ dead cells that were also successfully transduced (GFP+) at 40hr, 72hr, or 120hr after SALL4 KD. Experiments were performed in biological duplicates. **C)** Western blotting of lysates at 40hr after SALL4 KD, gel cut for clarity, molecular weight marker is indicated on the right. **D)** Immunofluorescence staining of HP1 of SNU398 liver cancer cells transduced with SCR control, or two shRNAs against SALL4; white scale bar denotes 33μm.

Figure 5. SALL4 KD resulted in more open chromatin regions near gene promoters in cancer cells. **A)** ATAC-seq results showing intervals that are unique to SCR control or shSALL4-2 transduced SNU398 liver cancer cells 72 hours after KD. **B)** Unique intervals found in shSALL4-2 tranduced cells showed on average wider intervals. **C)** The genomic distribution of ATAC-seq peaks in SCR control or shSALL4-2 cells.

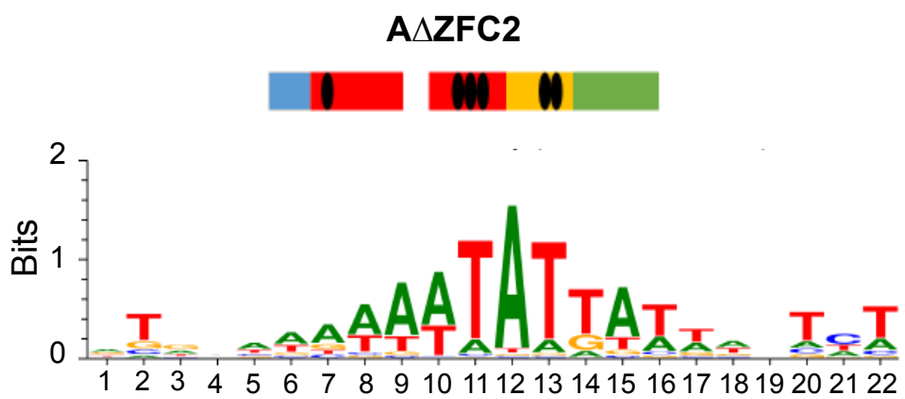
Figure 6. Model of SALL4 regulation of global chromatin landscape. Top panel: KDM3A and KDM4C function by demethylating H3K9me2/3 and thus generating open chromatin at select regions. Bottom panel: SALL4 binds to the promoters of *KDM3A* and *KDM4C*, recruiting the HDAC-containing NuRD complex, and thus repressing their transcription and indirectly maintaining the heterochromatin state of cells

Figure 1

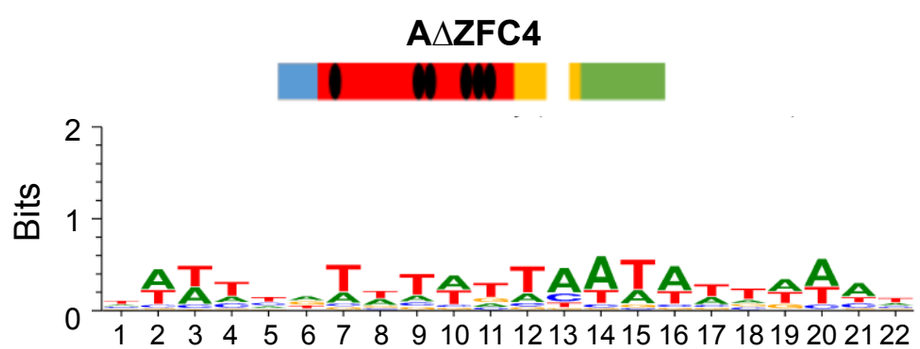
A.



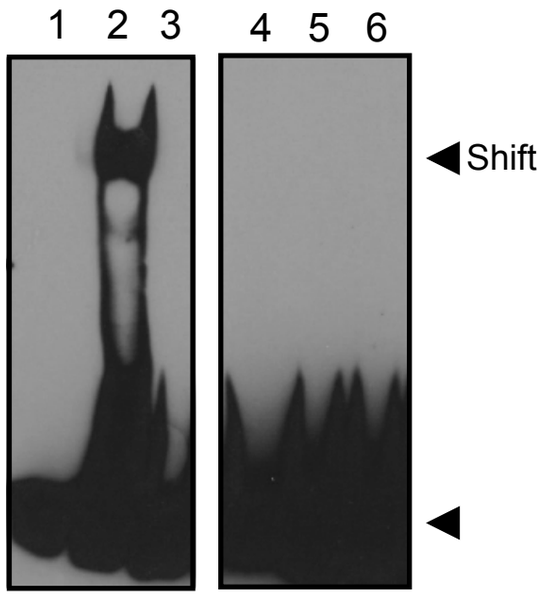
B.



C.

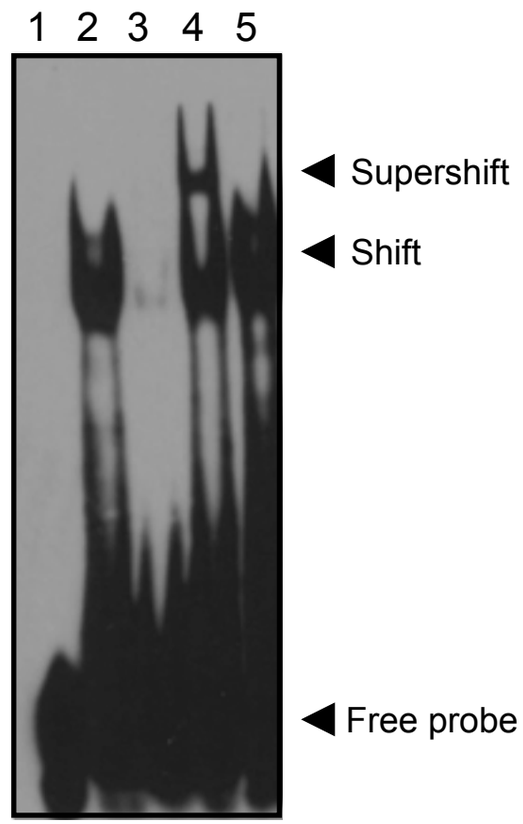


D.



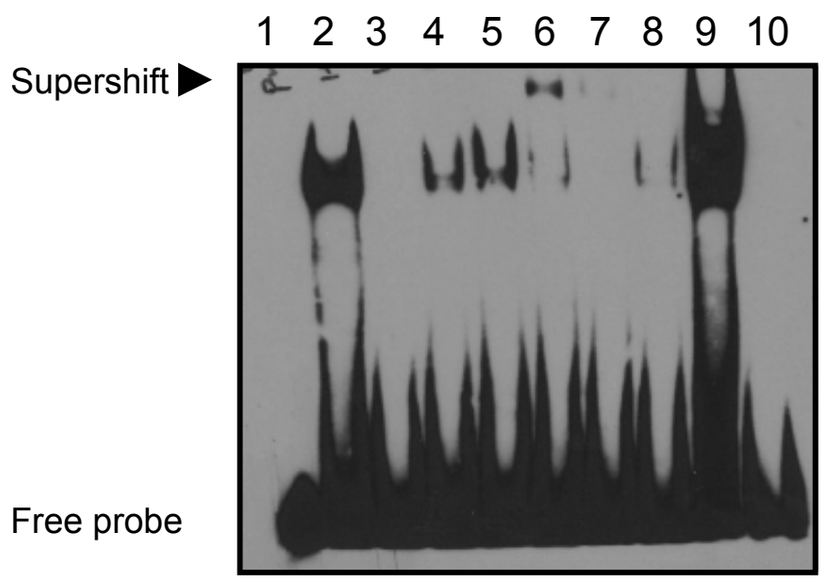
1: Free WT PBM probe
2: WT probe + SALL4A WT
3: 2+ unlabeled competitor
4: Free Mutant PBM probe
5: Mutant probe + SALL4A WT
6: 5+ unlabeled competitor

E.



1: Free PBM WT probe 1
2: WT probe + SALL4A WT
3: 2+ unlabeled competitor
4: 2+ SALL4 antibody (mouse)
5: 2+ mouse isotype control IgG

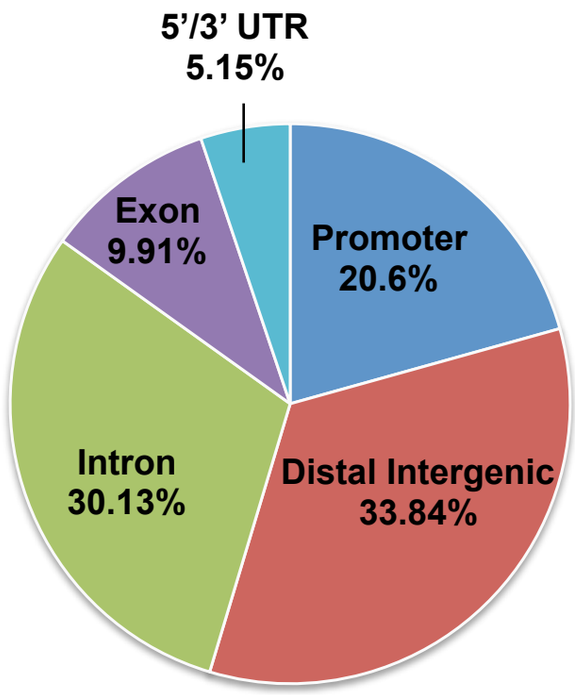
F.



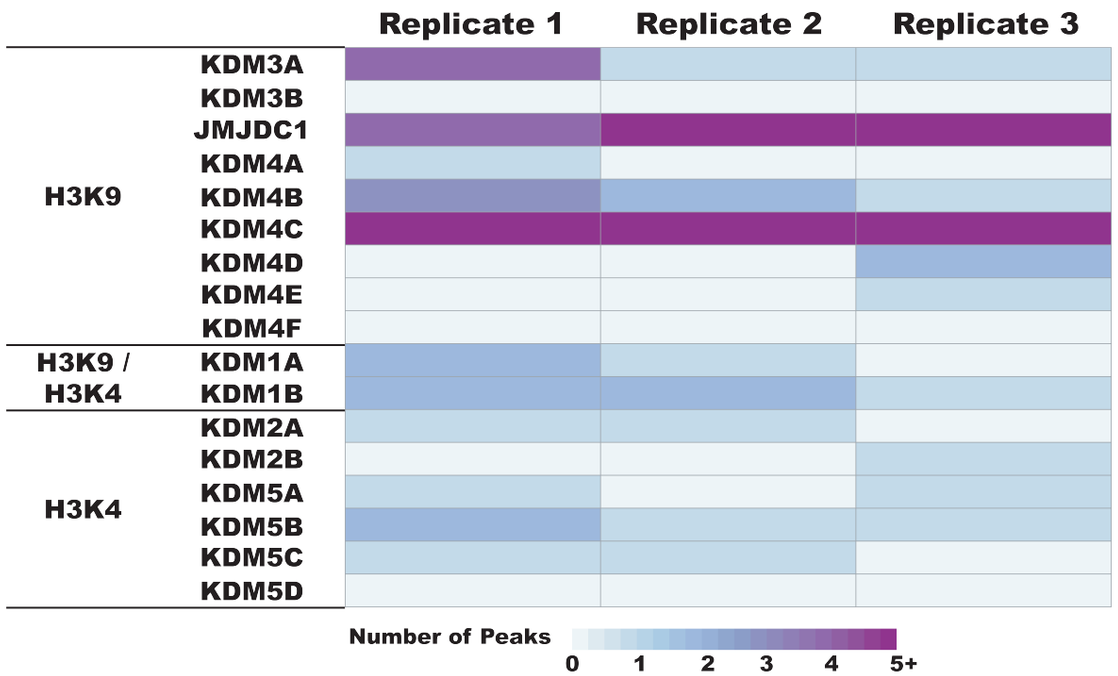
1: Free WT PBM probe
2: WT probe + SALL4A WT
3: 2+ unlabeled competitor
4: 2+ FLAG antibody
5: 2+ Cell Signaling SALL4 antibody
6: 2+ Santa Cruz SALL4 antibody
7: 2+ Abcam SALL4 antibody
8: probe + SALL4AΔZFC2
9: probe + SALL4AΔZFC3
10: probe + SALL4AΔZFC4

Figure 2

A.



C.



B.

Motifs from SALL4 CUT&RUN data

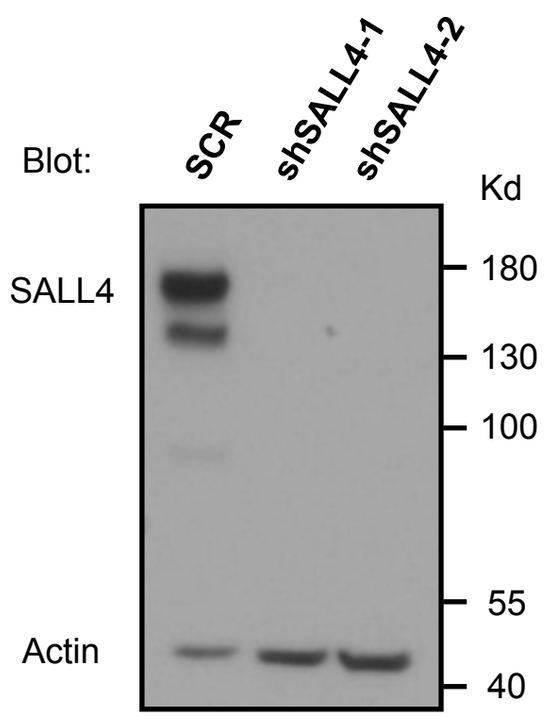
p-Value	% of peaks
1E-691	24.54%
1E-563	40.5%
1E-505	38.65%
1E-481	17.8%

D.

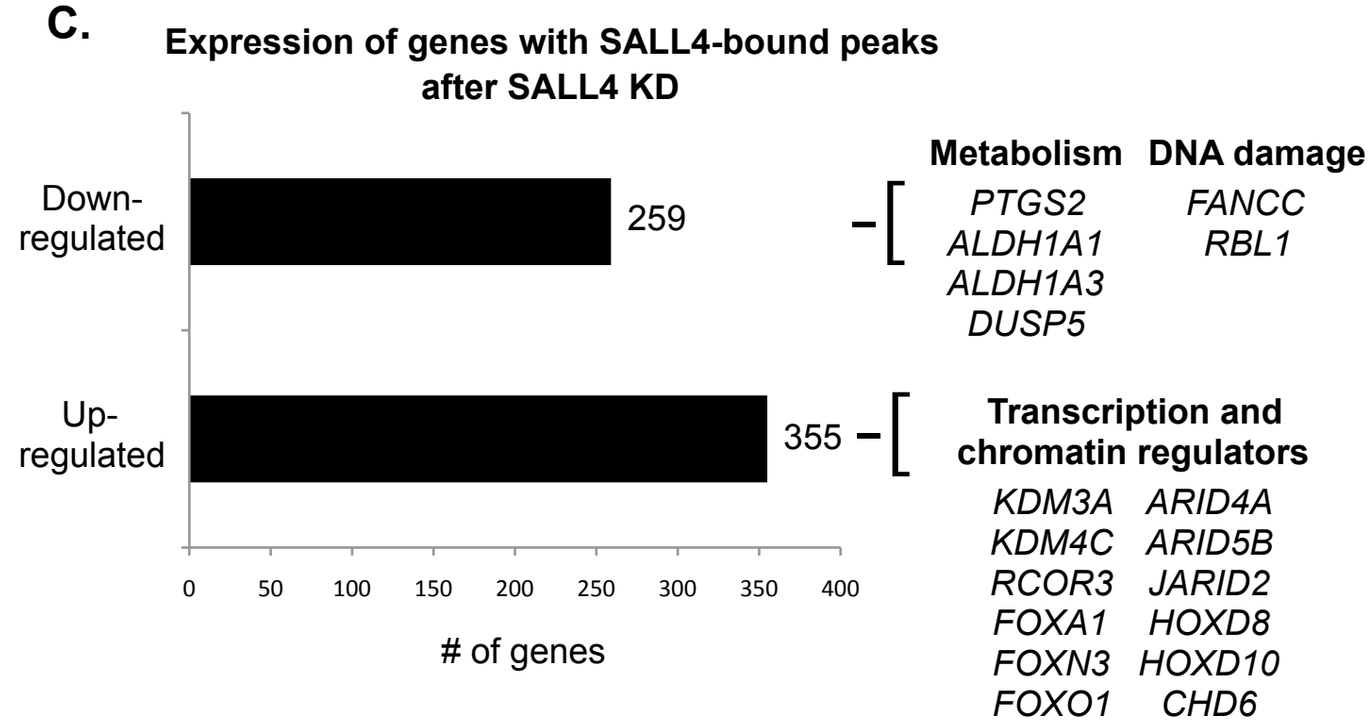
1: Free WT *KDM3A* probe
2: WT probe + SALL4A WT
3: 2+ unlabeled competitor
4: Free mutant *KDM3A* probe
5: Mutant probe + SALL4A WT
6: 5 + unlabeled competitor

Figure 3

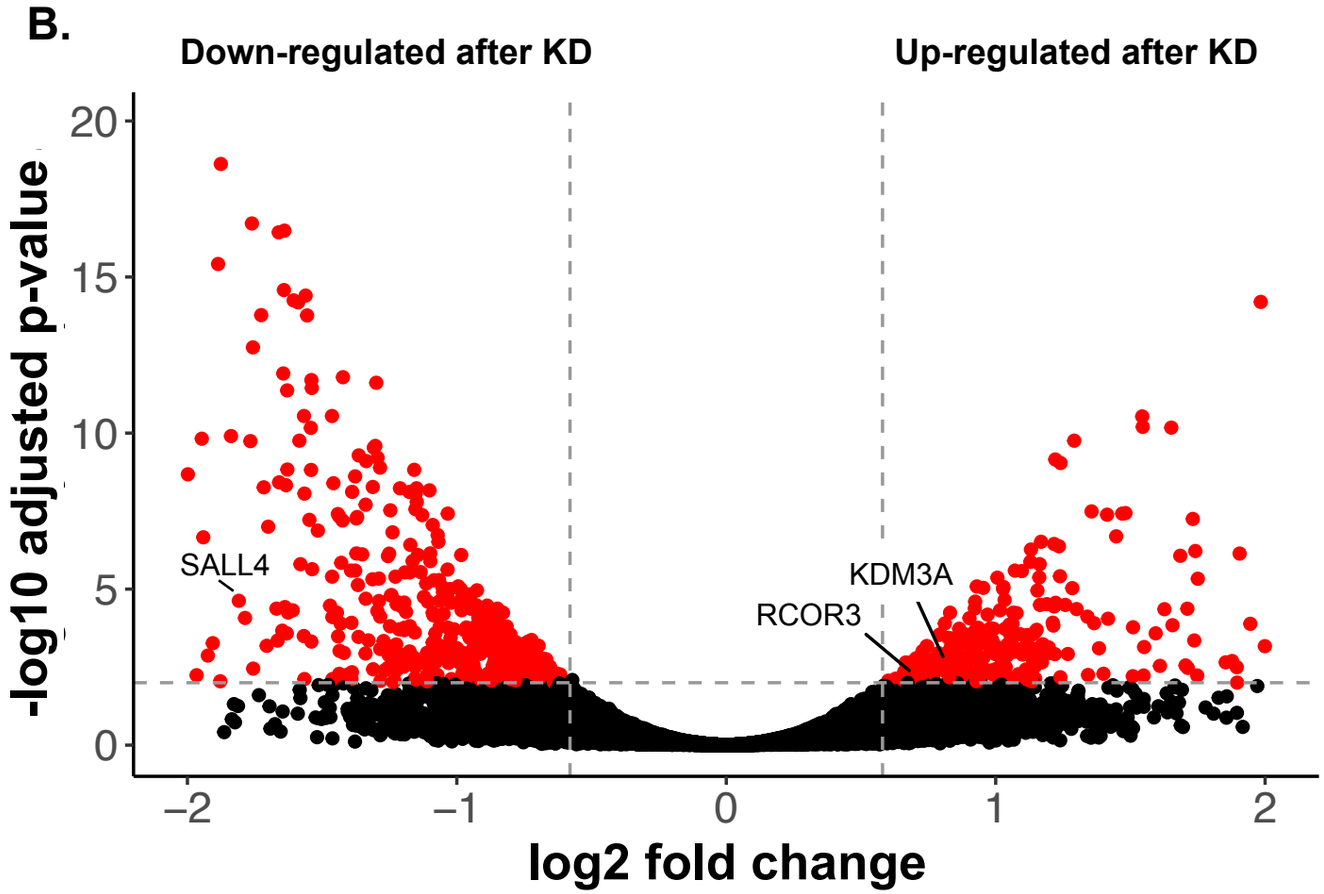
A.



C.



B.



D.

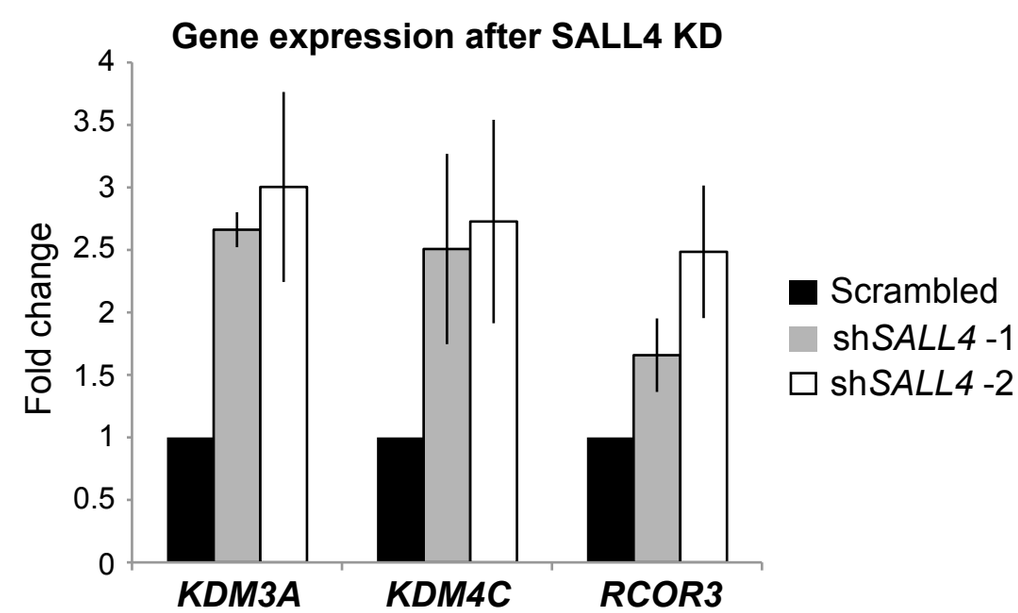


Figure 4

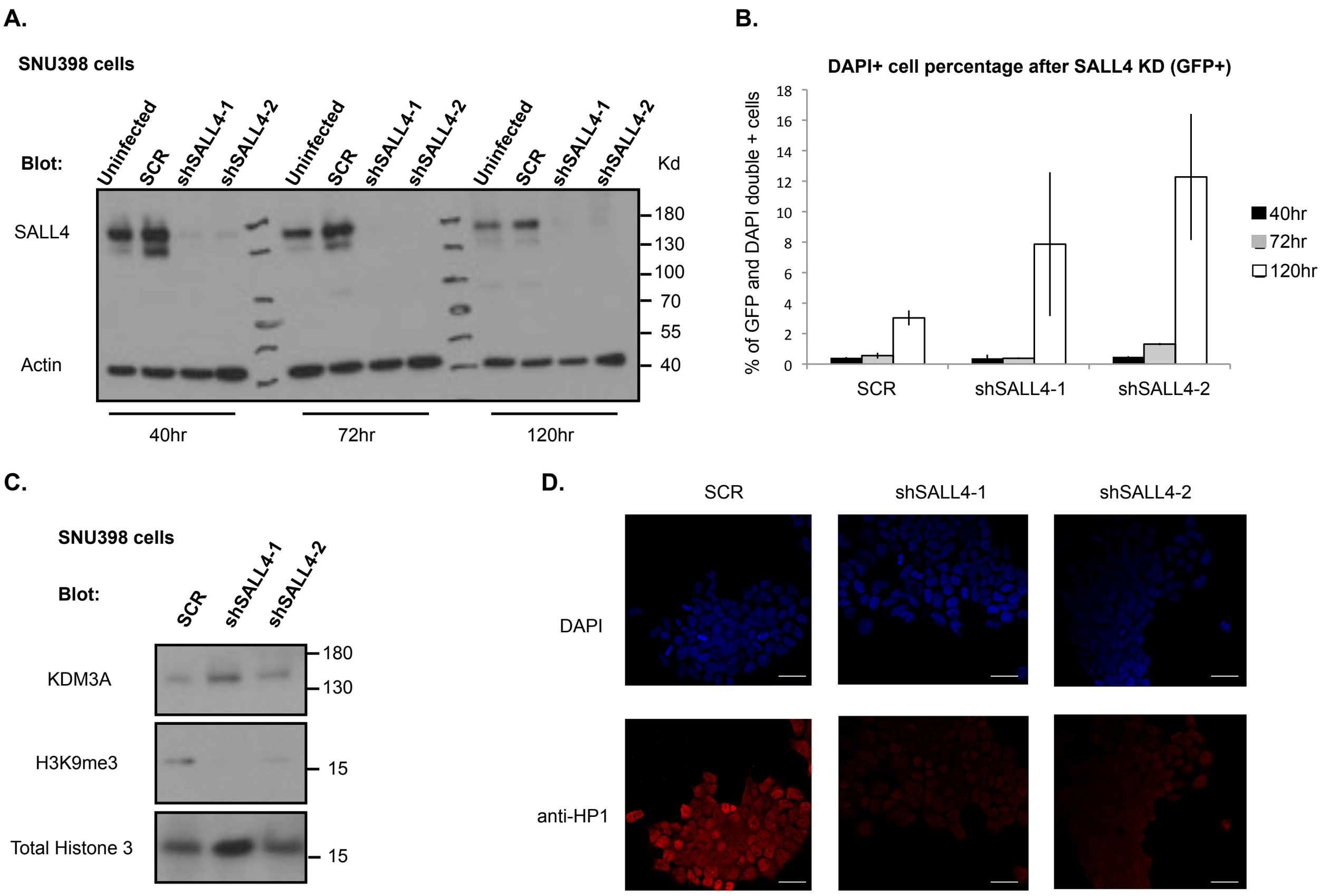
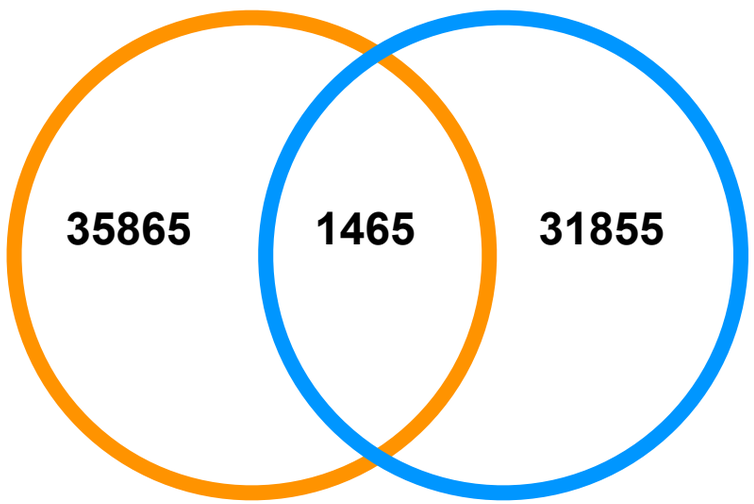

A.
SNU398 cells
Blot:
Uninfected
SCR
shSALL4-1
shSALL4-2
Uninfected
SCR
shSALL4-1
shSALL4-2
Uninfected
SCR
shSALL4-1
shSALL4-2
Kd
SALL4
Actin
180
130
100
70
55
40
40hr
72hr
120hr
B.
DAPI+ cell percentage after SALL4 KD (GFP+)
% of GFP and DAPI double + cells
18
16
14
12
10
8
6
4
2
0
SCR
shSALL4-1
shSALL4-2
40hr
72hr
120hr
C.
SNU398 cells
Blot:
SCR
shSALL4-1
shSALL4-2
KDM3A
H3K9me3
Total Histone 3
180
130
15
15
D.
SCR
shSALL4-1
shSALL4-2
DAPI
anti-HP1

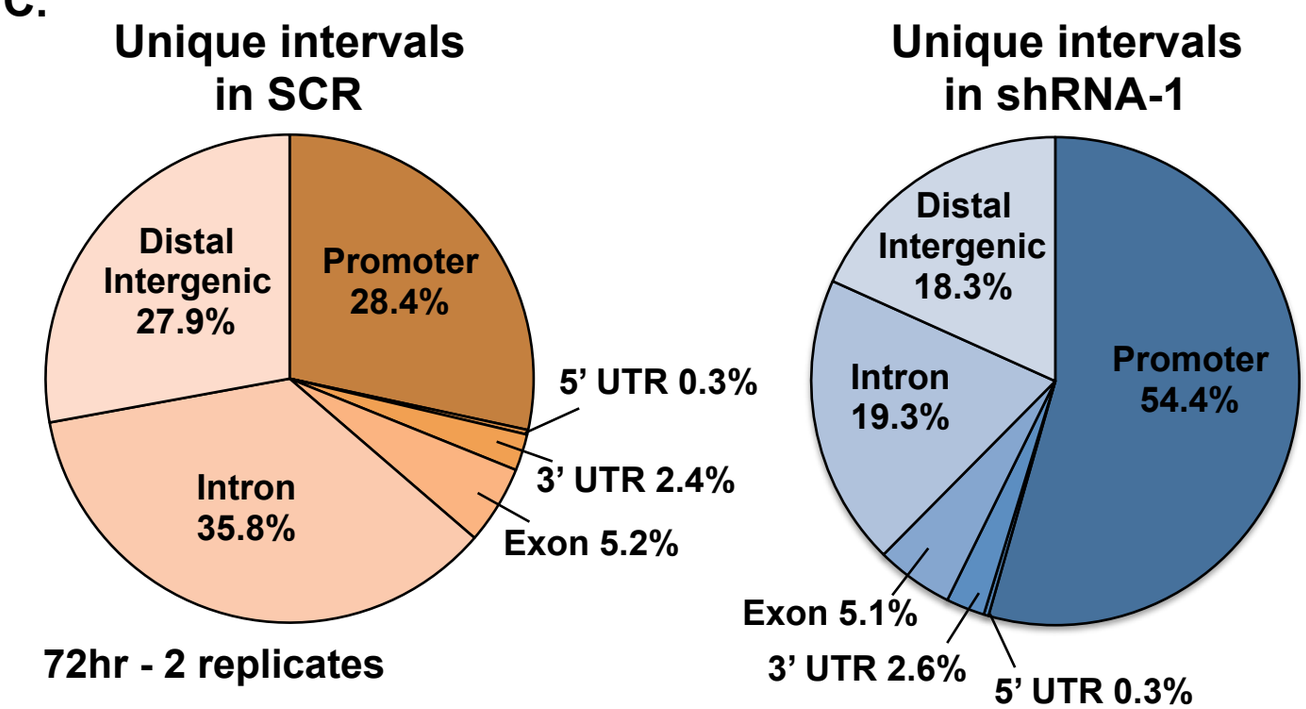
Figure 5

A.



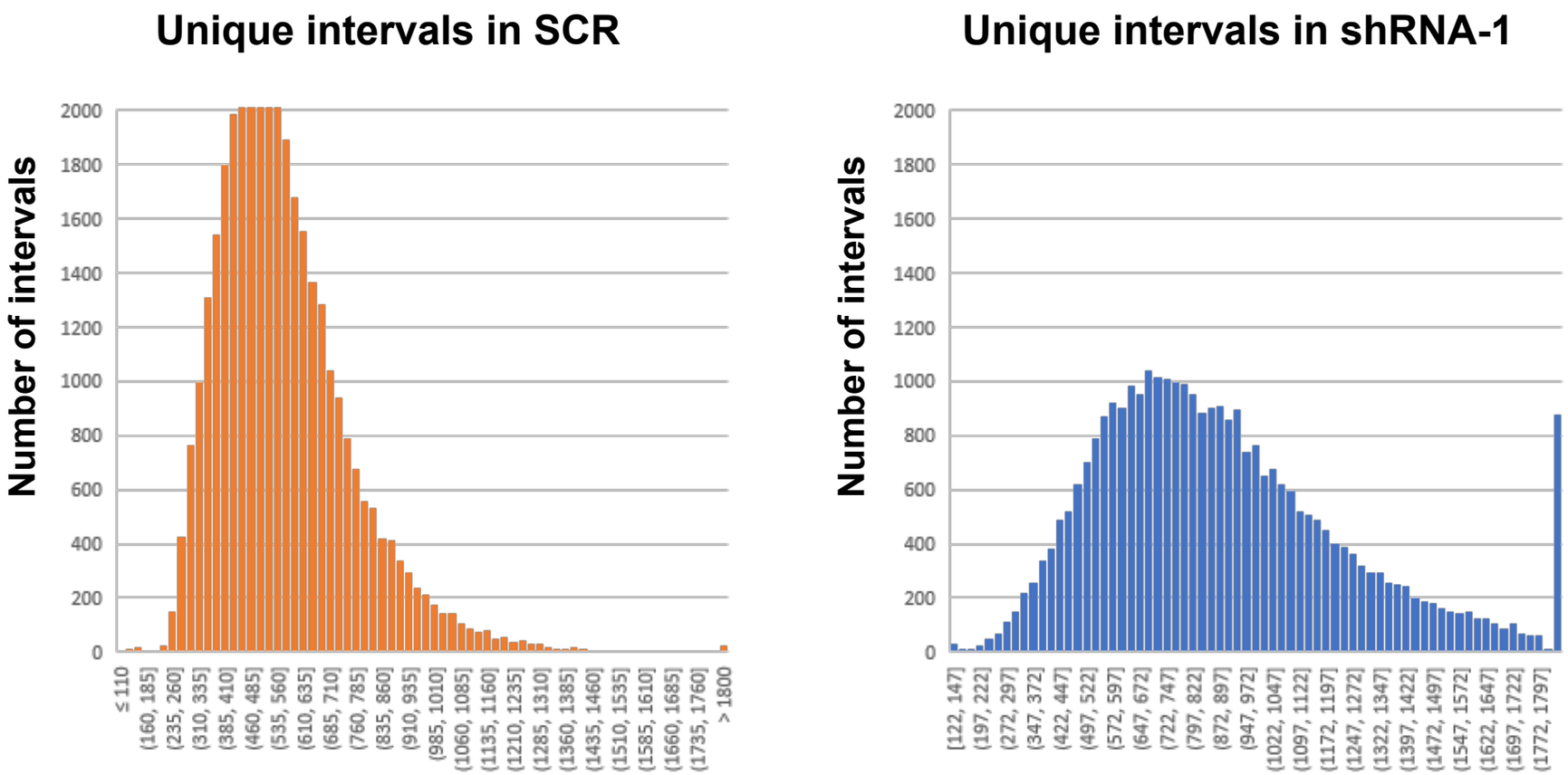
72hr - 2 replicates

C.



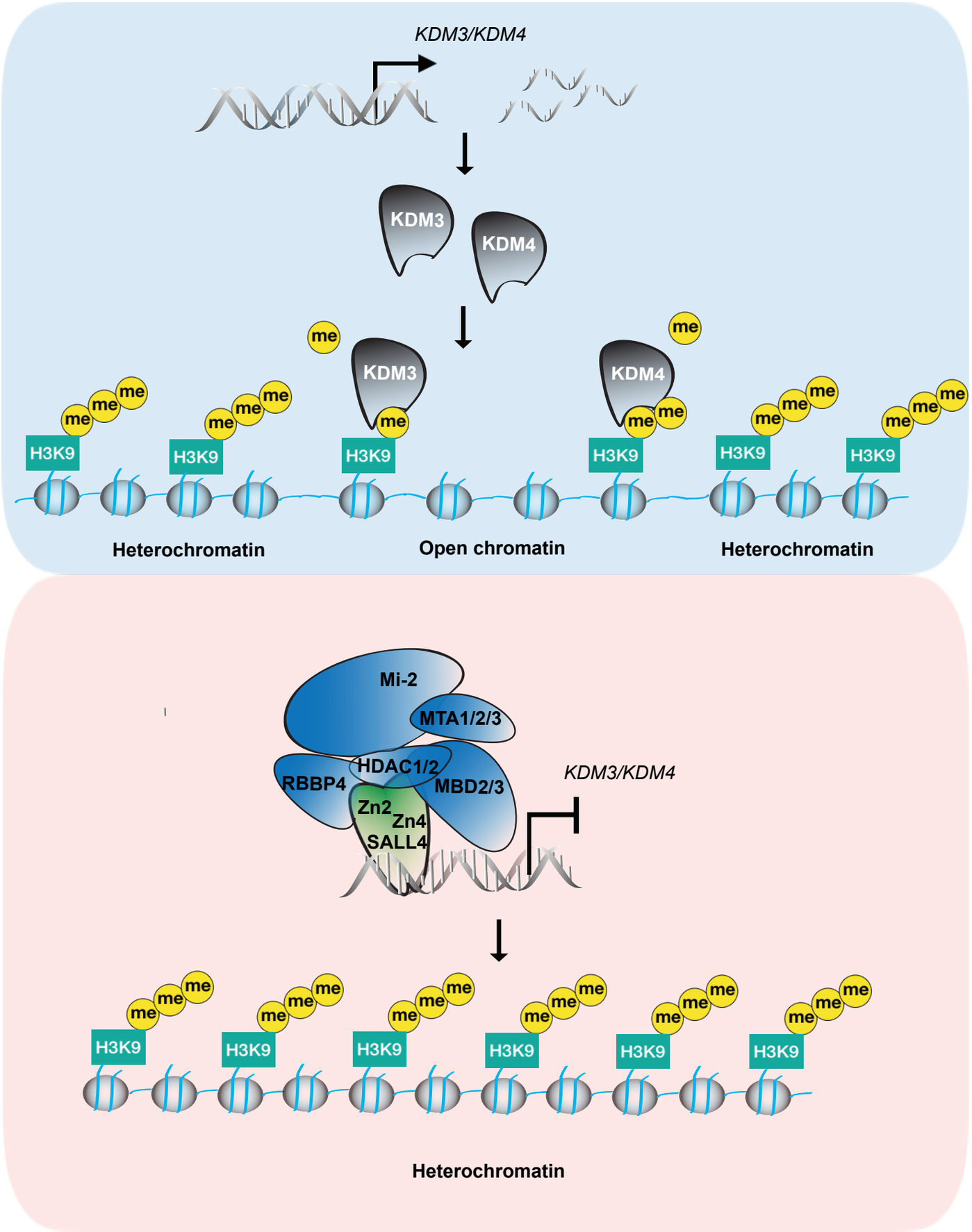
72hr - 2 replicates

B.



72hr - 2 replicates

Figure 6


KDM3/KDM4
KDM3
KDM4
me
H3K9
Heterochromatin
Open chromatin
Heterochromatin
Mi-2
MTA1/2/3
HDAC1/2
RBBP4
MBD2/3
Zn2
Zn4
SALL4
KDM3/KDM4
Heterochromatin