Zinc finger protein SALL4 functions through an AT-rich motif to regulate gene expression

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

Summary

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 of protein binding microarray (PBM) 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 new SALL4 target genes that it directly regulates in aggressive liver cancer cells, including genes encoding a family of Histone 3 Lysine 9-specific Demethylases (KDMs). Taken together, these results elucidated the mechanism of SALL4 DNA binding and revealed novel pathways and molecules to target in SALL4-dependent tumors.

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

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 (Al-Baradie et al., 2002; Kohlhase et al., 1996). *SALL4* knockout mice die at embryonic E6.5 shortly after implantation (Sakaki-Yumoto et al., 2006) and the protein plays an important role in maintaining the stemness of murine embryonic stem cells (ESCs) (Elling et al., 2006; Sakaki-Yumoto et al., 2006; Zhang et al., 2006). Along with Nanog, Lin28, and Esrrb, Sall4 can promote the reprogramming of mouse embryonic fibroblasts to pluripotent stem cells (Buganim et al., 2014). Furthermore, Sall4 interacts with Oct4 and Nanog directly to regulate the expression of the ESC-specific gene signature (Lim et al., 2008; Wu et al., 2006; Zhang et al., 2006). 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 (Al-Baradie et al., 2002; Borozdin et al., 2004; Kohlhase et al., 2003). Recently, SALL4 has been reported to be a neo-substrate of Cereblon E3 ligase complex in thalidomide-mediated teratogenicity (Donovan et al., 2018; Matyskiela et al., 2018). Normally, SALL4 is down-regulated in most adult tissues except germ cells (Chan et al., 2017; Hobbs et al., 2012; Yamaguchi et al., 2015) and hematopoietic stem cells (Gao et al., 2013b). However, it is dysregulated in hematopoietic pre-leukemias and leukemias, including high-grade myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) (Gao et al., 2013a; Ma et al., 2006). SALL4 is also reactivated in a significant fraction of almost all solid tumors, including lung cancer, endometrial cancer, germ cell tumors, and hepatocellular carcinomas (Li et al., 2013a; Miettinen et al., 2014; Yong et al., 2013; 2016). 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 (Oikawa et al., 2013; Yong et al., 2013).

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 (Al-Baradie et al., 2002). 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 (Borozdin et al., 2004; Kohlhase et al., 2003; Miertus et al., 2006; Terhal et al., 2006), 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 (Donovan et al., 2018; Matyskiela et al., 2018). 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) (Gao et al., 2013a; Lu et al., 2009; Yong et al., 2013), many of its target genes and downstream pathways have yet to be elucidated. Its association with the NuRD complex has led to the hypothesis that SALL4 may play a role in global chromatin regulation. However, the evidence for its direct involvement with heterochromatin or euchromatin has been unclear (Böhm et al., 2007; Kim et al., 2017; Sathyan et al., 2011).

Here, we have used an unbiased screen to discover that SALL4 binds AT-rich motif through its C-terminal ZFC. These result were further confirmed by a novel method of targeted in-situ genome wide profiling of SALL4 binding sites in liver cancer cells, which allowed us to discover SALL4's transcriptional regulation of a family of Hisone 3 lysine 9-specific demethylases (KDM3/4), through which it can regulate the chromatin landscape in 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 DNA through an AT-rich motif

In order to identify the SALL4 consensus binding sequence(s), we used the universal protein binding microarray (PBM) technology (Berger et al., 2006) 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 (Struhl, 1989), 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 2A, 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 ZFC4 domain of SALL4 with either WT probes containing the AT-rich motif or mutated probes with only the core motif changed (Supplementary Table 1). ITC experiments showed that while SALL4's C-terminal ZFC4 domain can bind WT probes with Kd of 6.4μM for probe 1 and 6.88μM for probe2, it cannot bind mutated probes (Figure 2B, Supplementary Figure 3). 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. Previously, we have performed SALL4 ChIP-seq experiments in human cells and found they were challenging because SALL4 is located in the chromatin fraction which can be difficult to sonicate (Supplementary Figure 4A). Further, our previous method of interrogating the chromatin fraction still required cross-linking, while it did not generate many SALL4 peaks or yield a SALL4 motif (Liu et al., 2018a). 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 (Skene et al., 2018), which is an in-situ profiling of protein/DNA binding that eliminates the cross-linking step while generates 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 ChIP data are available in these cells (Liu et al., 2018a); ii) SALL4 is required for the viability of a large fraction of hepatocellular carcinoma cells (Yong et al., 2013), and iii) SALL4 serves as a biomarker for worse prognosis in liver cancer (Oikawa et al., 2013; Yong et al., 2013).

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 (representative track shown in Figure 3A, Supplementary Table 2). Furthermore, SALL4 peaks were distributed 33.8% intergenic,

30% intronic, and 21% within the proximal promoter (Figure 3B) 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 was independently identified from PBM experiments (Figure 3C and Supplementary Figure 4C). This motif was present in similar number of peaks in all three SALL4 CUT&RUN replicates through both *de no* and direct searches using Motif 2, which is a shortened Motif 1 (Figure 3D).

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 (Gao et al., 2013a; Li et al., 2013b; Lu et al., 2009; Ma et al., 2006; Young et al., 2014; Zhang et al., 2006). 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 (Supplementary Figure 5A shows a representative western blot showing SALL4 expression after KD with two different shRNAs, Supplementary Table 3). RNA-seq was performed with shRNA-2. We then compared the RNA-seq results to our CUT&RUN peaks and found that among 2,696 significantly differentially expressed genes (red dots in Figure 5A, Supplementary Table 4), 430 genes had SALL4 peaks (totaling 1,192 peaks, Figure 5B, Supplementary Table 5), suggesting they were directly regulated by SALL4. When *de novo* motif search was performed on the peaks, the core WTATB motif was among the top hits (Figure 5C).

For genes with SALL4 CUT&RUN peaks and are differentially expressed, 240 were up-regulated (repressed by SALL4) and 190 were down-regulated (activated by SALL4) after KD. When gene ontology analysis was performed, we found that the most represented molecular pathway that was repressed by SALL4 was transcription regulation, accounting for 11.1% of the up-regulated genes after SALL4 KD, compared to 3.1% of down-regulated genes after SALL4 KD (Figure 5D). This categories of SALL4-repressed genes included chromatin modifiers such as *KDM3A* and several family of transcription factors such as Forkhead family (*FOXA1, FOXO1)*, BCL family (*BCL11A,* BCL11B, and *BCL6*), KLF family (*KLF10, KLF12*), as well as *TBX5*. Many of these targets were not previously reported in liver cancer cells(Liu et al., 2018a), which demonstrates the sensitivity of CUT&RUN technique. Taken together, these genome-wide binding assays have confirmed the bona fide SALL4 motif we found *in vitro*.

SALL4 regulates the expression of histone demethylases

Since SALL4 can interact with the HDAC-containing nucleosome remodeling and deacetylase (NuRD) complex (Lu et al., 2009) (Supplementary Figure 5B), 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 (Gao et al., 2013a; Liu et al., 2018a), we focused on the up-regulated genes after SALL4 KD. Furthermore, SALL4 has been shown to localize in and regulate chromatin in cells (Böhm et al., 2007; Chan et al., 2017; Hobbs et al., 2012; Kim et al., 2017; Xiong et al., 2016), thus our observation that it directly represses KDM3A, a Histone 3 Lysine 9 (H3K9)-specific Demethlyase (Gray et al., 2005), is of great interest. When we categorized all genes encoding lysine demethylases that had SALL4 CUT&RUN peaks, those that

specifically encode demethylases of di- or tri-methylated H3K9 were particularly enriched, compared to those encoding demethylases of tri-methylated H3K4 (Figure 4A).

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 (Liu et al., 2018a). 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* and *KDM4C* (Cloos et al., 2006; Ponnaluri et al., 2011), which was further confirmed by our CUT&RUN data (Supplementary Figure 6A, B). 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 6C and Supplementary Table 6). 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 4B, Supplementary Table 1).

When we compared our RNA-seq results after SALL4 KD to the KDM family of genes containing SALL4 binding sites, we found several others to have changed gene expression, albeit were filtered out in our analyses, such as KDM4C. We therefore repeated the KD experiments and used quantitative real-time PCR (qPCR) to validate both *KDM3A* and *KDM4C* to be negatively regulated targets of SALL4 (Figure 4D, Supplementary Table 2). In order to confirm the importance of SALL4 binding on *KDM3A* gene expression, we further used CRISPR/Cas9 to disrupt one of the SALL4 motifs in its binding site on the *KDM3A* promoter (targeted region 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 6D).

Our combined analyses of RNA-seq and CUT&RUN data, in addition to EMSA and gene editing experiments, 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 and heterochromatin

It has been shown previously that SALL4 KD in liver cancer cells led to cell death (Yong et al., 2013) and we were able to confirm that cells could not survive after prolonged loss of SALL4 expression via KD by two different shRNAs (Supplementary Figure 7). However, at an earlier time point (40 hours post-viral transduction), SALL4 KD was highly efficient (Figure 7A, 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 7B), allowing better assessment of true SALL4-dependent cellular functions and fewer potential secondary effects.

Because of SALL4's known association with the NuRD complex, the HP-1 heterochromatin pathway (Kim et al., 2017), and its regulation of *KDM3/4* pathway (Figure 4), 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 6A). 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 6A).

While shRNA-mediated KD of SALL4 expression is highly efficient at 40hr, this was still a long time point, during which secondary gene expression could occur. It has been shown that SALL4 is a neo-subtrate of E3 ligase CRBN, the cellular degradation of which can be induced via treatment with immunomodulatory drugs such as thalidomide (Donovan et al., 2018; Matyskiela et al., 2018). Therefore, we treated SNU398 liver cancer cells with a thalidomide analog, pomalidomide, for 6, 12, or 24hrs and collected protein lysates. We observed robust SALL4 degradation as soon as 6-12hr after treatment, at which points H3K9me2/3 marks were already diminished significantly, while total histone levels remained unchanged (Figure 6B), suggesting that SALL4's regulation of KDM proteins is direct and has fast kinetics. We further confirmed SALL4's importance in heterochromatin regulation by performing immunofluoresence for HP1α protein after SALL4 KD. HP1 binds H3K9me2/3 and is a hallmark of heterochromatin (Zeng et al., 2010). 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 6C) along with reduction in SALL4 staining (Supplementary Figure 8).

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. Therefore, SALL4 is part of a transcriptional regulatory network in liver cancer cells, where it is required for both cell survival and maintenance of 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.

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 (Xiong et al., 2016). 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. In addition, through biochemical approaches we showed that the 4th ZFC of SALL4 is responsible for DNA recognition and strong binding. Further, our novel motif was supported by CUT&RUN assays, which resulted in much smaller DNA fragments than previous 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. While this publication was in preparation, we learned that Dr. Adrian Bird's laboratory independently uncovered an AT-rich motif for murine SALL4 through a DNA pulldown screen followed by validation by SELEX (biorXiv doi: 10.1101/2020.06.30.179481). Further, through CRISPR deletion experiment in mouse ESCs, they independently found that the 4th ZFC of SALL4 is important for DNA binding. These results orthogonally confirmed our mechanism of SALL4 DNA binding in human cancer cells.

Our novel CUT&RUN 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 and decreased heterchromatin marks presents a novel mechanism by which SALL4 acts as a 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 (Okada et al., 2007). 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 (Hobbs et al., 2012). 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 (Kuroki et al., 2018). 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 (Pedanou et al., 2016).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.

STAR Methods

Lead contact and materials availability

Further information and requests for resources such as recombinant DNA plasmids generated in this study should be directed to and will be fulfilled by the Lead Contact, Daniel G. Tenen (Daniel.tenen@nus.edu.sg)

Experimental model and subject details

SNU398 and SNU387 hepatocellular carcinoma cells (ATCC) were cultured in RPMI media with 10% fetal bovine serum (FBS, Gibco). HEK293T and HeLa cells were cultured in DMEM media with 10% FBS.

Method Details

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 (Berger and Bulyk, 2009; Berger et al., 2006). 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).

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 (ITC)

The DNA fragments encoding SALL4 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).

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) (Skene et al., 2018). 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.) (Schmid et al., 2004) 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) (Liu et al., 2018b). Pair-end (42bp) Illumina sequencing was performed on the bar-coded and amplified libraries.

Lentivirus-mediated gene expression knockdown and western blotting

Two shRNAs targeting SALL4 were previously described (Gao et al., 2013b)**:** E5 and 507 (Supplementary Table 3); 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. 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 in three biological replicates 72 hours after SALL4 KD and libraries were made following manufacturer's instructions (Illumina). Pair-end Illumina sequencing was performed on the bar-coded and amplified libraries.

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 6hr 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.

Quantification and statistical analysis

PBM analysis

PBM 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) (Berger et al., 2006; Dudley et al., 2002) 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(Berger and Bulyk, 2009; Struhl, 1989). Motifs were derived using the Seed-and-Wobble algorithm, and Enologos was used to generate logos from PWMs, as previously described (Berger and Bulyk, 2009; Berger et al., 2006).

ITC analysis

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.

CUT&RUN analysis

Detailed data analysis combining Henikoff (Skene et al., 2018) and Orkin (Liu et al., 2018b) labs' pipelines can be found on github (https://github.com/mbassalbioinformatics/CnRAP). Briefly, raw fastq files were trimmed with Trimmomatic v0.36 (Bolger et al., 2014) in pair-end mode. Next, the kseq trimmer developed by the Orkin lab was run on each fastq file. BWA (v0.7.17-r1188) (Li and Durbin, 2009) 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) (Lunter and Goodson, 2011) was in "—sensitive" mode. Next, using SAMtools (v1.5)(Li et al., 2009), 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) (Ramírez et al., 2014). 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 tool from UCSC (v4) (Kent et al., 2010). For peak calling, the recently developed SEACR (v.1.1) (Meers et al., 2019)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) . Peaks were annotate using the R package ChIPSeeker (v.1.20.0)(Yu et al., 2015). Overlapping peak subsets within 3kb of each other were generated using mergePeaks.py from the HOMER suite (Heinz et al., 2010). 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.

RNA-seq analysis

Raw fastq files had optical duplicates removed using clumpify form BBMap (https://sourceforge.net/projects/bbmap/). Next, adapt trimming was performed using BBDuk (from BBMap) and reads were trimmed using trimmomatic (Bolger et al., 2014). After read cleanup, reads were aligned to hg38 genome using STAR (Dobin et al., 2013). BamCoverage (Ramírez et al., 2014) maps were generated using default parameters and read distributions were calculated using read_distribution.py from the RseQC suite of tools (Wang et al., 2012). Counts tables were generated using htseq-count (Anders et al., 2015). The fold change was plotted as a volcano plot using EnhancedVolcano (https://github.com/kevinblighe/EnhancedVolcano) with FDR cut-off of 0.05. Quantitative real-time PCR for selected targets were performed with primer sequences found in Supplementary Table 6.

Motif search

De novo motif search was performed using both HOMER (Heinz et al., 2010) with the flags "-size given –mask –S 50" and MEME-ChIP (Bailey et al., 2009) with the flags "-drene-m 50 –meme-nmotifs 50". For directed motif search wherein we searched for the abundance of our motif of interest (Motif 2 in Figure 3C) in called CUT&RUN peaks, HOMER was utilized with a calculated motif position weight matrix and the flags "-find pos_weight_matrix.motif".

Data and code availability

The datasets and code utilized in this study are available at GEO (accession GSE136332) and on GitHub at https://github.com/mbassalbioinformatics/CnRAP

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., MA.B., L.C., and D.G.T. prepared the manuscript.

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

Al-Baradie, R., Yamada, K., St Hilaire, C., Chan, W.-M., Andrews, C., McIntosh, N., Nakano, M., Martonyi, E.J., Raymond, W.R., Okumura, S., et al. (2002). 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. *71*, 1195–1199.

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

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc *4*, 393–411.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

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

Borozdin, W., Wright, M.J., Hennekam, R.C.M., Hannibal, M.C., Crow, Y.J., Neumann, T.E., and Kohlhase, J. (2004). 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. *41*, e102.

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

Buganim, Y., Markoulaki, S., van Wietmarschen, N., Hoke, H., Wu, T., Ganz, K., Akhtar-

Zaidi, B., He, Y., Abraham, B.J., Porubsky, D., et al. (2014). The developmental potential of iPSCs is greatly influenced by reprogramming factor selection. Cell Stem Cell *15*, 295–309.

Chan, A.-L., La, H.M., Legrand, J.M.D., Mäkelä, J.-A., Eichenlaub, M., De Seram, M., Ramialison, M., and Hobbs, R.M. (2017). Germline Stem Cell Activity Is Sustained by SALL4-Dependent Silencing of Distinct Tumor Suppressor Genes. Stem Cell Reports *9*, 956–971.

Cloos, P.A.C., Christensen, J., Agger, K., Maiolica, A., Rappsilber, J., Antal, T., Hansen, K.H., and Helin, K. (2006). The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. Nature *442*, 307–311.

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

Donovan, K.A., An, J., Nowak, R.P., Yuan, J.C., Fink, E.C., Berry, B.C., Ebert, B.L., and Fischer, E.S. (2018). Thalidomide promotes degradation of SALL4, a transcription factor implicated in Duane Radial Ray syndrome. Elife *7*.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci USA *99*, 7554–7559.

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

Gao, C., Dimitrov, T., Yong, K.J., Tatetsu, H., Jeong, H.-W., Luo, H.R., Bradner, J.E., Tenen, D.G., and Chai, L. (2013a). Targeting transcription factor SALL4 in acute myeloid leukemia by interrupting its interaction with an epigenetic complex. Blood *121*, 1413–1421.

Gao, C., Kong, N.R., Li, A., Tatetu, H., Ueno, S., Yang, Y., He, J., Yang, J., Ma, Y., Kao, G.S., et al. (2013b). SALL4 is a key transcription regulator in normal human hematopoiesis. Transfusion *53*, 1037–1049.

Gray, S.G., Iglesias, A.H., Lizcano, F., Villanueva, R., Camelo, S., Jingu, H., Teh, B.T., Koibuchi, N., Chin, W.W., Kokkotou, E., et al. (2005). Functional characterization of JMJD2A, a histone deacetylase- and retinoblastoma-binding protein. J Biol Chem *280*, 28507–28518.

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

Hobbs, R.M., Fagoonee, S., Papa, A., Webster, K., Altruda, F., Nishinakamura, R., Chai, L., and Pandolfi, P.P. (2012). Functional antagonism between Sall4 and Plzf defines

germline progenitors. Cell Stem Cell *10*, 284–298.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207.

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

Kohlhase, J., Schubert, L., Liebers, M., Rauch, A., Becker, K., Mohammed, S.N., Newbury-Ecob, R., and Reardon, W. (2003). 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. *40*, 473–478.

Kohlhase, J., Schuh, R., Dowe, G., Kühnlein, R.P., Jäckle, H., Schroeder, B., Schulz-Schaeffer, W., Kretzschmar, H.A., Köhler, A., Muller, U., et al. (1996). Isolation, characterization, and organ-specific expression of two novel human zinc finger genes related to the Drosophila gene spalt. Genomics *38*, 291–298.

Kuroki, S., Nakai, Y., Maeda, R., Okashita, N., Akiyoshi, M., Yamaguchi, Y., Kitano, S., Miyachi, H., Nakato, R., Ichiyanagi, K., et al. (2018). Combined Loss of JMJD1A and JMJD1B Reveals Critical Roles for H3K9 Demethylation in the Maintenance of Embryonic Stem Cells and Early Embryogenesis. Stem Cell Reports *10*, 1340–1354.

Li, A., Jiao, Y., Yong, K.J., Wang, F., Gao, C., Yan, B., Srivastava, S., Lim, G.S.D., Tang, P., Yang, H., et al. (2013a). SALL4 is a new target in endometrial cancer. Oncogene.

Li, A., Yang, Y., Gao, C., Lu, J., Jeong, H.-W., Liu, B.H., Tang, P., Yao, X., Neuberg, D., Huang, G., et al. (2013b). A SALL4/MLL/HOXA9 pathway in murine and human myeloid leukemogenesis. J. Clin. Invest. *123*, 4195–4207.

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

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

Lim, C.Y., Tam, W.-L., Zhang, J., Ang, H.S., Jia, H., Lipovich, L., Ng, H.-H., Wei, C.-L., Sung, W.K., Robson, P., et al. (2008). Sall4 regulates distinct transcription circuitries in different blastocyst-derived stem cell lineages. Cell Stem Cell *3*, 543–554.

Liu, B.H., Jobichen, C., Chia, C.S.B., Chan, T.H.M., Tang, J.P., Chung, T.X.Y., Li, J., Poulsen, A., Hung, A.W., Koh-Stenta, X., et al. (2018a). Targeting cancer addiction for SALL4 by shifting its transcriptome with a pharmacologic peptide. Proceedings of the National Academy of Sciences *115*, E7119–E7128.

Liu, N., Hargreaves, V.V., Zhu, Q., Kurland, J.V., Hong, J., Kim, W., Sher, F., Macias-Trevino, C., Rogers, J.M., Kurita, R., et al. (2018b). Direct Promoter Repression by

BCL11A Controls the Fetal to Adult Hemoglobin Switch. Cell *173*, 430–442.e17.

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

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

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

Matyskiela, M.E., Couto, S., Zheng, X., Lu, G., Hui, J., Stamp, K., Drew, C., Ren, Y., Wang, M., Carpenter, A., et al. (2018). SALL4 mediates teratogenicity as a thalidomide-dependent cereblon substrate. Nat. Chem. Biol. *14*, 981–987.

Meers, M.P., Tenenbaum, D., and Henikoff, S. (2019). Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics Chromatin *12*, 42.

Miertus, J., Borozdin, W., Frecer, V., Tonini, G., Bertok, S., Amoroso, A., Miertus, S., and Kohlhase, J. (2006). 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. *119*, 154–161.

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

Oikawa, T., Kamiya, A., Zeniya, M., Chikada, H., Hyuck, A.D., Yamazaki, Y., Wauthier, E., Tajiri, H., Miller, L.D., Wang, X.W., et al. (2013). Sal-like protein 4 (SALL4), a stem cell biomarker in liver cancers. Hepatology *57*, 1469–1483.

Okada, Y., Scott, G., Ray, M.K., Mishina, Y., and Zhang, Y. (2007). Histone demethylase JHDM2A is critical for Tnp1 and Prm1 transcription and spermatogenesis. Nature *450*, 119–123.

Pedanou, V.E., Gobeil, S., Tabariès, S., Simone, T.M., Zhu, L.J., Siegel, P.M., and Green, M.R. (2016). The histone H3K9 demethylase KDM3A promotes anoikis by transcriptionally activating pro-apoptotic genes BNIP3 and BNIP3L. Elife *5*.

Ponnaluri, V.K.C., Vavilala, D.T., and Mukherji, M. (2011). Studies on substrate specificity of Jmjd2a-c histone demethylases. Biochemical and Biophysical Research Communications *405*, 588–592.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Sakaki-Yumoto, M., Kobayashi, C., Sato, A., Fujimura, S., Matsumoto, Y., Takasato, M.,

Kodama, T., Aburatani, H., Asashima, M., Yoshida, N., et al. (2006). 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 *133*, 3005–3013.

Sathyan, K.M., Shen, Z., Tripathi, V., Prasanth, K.V., and Prasanth, S.G. (2011). A BEN-domain-containing protein associates with heterochromatin and represses transcription. J Cell Sci *124*, 3149–3163.

Schmid, M., Durussel, T., and Laemmli, U.K. (2004). ChIC and ChEC; genomic mapping of chromatin proteins. Molecular Cell *16*, 147–157.

Skene, P.J., Henikoff, J.G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat Protoc *13*, 1006–1019.

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

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

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.

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

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

Yamaguchi, Y.L., Tanaka, S.S., Kumagai, M., Fujimoto, Y., Terabayashi, T., Matsui, Y., and Nishinakamura, R. (2015). Sall4 is essential for mouse primordial germ cell specification by suppressing somatic cell program genes. Stem Cells *33*, 289–300.

Yong, K.J., Gao, C., Lim, J.S.J., Yan, B., Yang, H., Dimitrov, T., Kawasaki, A., Ong, C.W., Wong, K.-F., Lee, S., et al. (2013). Oncofetal gene SALL4 in aggressive hepatocellular carcinoma. N. Engl. J. Med. *368*, 2266–2276.

Yong, K.J., Li, A., Ou, W.-B., Hong, C.K.Y., Zhao, W., Wang, F., Tatetsu, H., Yan, B., Qi, L., Fletcher, J.A., et al. (2016). Targeting SALL4 by entinostat in lung cancer. Oncotarget *7*, 75425–75440.

Young, J.J., Kjolby, R.A.S., Kong, N.R., Monica, S.D., and Harland, R.M. (2014). Spalt-like 4 promotes posterior neural fates via repression of pou5f3 family members in Xenopus. Development *141*, 1683–1693.

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

Zeng, W., Ball, A.R., and Yokomori, K. (2010). HP1: heterochromatin binding proteins working the genome. Epigenetics *5*, 287–292.

Zhang, J., Tam, W.-L., Tong, G.Q., Wu, Q., Chan, H.-Y., Soh, B.-S., Lou, Y., Yang, J., Ma, Y., Chai, L., et al. (2006). Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. Nat Cell Biol *8*, 1114–1123.

Figure 1. Discovery of a novel SALL4 DNA binding motif. **A)** Specific DNA sequence motif bound by SALL4A in universal PBM assays (Berger et al., 2006).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); S4A=SALL4A, UC=unlabeled WT-1 probe **E)** SALL4-DNA complex is super-shifted by a SALL4-specific monoclonal antibody (lane 4) but not by mouse IgG isotype control (lane 5); S4 mAb=mouse SALL4 antibdy (Santa Cruz EE-30), mIgG=mouse isotype control IgG. All EMSA reactions contain poly dI:dC competitor to reduce background binding.

Figure 2. SALL4 binds DNA through its 4th zinc finger cluster. **A)** 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); S4A=SALL4A, UC=unlabeled WT-1 probe, S4 rAb-1=SALL4 rabbit antibody (Cell Signaling D16H12), S4 rAb-2=SALL4 rabbit antibody (Abcam ab57577). 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, all EMSA reactions contain poly dI:dC competitor. **B)** Isothermal titration calorimetry (ITC) experiments showing purified SALL4 ZFC4 (amino acids 864-929) binds DNA oligos containing the WTATB motif (left, WT sequence: GAGTTATTAATG) and not when the motif was mutated (right, mutant sequence: GAGTCGCTAATG).

Figure 3. SALL4 CUT&RUN showing its binding genome-wide in liver cancer cells. **A)** Representative genomic tracks of three SALL4 CUT&RUN replicates and their isotype rabbit IgG control experiments. Scale is 0-25 **B)** The genomic distribution of SALL4 CUT&RUN peaks from three biological replicates (rep). **C)** The top five HOMER motifs from *de novo* analysis of three CUT&RUN replicates with their respective p-values, resulting in the two composite motifs shown on right. **D)** Bar graph showing the percentage of peaks containing Motif 2 in *de novo* (black bars) and direct (gray bars) from three SALL4 CUT&RUN replicates.

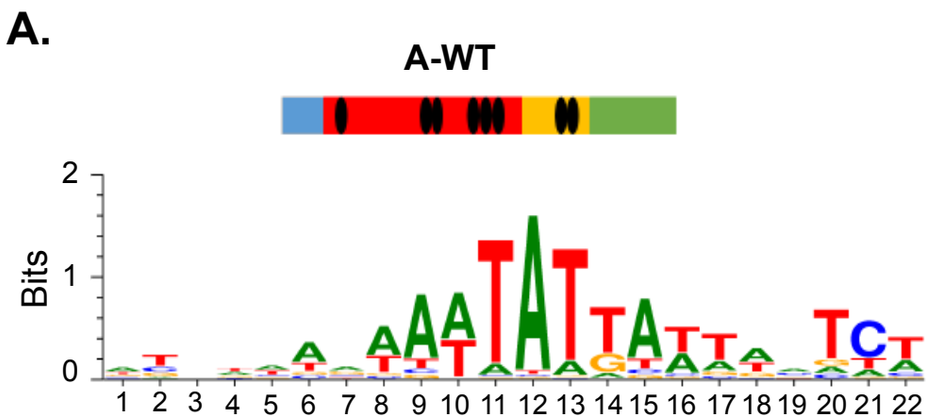
Figure 4. RNA-seq data revealed direct SALL4 target genes. **A)** Volcano plot showing genes that are down- or up-regulated significantly after SALL4 KD with log2 fold change (FC) represented on the x-axis; red circles denote differentially expressed genes (DGEs) with FDR <0.05; experiment performed in biological triplicates with shSALL4-2 compared to scrambled control shRNA (SCR) 72 hours after transduction. **B)** Number of differentially expressed genes after SALL4 KD as well as those containing SALL4 peaks (Supplementary Table 4 and 5). **C)** *De novo* motif search focusing on 1,192 SALL4 CUT&RUN peaks near the differentially expressed genes. **D)** Bar graph representing number of up- and down-regulated genes after SALL4 KD (Supplementary Table 5), and their gene ontology (GO) molecular pathway analysis (www.pantherdb.org) focusing on GO 0140110.

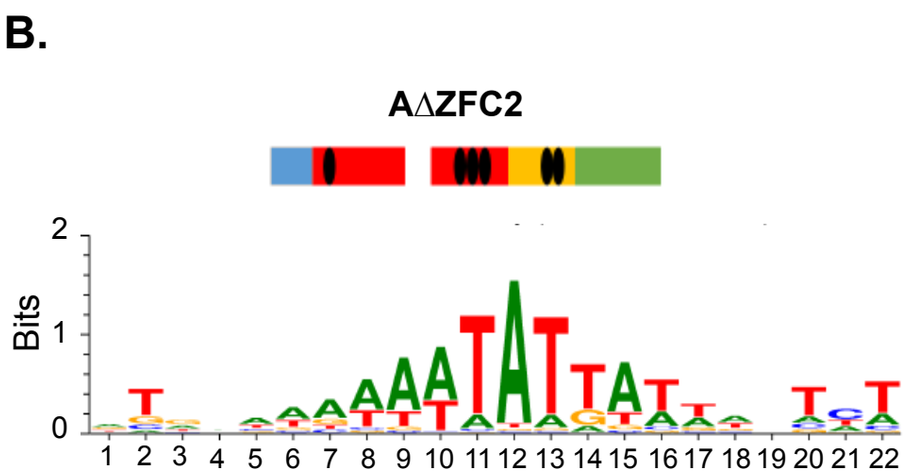
Figure 5. SALL4 regulates histone lysine demethylases. **A)** Heatmap showing SALL4 CUT&RUN peaks that are enriched above IgG isotype control at genes that encode histone lysine demethylases, which were categorized as H3K9-specific, H3K4-specific, and H3K4/9-specific. **B)** EMSA showing that SALL4 binds *KDM3A* promoter region containing the AT-rich motif (WT sequence: TCTTCATTTATCCTTCAAAA), but not the

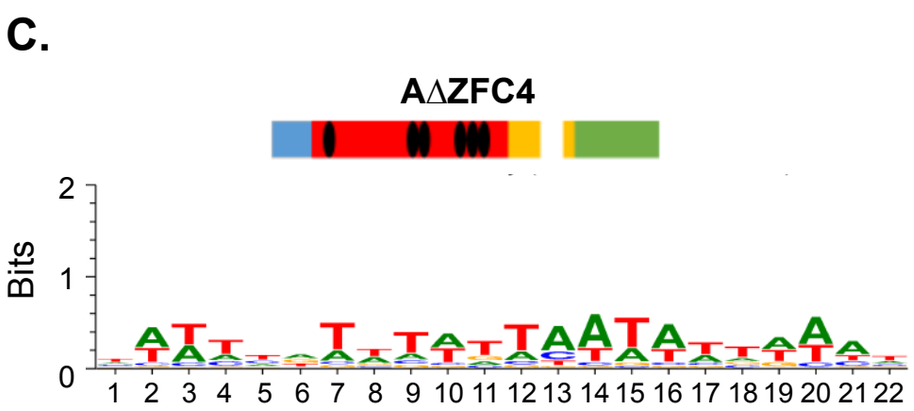
mutated sequence (TCTTCTTTTAACCTTCAAAA). **C)** Volcano plot from Figure 4A with genes encoding KDM proteins labeled. Green circles denote genes containing SALL4 CUT&RUN peaks, the size of the circles corresponds to $\log_2$FC in expression. **D)** Quantitative real-time PCR (qPCR) analysis of three SALL4 direct targets 40 hours after SALL4 KD, results are summarized from either three or four independent experiments (primer sequences found in Supplementary Table 3).

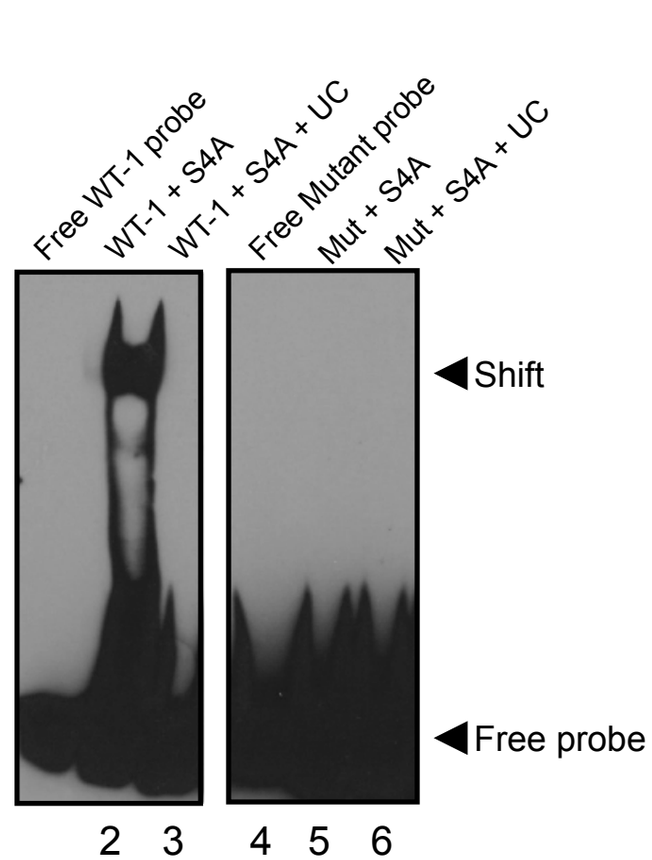
Figure 6. SALL4 and heterochromatin. **A)** Western blotting of SNU398 liver cancer cell lysates at 40hr after SALL4 KD, gel cut for clarity, molecular weight marker is indicated on the right. Bottom: summary of western blot densitometry from two separate SALL4 KD experiments of H3K9me2/3 protein expression (black bars) and KDM3A expression (gray bars). **B)** Western blotting of SNU398 cell lysates collected after either control DMSO or pomalidomide (Pom, 10μM) treatments at the indicated time points. **C)** Immunofluorescence staining of HP1 of SNU398 liver cancer cells transduced with SCR control, or two shRNAs against SALL4; white scale bar denotes 33μm.

Figure 1


A.
A-WT
Bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


B.
AΔZFC2
Bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


C.
AΔZFC4
Bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


D.
Free WT-1 probe
WT-1 + S4A
WT-1 + S4A + UC
Free Mutant probe
Mut + S4A
Mut + S4A + UC
Shift
Free probe
2 3 4 5 6

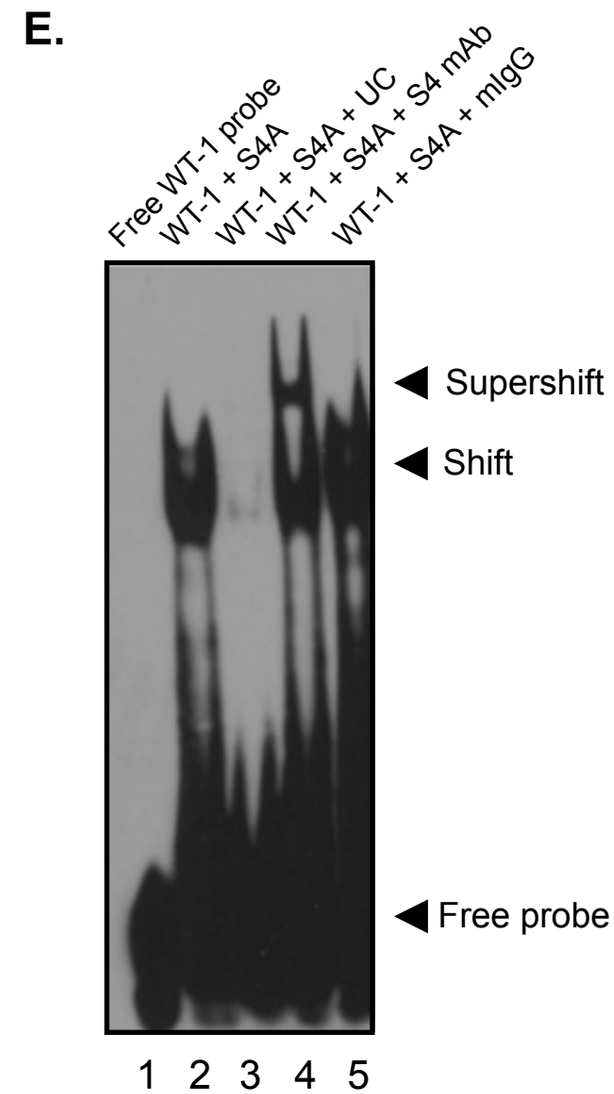
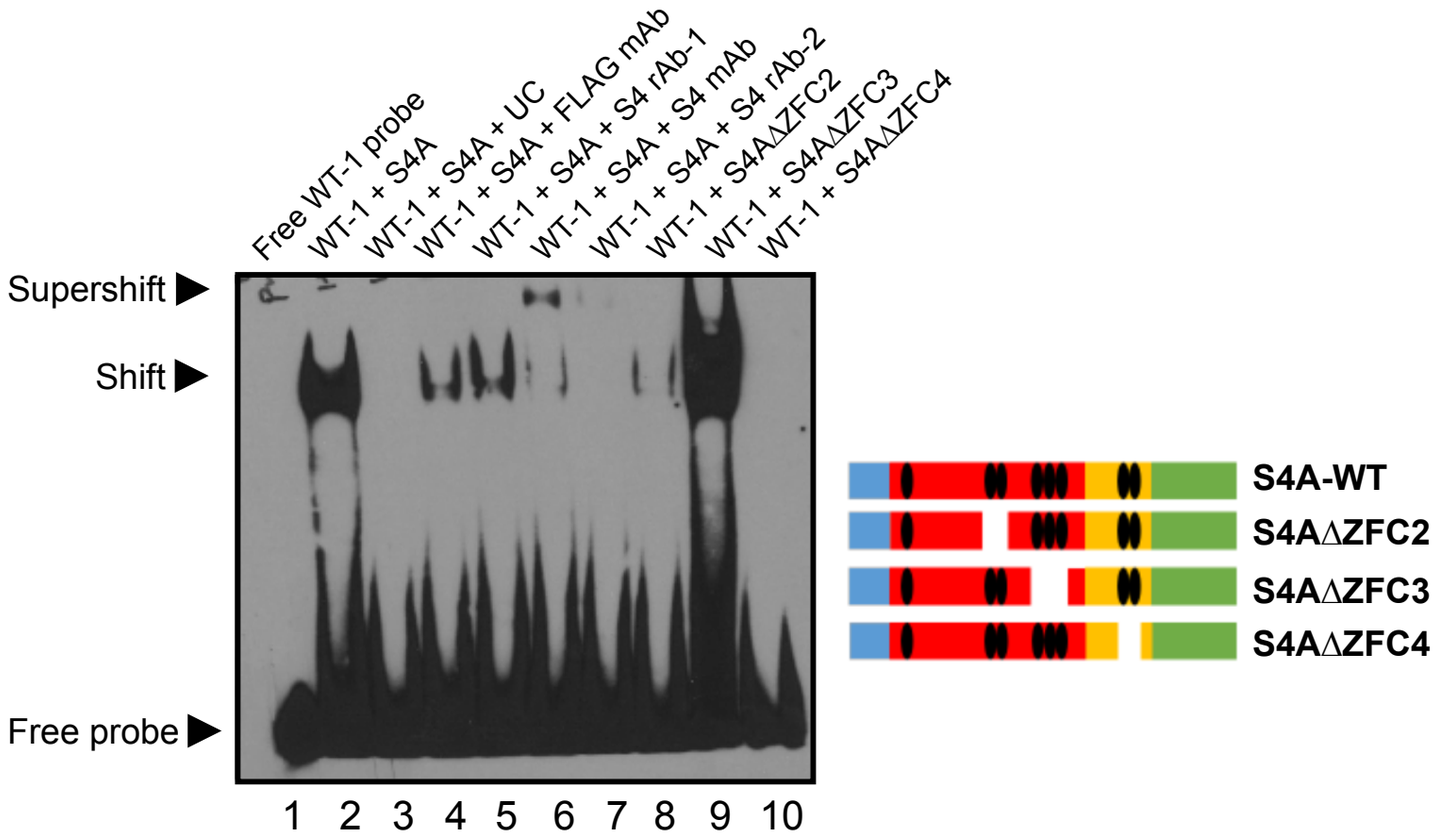

E.
Free WT-1 probe
WT-1 + S4A
WT-1 + S4A + UC
WT-1 + S4A + S4 mAb
WT-1 + S4A + mIgG
Supershift
Shift
Free probe
1 2 3 4 5

Figure 2

A.


Free WT-1 probe
WT-1 + S4A
WT-1 + S4A + UC
WT-1 + S4A + FLAG mAb
WT-1 + S4A + S4 rAb-1
WT-1 + S4A + S4 mAb
WT-1 + S4A + S4 rAb-2
WT-1 + S4AΔZFC2
WT-1 + S4AΔZFC3
WT-1 + S4AΔZFC4
Supershift
Shift
Free probe
1 2 3 4 5 6 7 8 9 10
S4A-WT
S4AΔZFC2
S4AΔZFC3
S4AΔZFC4

B.

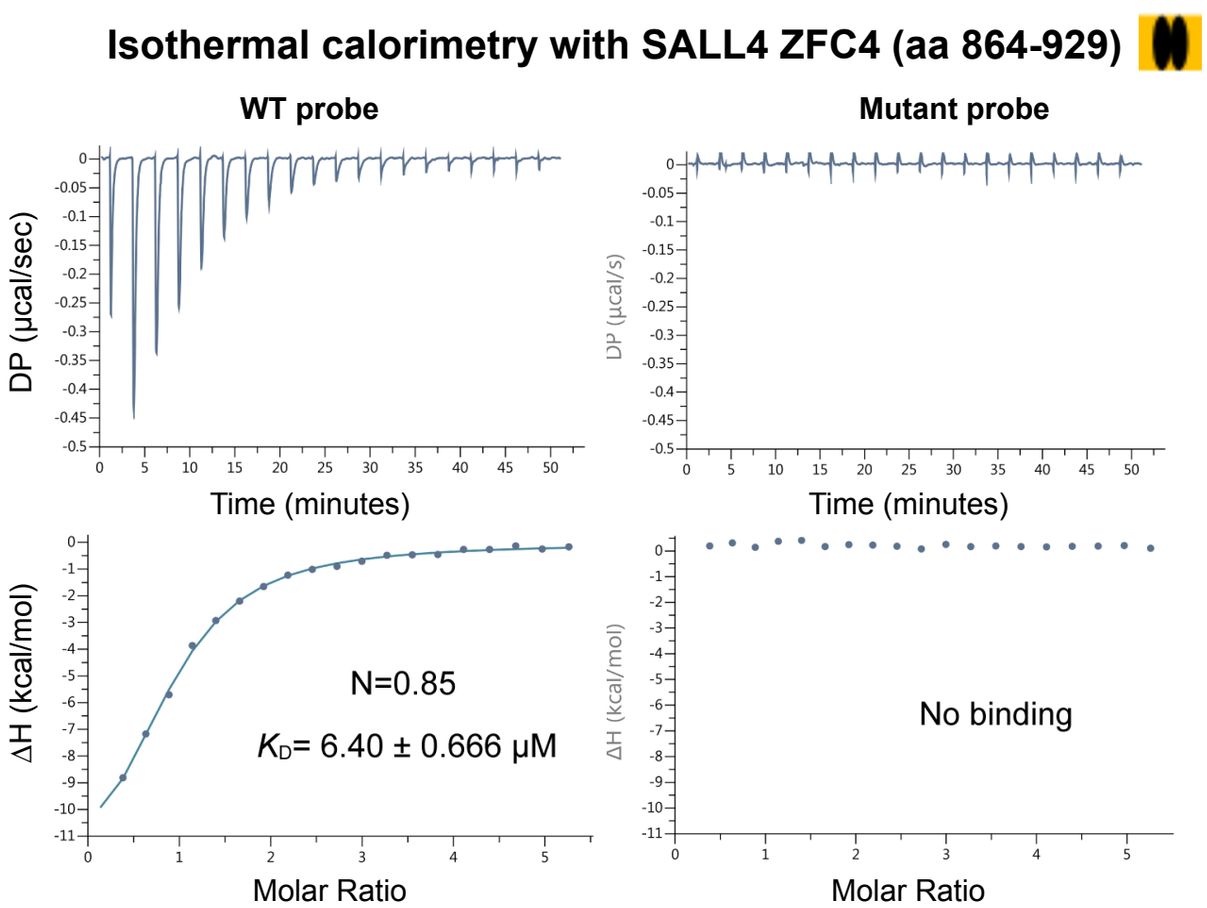
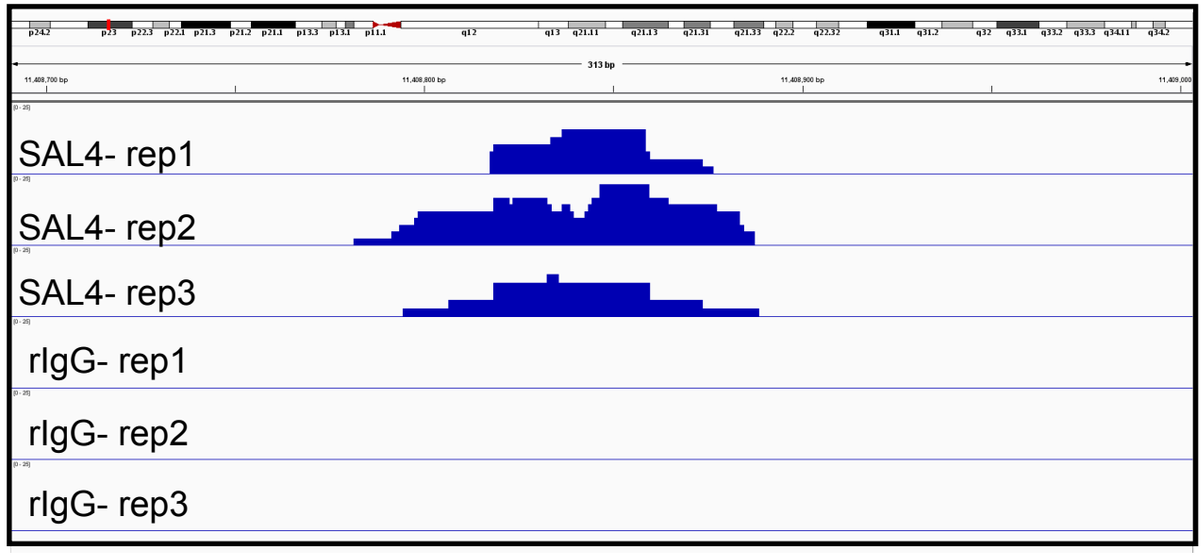

Isothermal calorimetry with SALL4 ZFC4 (aa 864-929)
WT probe
Mutant probe
DP (µcal/sec)
DP (µcal/s)
Time (minutes)
ΔH (kcal/mol)
N=0.85
K_D= 6.40 ± 0.666 µM
No binding
Molar Ratio

Figure 3

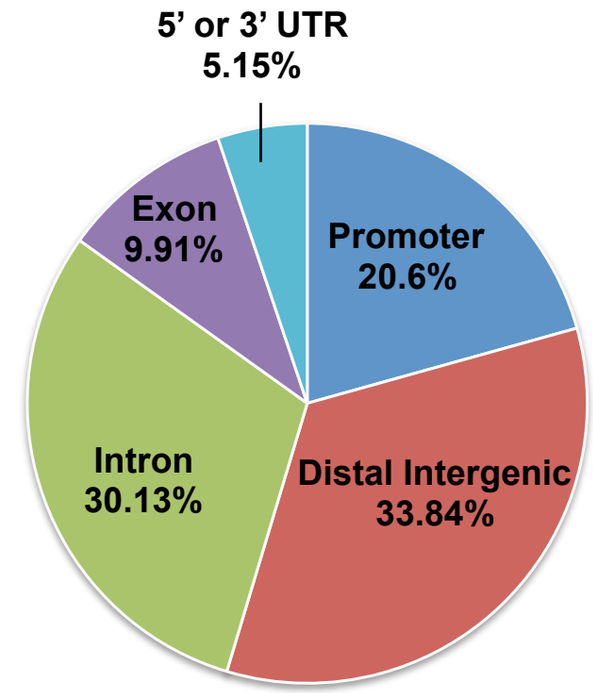
A.

SAL4 CUT&RUN in SNU398 cells



Distal intergenic, Chr9: 11,408,694-11,409,007

B.



C.

SALL4 rep1

Rank	Motif	P-value
1	ATCGAATGGAAT	1e-1774
2	TGGAATCATC	1e-801
3	ATCATCAAATGG	1e-613
4	TCGAACGG	1e-248
5	AAGTTATTAA	1e-214

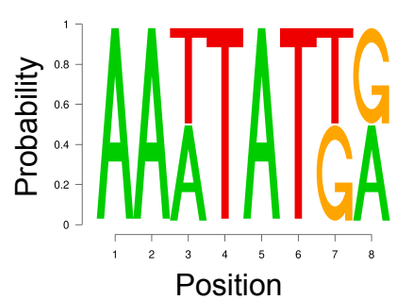
SALL4 rep2

Rank	Motif	P-value
1	TCGAATGGAATC	1e-2145
2	AAAAATTTATTT	1e-846
3	AGTGATCT	1e-794
4	AAGGCATTTG	1e-604
5	ATGTGTTT	1e-573

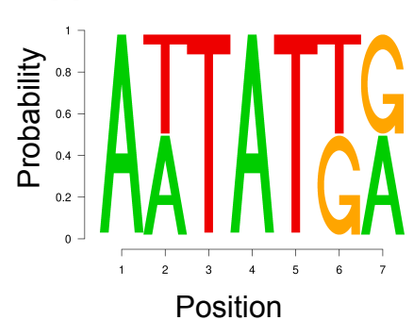
SALL4 rep3

Rank	Motif	P-value
1	TCGAATGGAATC	1e-2246
2	ATAAATGTCA	1e-674
3	AAATATTAATTT	1e-654
4	GTACAATTGA	1e-450
5	TGTGAGCCACTG	1e-439

Motif 1



Motif 2



D.

Percentage of CUT&RUN peaks containing the Motif 2 (partial Motif 1)

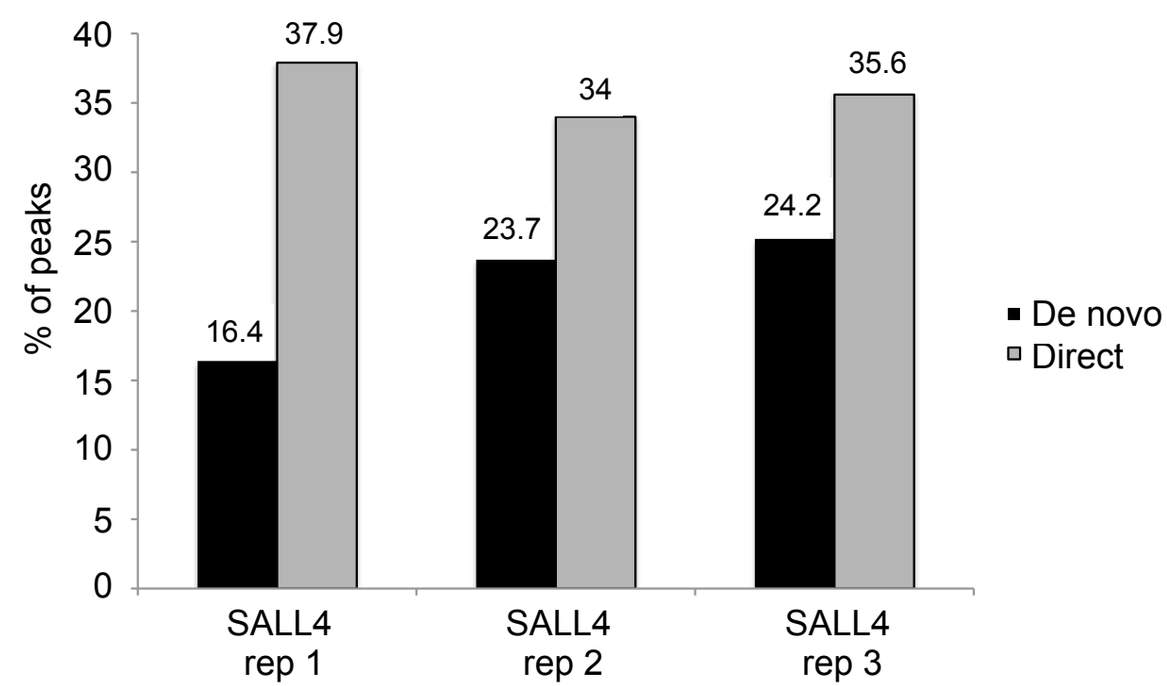
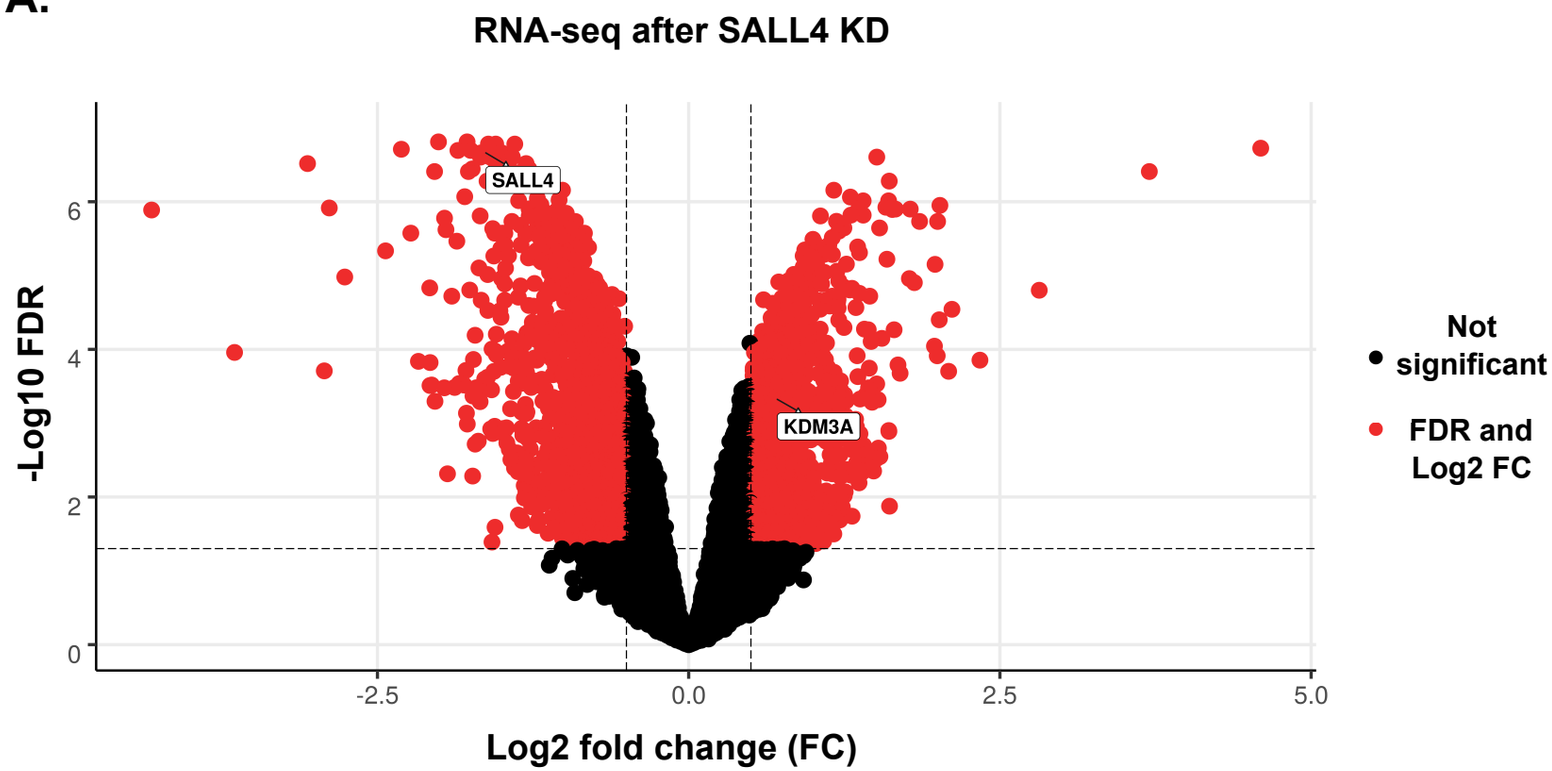
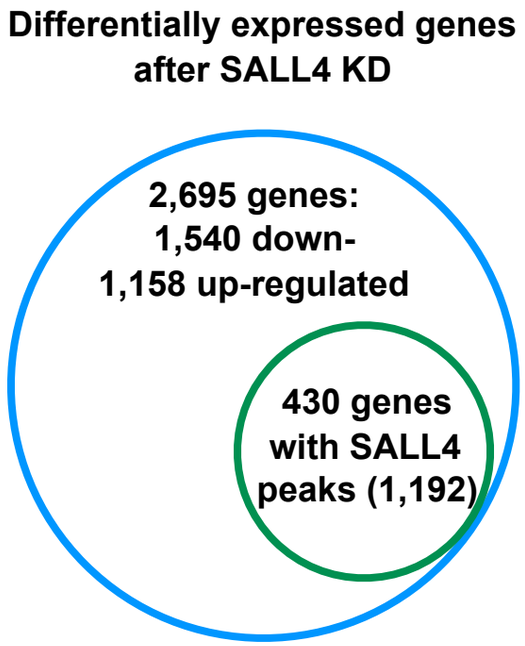


Figure 4

A.



B.



C.

De novo motif search
SALL4-bound differentially expressed genes

Rank	Motif	P-value	log P-pvalue
1	GAAATTTTCATG	1e-26	-6.097e+01
2	ACAATTTGTGGC	1e-22	-5.079e+01
3	TAATATTGCAAT	1e-21	-4.968e+01
4	GCAATTATTGCT	1e-20	-4.827e+01
5	GTTTAACTGA	1e-19	-4.484e+01

D.

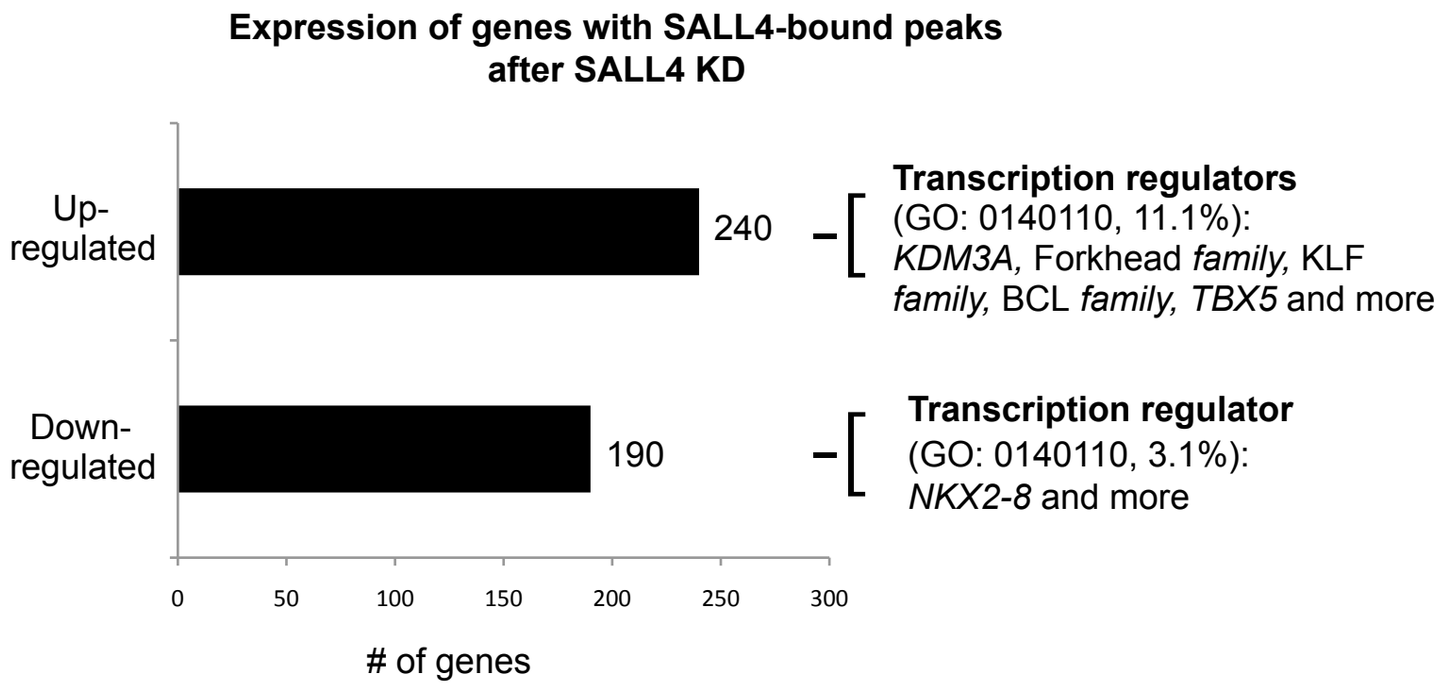
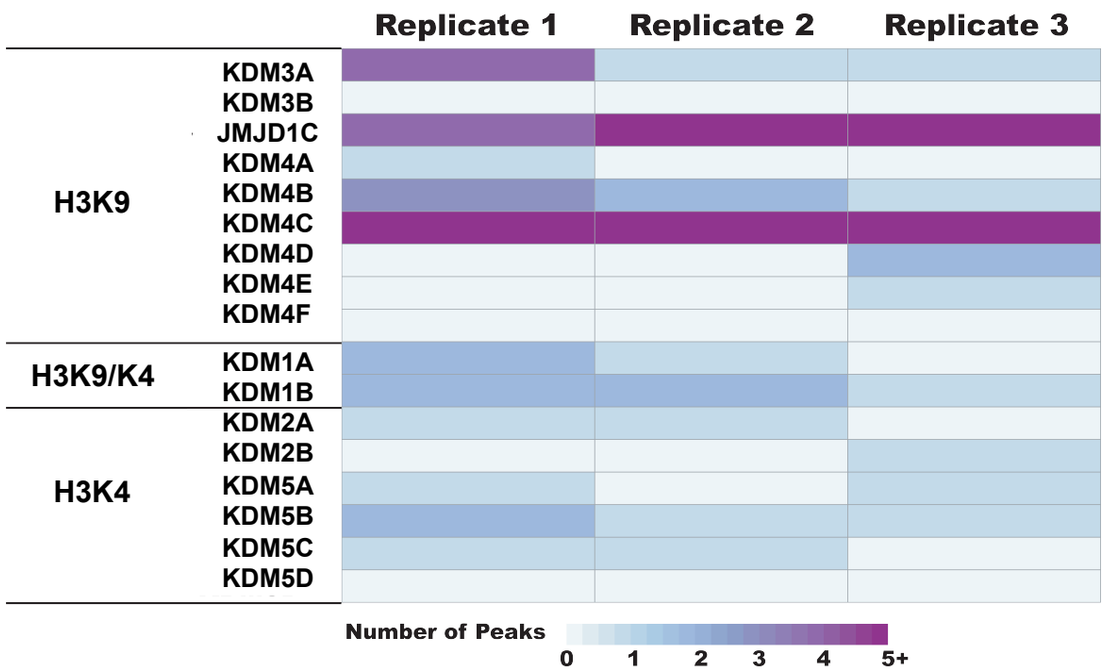
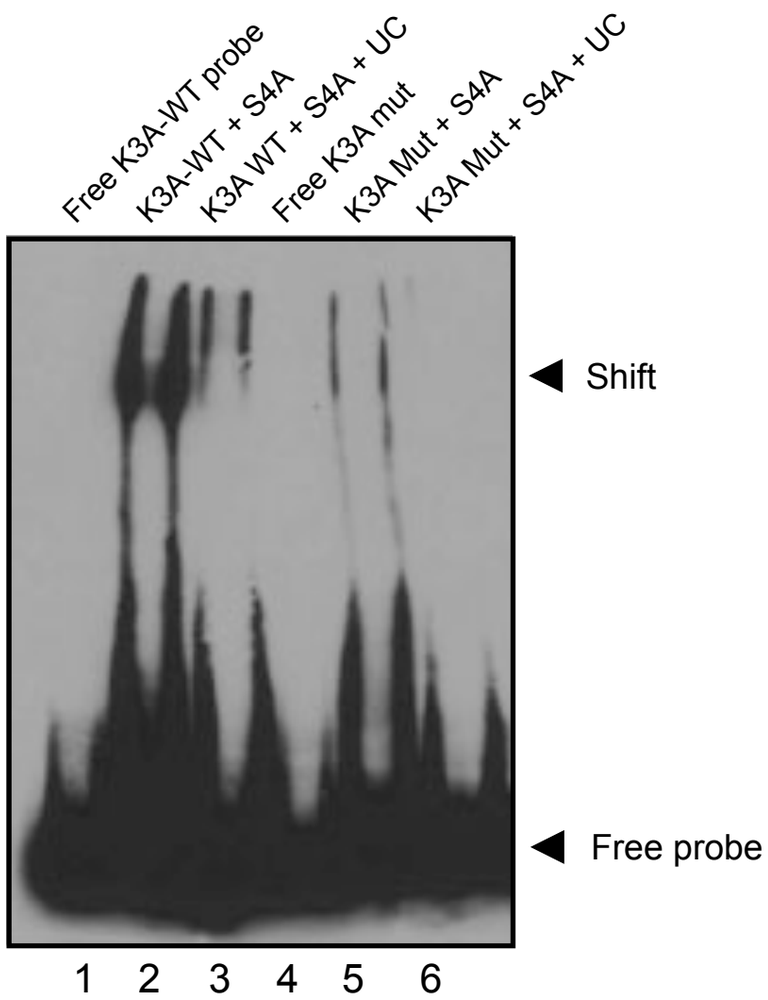


Figure 5

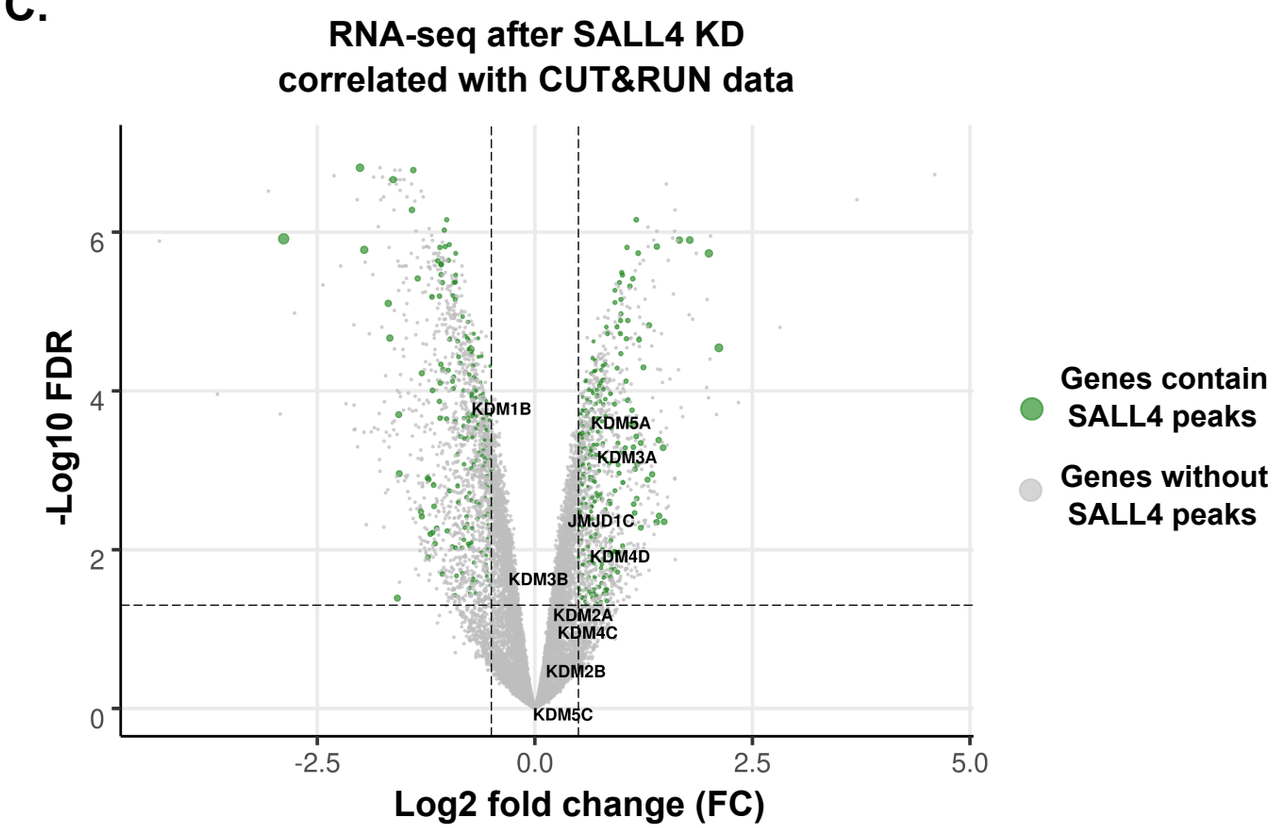
A.



B.



C.



D.

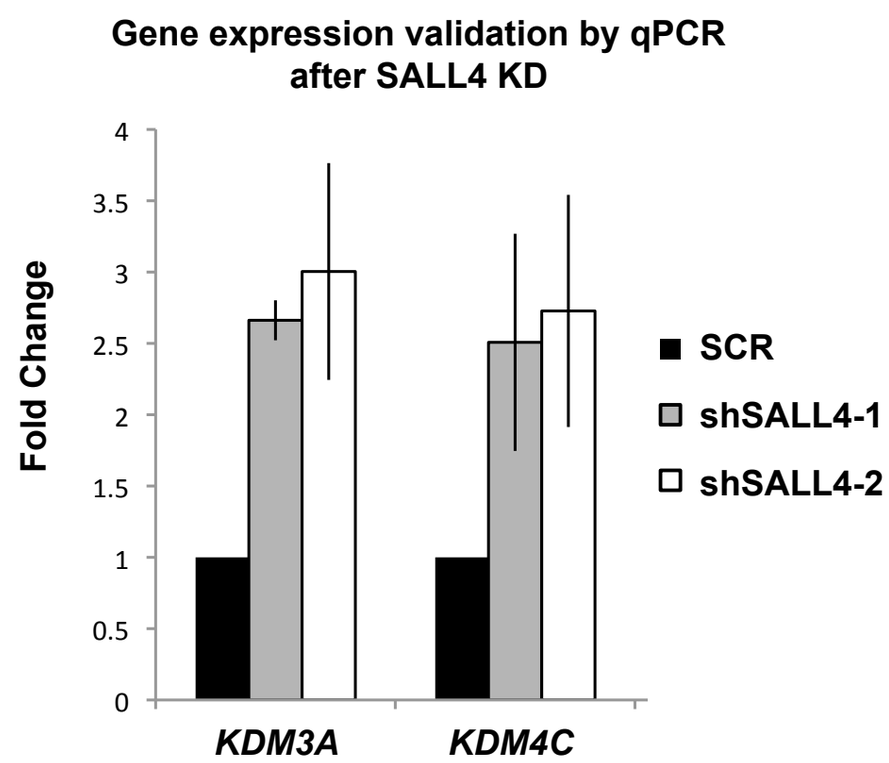
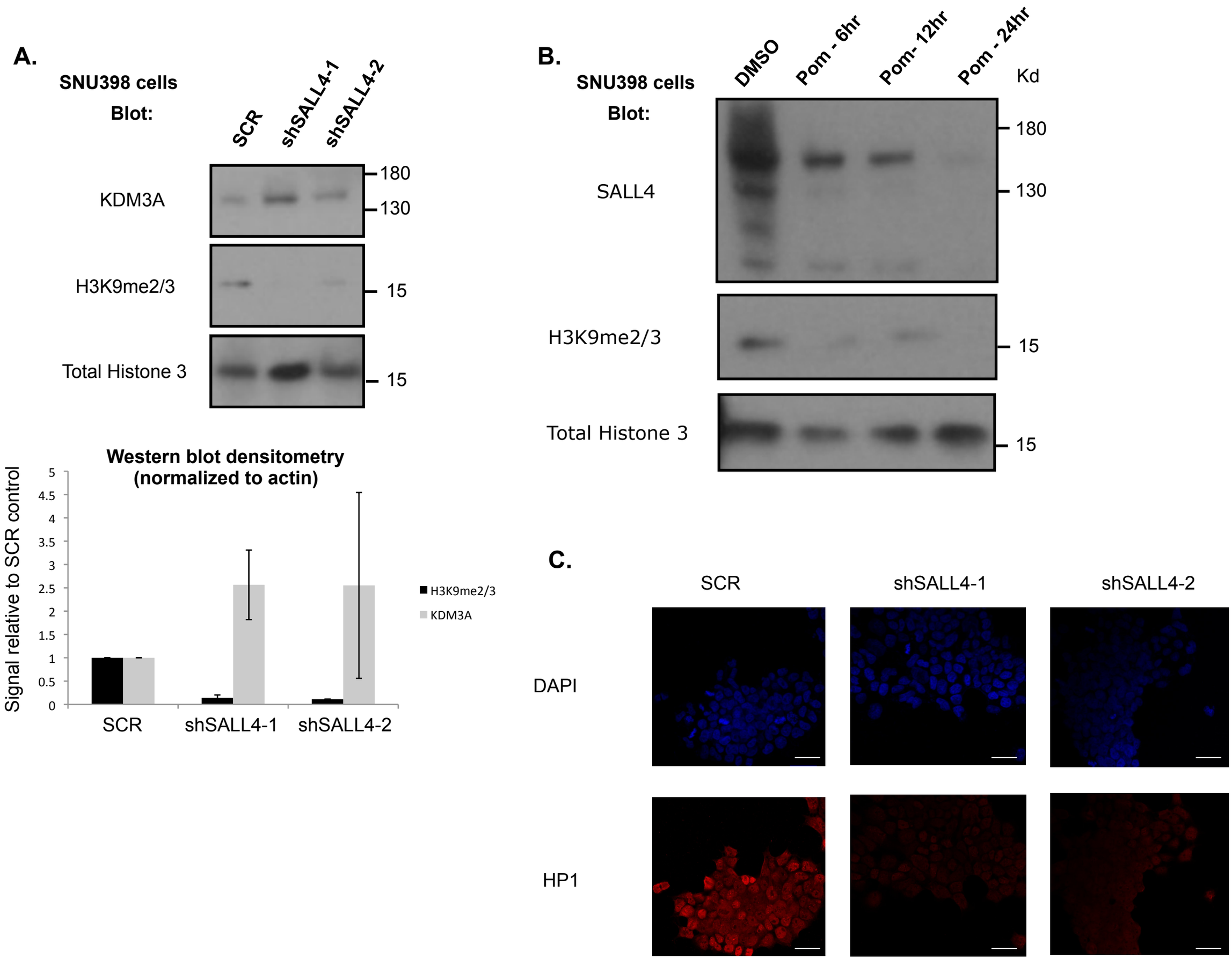


Figure 6


A.
SNU398 cells
Blot:
SCR
shSALL4-1
shSALL4-2
KDM3A
H3K9me2/3
Total Histone 3
180
130
15
15
Western blot densitometry
(normalized to actin)
Signal relative to SCR control
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
SCR
shSALL4-1
shSALL4-2
H3K9me2/3
KDM3A
B.
SNU398 cells
Blot:
DMSO
Pom - 6hr
Pom- 12hr
Pom - 24hr
Kd
SALL4
H3K9me2/3
Total Histone 3
180
130
15
15
C.
SCR
shSALL4-1
shSALL4-2
DAPI
HP1