



Structural studies of SALL family protein zinc finger cluster domains in complex with DNA reveal preferential binding to an AATA tetranucleotide motif

Received for publication, July 20, 2022, and in revised form, October 10, 2022. Published, Papers in Press, October 17, 2022.

<https://doi.org/10.1016/j.jbc.2022.102607>

Wenwen Ru^{1,‡}, Tomoyuki Koga^{2,‡}, Xiaoyang Wang¹, Qiong Guo¹, Micah D. Gearhart³, Shidong Zhao¹, Mark Murphy³, Hiroko Kawakami³, Dylan Corcoran³, Jiahai Zhang¹, Zhongliang Zhu¹, Xuebiao Yao¹, Yasuhiko Kawakami^{3,4,*}, and Chao Xu^{1,*}

From the ¹MOE Key Laboratory for Cellular Dynamics, Hefei National Center for Cross-disciplinary Sciences, School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, P. R. China; ²Department of Neurosurgery, ³Department of Genetics, Cell Biology and Development, and ⁴Stem Cell Institute, University of Minnesota, Minneapolis, Minnesota, USA

Edited by Wolfgang Peti

The Spalt-like 4 transcription factor (SALL4) plays an essential role in controlling the pluripotent property of embryonic stem cells *via* binding to AT-rich regions of genomic DNA, but structural details on this binding interaction have not been fully characterized. Here, we present crystal structures of the zinc finger cluster 4 (ZFC4) domain of SALL4 (SALL4^{ZFC4}) bound with different dsDNAs containing a conserved AT-rich motif. In the structures, two zinc fingers of SALL4^{ZFC4} recognize an AATA tetranucleotide. We also solved the DNA-bound structures of SALL3^{ZFC4} and SALL4^{ZFC1}. These structures illuminate a common preference for the AATA tetranucleotide shared by ZFC4 of SALL1, SALL3, and SALL4. Furthermore, our cell biology experiments demonstrate that the DNA-binding activity is essential for SALL4 function as DNA-binding defective mutants of mouse *Sall4* failed to repress aberrant gene expression in *Sall4*^{-/-} mESCs. Thus, these analyses provide new insights into the mechanisms of action underlying SALL family proteins in controlling cell fate *via* preferential targeting to AT-rich sites within genomic DNA during cell differentiation.

Transcription factors (TFs) play essential roles in embryo development through binding to the specific regions of genomic DNA to direct different complexes in mediating programmable gene transcription (1–3). The occupancy of sequence-specific TFs is typically determined by the base composition within a genomic DNA region (4–6). It has been well known that unmodified CpG dinucleotide serves as signaling motif to recruit epigenetic regulators containing CXXC domain, a CpG-binding module (7–9). Although AT-rich regions are also highly enriched in important regulatory genomic DNA elements, including TATA box (10, 11), whether they are also recognized by sequence-specific TFs and

how they function in embryo development are largely unknown (12, 13).

Very recently, two lines of work independently identified Spalt-like transcription factor 4 (SALL4) as the AT-rich DNA-binding protein *via* pull-down mass spectrometry screen and protein binding microarray, respectively (14, 15). SALL4 belongs to the Spalt-like transcription factors (SALLs) family, which includes SALL1–4 (16). Both SALL1 and SALL3 contain four zinc finger clusters (ZFCs), termed as ZFC1–4, whereas SALL2 and SALL4 contain three ZFCs and lack ZFC4 and ZFC3, respectively (14). SALL4 is highly expressed in embryonic stem cells (ESCs) and several tumors but absent in most adult tissues. Dysfunctional SALL4 pathway is associated with severe human diseases, including Holt-Oram syndrome (17), acro-renal-ocular syndrome (18), leukemogenesis, and other cancers (19). SALL4 contains seven zinc fingers within its three clusters and an additional single zinc finger near the N-terminus (14). The ZFC4 is essential for SALL4 to recognize AT-rich sequence to repress expression of a variety of genes, and its mutation results in abnormal differentiation and embryonic lethality (14, 15). All SALL proteins except SALL2 recognize AT-rich DNAs *via* ZFC4 (14, 15). Despite the critical role of SALL4^{ZFC4} in embryo development and its conservation in other SALL proteins, how it binds to AT-rich DNAs and how the SALL4 occupancy at AT-rich regions influences gene expression remain elusive.

By using isothermal titration calorimetry (ITC)-binding assays, we ascertain that the ZFC4 of SALL3 and SALL4 prefer AT-rich DNAs. By solving the structures of SALL3 and SALL4 ZFC4 bound with different AT-rich DNAs, we revealed that the SALL3 and SALL4 ZFC4 recognize AT-rich DNAs through the Gln-Ade residue base pairing and thymine-mediated van der Waals interactions. In addition, we found that SALL4^{ZFC1} also serves as a binder of AT-rich DNAs, albeit with weaker binding affinity. Inspired by previous finding that loss of *Sall4* in ESCs causes aberrant neural gene expression (20), we evaluated the functional relevance of DNA-binding and found that DNA-binding deficient SALL4 mutant fails

[‡] These authors contributed equally to this work.

* For correspondence: Chao Xu, xuchaor@ustc.edu.cn; Yasuhiko Kawakami, kawak005@umn.edu.

Structures of SALL4 bound with AT-rich DNAs

to repress aberrant expression of several genes, such as *Irx3* and *Irx5*. Therefore, our study not only unveils how ZFC4 of SALL proteins preferentially recognizes AT-rich DNAs but also sheds light on the biological importance of their binding to AT-rich DNA sequences in ESCs.

Results

ZFC4 of SALL4 and SALL3 selectively recognize AT-rich DNAs

To quantitatively study the binding activity of SALL4 to DNAs with different base compositions, we cloned, expressed, purified a fragment of human SALL4 spanning SALL4^{ZFC4} (residues 856–930) (Fig. 1A), and measured its binding affinities towards different dsDNAs by using ITC. Consistent with previous studies (14, 15), SALL4^{ZFC4} binds to different 12-mer DNAs, containing AATATT with K_D s in a range of 6.9 to 9.0 μ M (Fig. 1, B and C and Table S1). In contrast, the binding was abolished when central four (ATAT) or six (AATATT) nucleotides were replaced by CG-rich nucleotides (Table S1).

To understand whether other SALL family members possess similar DNA-binding selectivity, we cloned, expressed, and purified the ZFC4 domain of SALL3 (SALL3^{ZFC4}) spanning aa 1102 to 1167 and examined its DNA-binding property by ITC. Our binding analyses indicate that SALL3^{ZFC4} binds to the AATATT-containing 12-mer DNA with a K_D of 8.0 μ M (Fig. 1D), comparable to that of SALL4^{ZFC4} (K_D = 6.9 μ M). Like SALL4^{ZFC4}, SALL3^{ZFC4} displayed no binding affinity towards CGCG- or CGCGCG-containing DNAs (Table S1). Thus, we conclude that ZFC4 domains of SALL3 and SALL4 specifically recognize AT-rich DNAs judged by an *in vitro* binding assay.

The structures of SALL4 with different AT-rich DNAs

To gain insight into the molecular mechanism underlying AT-rich DNA recognition by SALL4^{ZFC4}, we solved the crystal structure of the SALL4^{ZFC4} with a 12-mer dsDNA (5'-GGT AATATTTC-3') in a 2.45 Å resolution (Table S2). The density maps of protein and the dsDNA are of high quality

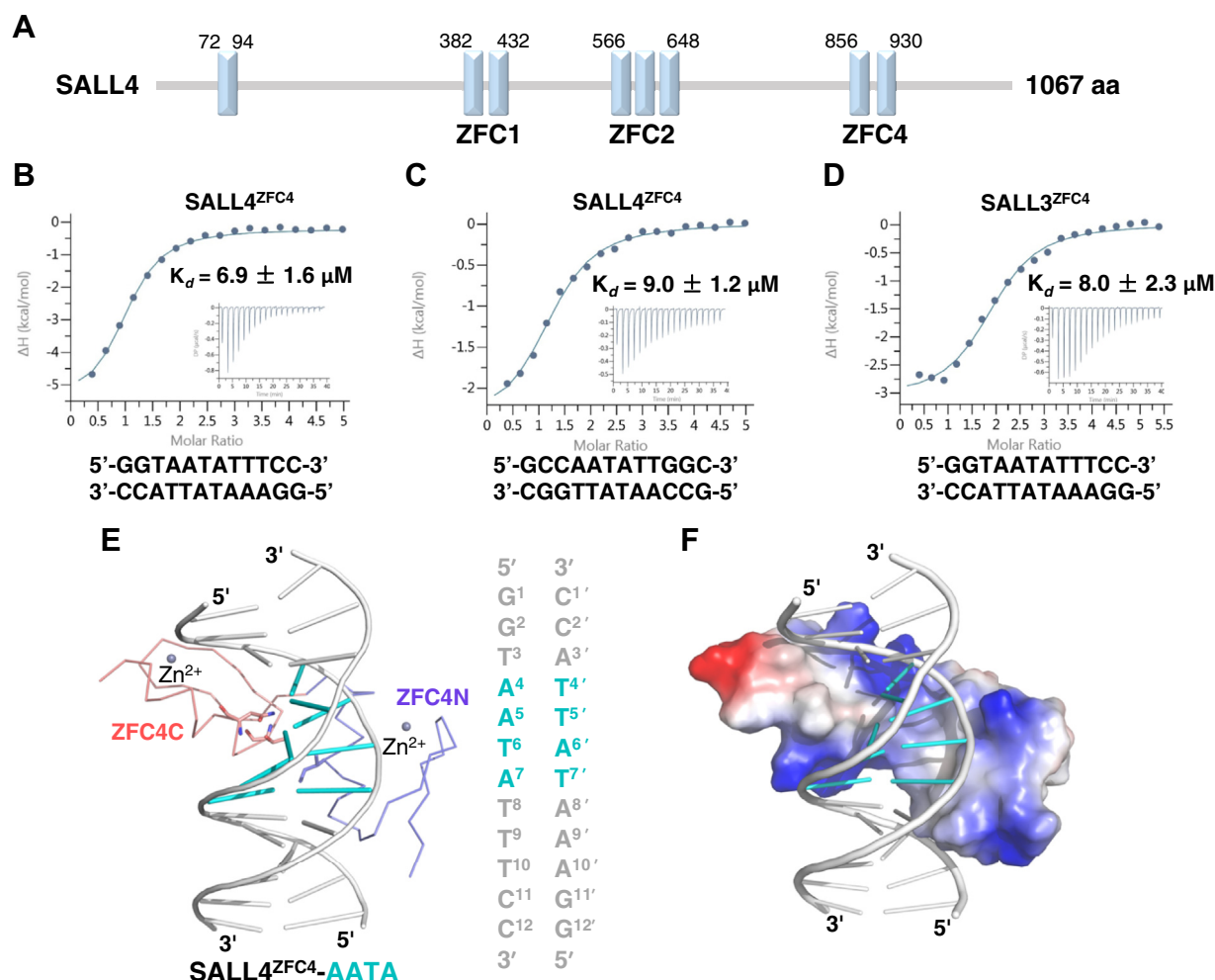


Figure 1. Structure of SALL4^{ZFC4} bound with the 12-mer dsDNA. A, domain architecture of human SALL4 containing ZFC1, ZFC2, and ZFC4, with the boundaries indicated. B, ITC-binding curves for SALL4^{ZFC4} binding to the 12-mer dsDNA (5'-GGTAATATTTC-3'). C, ITC-binding curves for SALL4^{ZFC4} binding to the 12-mer dsDNA (5'-GCCAATATTGGC-3'). D, ITC-binding curves for SALL3^{ZFC4} binding to the 12-mer dsDNA (5'-GGTAATATTTC-3'). E, crystal structure of SALL4^{ZFC4} with the 12-mer dsDNA (5'-GGTAATATTTC-3'). The DNA is shown in gray cartoon except the central AATA base pair, which is shown in cyan. Two zinc fingers of SALL4^{ZFC4}, ZFC4N, and ZFC4C are shown in purple and salmon ribbon, respectively. SALL4^{ZFC4} residues involved in base-specific DNA recognition are shown in sticks. F, electrostatic surface of SALL4^{ZFC4} bound with the 12-mer dsDNA, with the DNA shown in the same orientation and color as shown in (E). ITC, isothermal titration calorimetry; SALL, Spalt-like transcription factor; ZFC, zinc finger cluster domain.

(Fig. S1, A and B). There are two dsDNA molecules in an asymmetric unit, with each of them bound with two SALL4^{ZFC4} molecules. SALL4^{ZFC4} is comprised of two zinc fingers, termed as ZFC4N and ZFC4C (Figs. 1E and 2A). In each complex, one SALL4^{ZFC4} molecule binds to the central major groove of the dsDNA with both zinc fingers visible, whereas the other one binds to the end of the dsDNA with only ZFC4C visible. Given that ITC binding data suggest that SALL4^{ZFC4} binds to the 12-mer dsDNA in a molecular ratio of 1:1 (Table S1), binding of the second SALL4^{ZFC4} to the dsDNA is likely due to the crystal packing, and the invisible ZFC4N might be due to its intrinsic flexibility. Therefore, our structural analysis focuses on the SALL4^{ZFC4} molecule bound at the central major groove of the dsDNA. Upon binding to SALL4^{ZFC4}, the major groove of 12-mer DNA becomes narrower (16.7 Å vs. 20.0 Å) (Fig. S1C)

SALL4^{ZFC4} wraps around the 12-mer DNA and interacts with it *via* the positively charged surface (Fig. 1, E and F). ZFC4N and ZFC4C each belongs to Cys2-His2 (C2H2) finger motif that adopts a canonical β - β - α architecture (Fig. 2A). Although the eight Zn²⁺-coordinating residues are conserved in SALL1-4, the spacing between the last two histidines is altered in SALL2 (Fig. 2A), further suggesting that C-terminal ZFC of SALL2 possesses distinct DNA-binding property.

Structural basis for ⁴AATA⁷-specific recognition by SALL4^{ZFC4}

In the structure, the central ⁴AATA⁷ tetranucleotide is recognized by SALL4^{ZFC4} *via* base-specific hydrogen bonding and van der Waals interactions (Fig. 2B). The side chain carboxyl and nitrogen groups of SALL4 Asn912 are hydrogen bonded to the N6 and N7 atoms of A5, respectively. The side chain carboxyl group of Asn912 also forms two water-mediated hydrogen bonds with N6 atoms of A4 and A6', respectively (Fig. 2, B–D). The Asn-Gua residue base pairing is analogous to the Arg-Gua pair in the DNA-bound CXXC domain structures (9, 21). The side chains of Ile887 and Thr909 make van der Waals interactions with the methyl group of T6, which forms one water-mediated hydrogen bond with the Thr909 side chain (Fig. 2, B and C). Furthermore, Gly911 of SALL4^{ZFC4} makes van der Waals interaction with methyl group of T7', allowing A7 to be favored in the complementary strand (Fig. 2, B and E). Collectively, above hydrogen bonding and van der Waals interactions render SALL4^{ZFC4} the ability to recognize the AATA motif within the 12-mer DNA.

In addition to the base-specific interactions, the SALL4^{ZFC4}-DNA complex is further stabilized by extensive electrostatic interactions between DNA backbone and the basic residues of SALL4. The Arg905 and His916 side chains

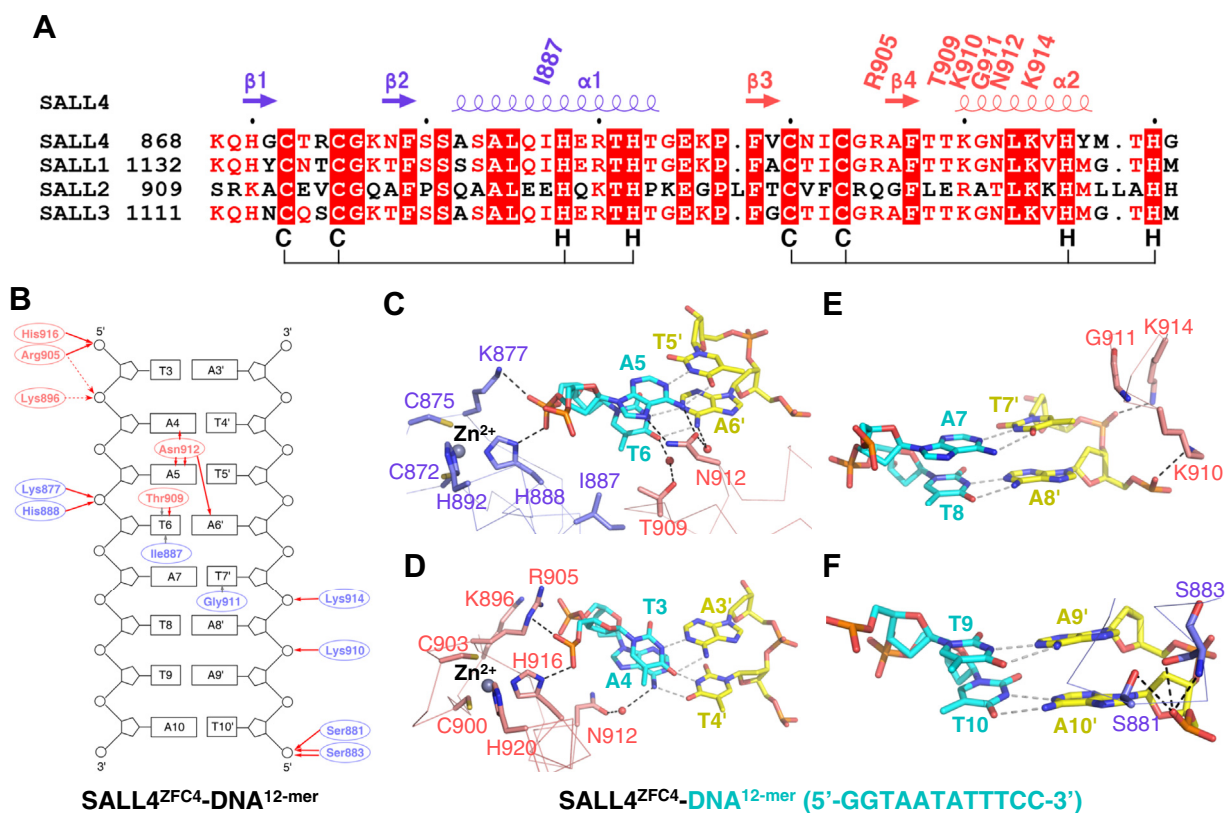


Figure 2. SALL4^{ZFC4} selectively recognizes AT-rich dsDNA. A, sequence alignment of human SALL family members, including SALL4 (Uniprot ID: Q9UJQ4), SALL1 (Uniprot ID: Q9NSC2), SALL2 (Uniprot ID: Q9Y467), and SALL3 (Uniprot ID: Q9BXA9). The secondary structure and DNA-binding residues are indicated at the top of sequences, while Zn²⁺-binding residues are indicated at the bottom of sequences. The black dots indicate 10th positions. B, schematic of the detailed interactions between SALL4^{ZFC4} and DNA. Residues in ZFC4N and ZFC4C are colored in purple and salmon, respectively. Inter-molecular hydrogen bonding and hydrophobic interactions are shown in red and gray arrows, respectively. C–F, detailed interactions between SALL4^{ZFC4} and (C) central ApT (A5-T5'/T6-A6'), (D) T3-A3'/A4-T4', (E) A7-T7'/T8-A8', (F) T9-A9'/A10-T10'. Nucleotides from two strands are shown in cyan and yellow sticks, respectively. The ZFC4N and ZFC4C residues involved in DNA binding are shown in purple and salmon, respectively. SALL, Spalt-like transcription factor; ZFC, zinc finger cluster.

Structures of SALL4 bound with AT-rich DNAs

form hydrogen bonds with T3 (Fig. 2, B and D); the side chains of Lys896 and Arg905 make electrostatic interactions with A4 (Fig. 2, B and D); the Lys877 and His888 side chains form two hydrogen bonds with T6 (Fig. 2, B and C); the side chains of Lys910 and Lys914 form are hydrogen bonded to the backbones of A8' and T7' (Fig. 2, B and E), respectively; Ser881 and Ser883 form several hydrogen bonds with the T10' backbone (Fig. 2, B and F).

Next, we applied structure-guided mutagenesis to evaluate the roles of SALL4 interfacial residues. While N912D abolished the binding, N912A reduced the DNA binding by > 55-fold (K_{D} s: >400 μM vs. 6.9 μM). In contrast, N912Q binds to the DNA with affinity comparable to the WT (K_{D} s: 8.0 μM vs. 6.9 μM), underscoring the critical role of base-specific hydrogen bonds between Asn912 and A5. The double mutation I887A/T909A weakened the DNA binding affinity by > 10-fold (K_{D} s: 75 μM vs. 6.9 μM), indicating the importance of the van der Waals interactions between Ile887, Thr909, and Thymine (T6); the triple mutant R905A/K910A/K914A disrupted the DNA binding, indicating the essential role of electrostatic interactions between basic residues and DNA backbone (Table S2). Collectively, mutagenesis and ITC-binding experiments further pinpointed the key interactions at the protein-DNA interface.

SALL4^{ZFC4} disfavors T or G upstream of ApT

Given that A5 and T6 are engaged in most of base-specific interactions, we sought to replace nucleotides flanking A5 to see how it could impact on its DNA binding. All nucleotide

replacements were based on the 12-mer dsDNA (5'-¹GGTAA-TATTTCC¹²-3'). ITC binding assay demonstrated that T3C/A3'G and A4C/T4'G only slightly weakened the binding to 12-mer AT-rich DNA (K_{D} s: 9.0–11 μM vs. 6.9 μM), whereas A4T/T4'A and A4G/T4'C decreased the SALL4-binding affinity by ~2.7 to 5 fold (K_{D} s: 19–36 μM vs. 6.9 μM).

To understand why SALL4 favors AAT and CAT, but not TAT or GAT, we modeled the structures of SALL4 bound with ⁴TAT⁷, ⁴GAT⁷, and ⁴CAT⁷, respectively (Fig. 3). When A4 is replaced by a thymine, the distance between the methyl moiety of T4 and C β of Asn912 is 3.2 Å, which likely results in the repulsion of the Asn912 side chain and the impaired hydrogen bonds between Asn912 and A5 (Fig. 3, A and B). In addition, A4T disrupts the water-mediated hydrogen bond between Asn912 and A4 (Fig. 3B). In the modeled structure of SALL4 bound with ⁴GAT⁷, N7 and O6 of G4, and carboxyl oxygen of Asn912, are all hydrogen bond acceptors, which disrupts the water-mediated hydrogen bond observed between Asn912 and A4 (Fig. 3C). In contrast, A4C did not affect the hydrogen bond between Asn912 and A4 and also maintains the water-mediated hydrogen bond (Fig. 3D). However, if C4 is methylated to mC4, mC4 would weaken the Asn912-A5 hydrogen bond as T4 does. Thus, our structural analysis and binding data further reveal that SALL4 prefers an ⁴AATA⁷ or a ⁴CATA⁷ motif within the 12-mer dsDNA. The AATA-specific recognition is achieved by the base-specific hydrogen bonds and the water-mediated hydrogen bonds, as well as the Thymine-specific hydrophobic interactions.

To study whether the AATA recognition by SALL4^{ZFC4} also applies for dsDNAs of different lengths, we determined the

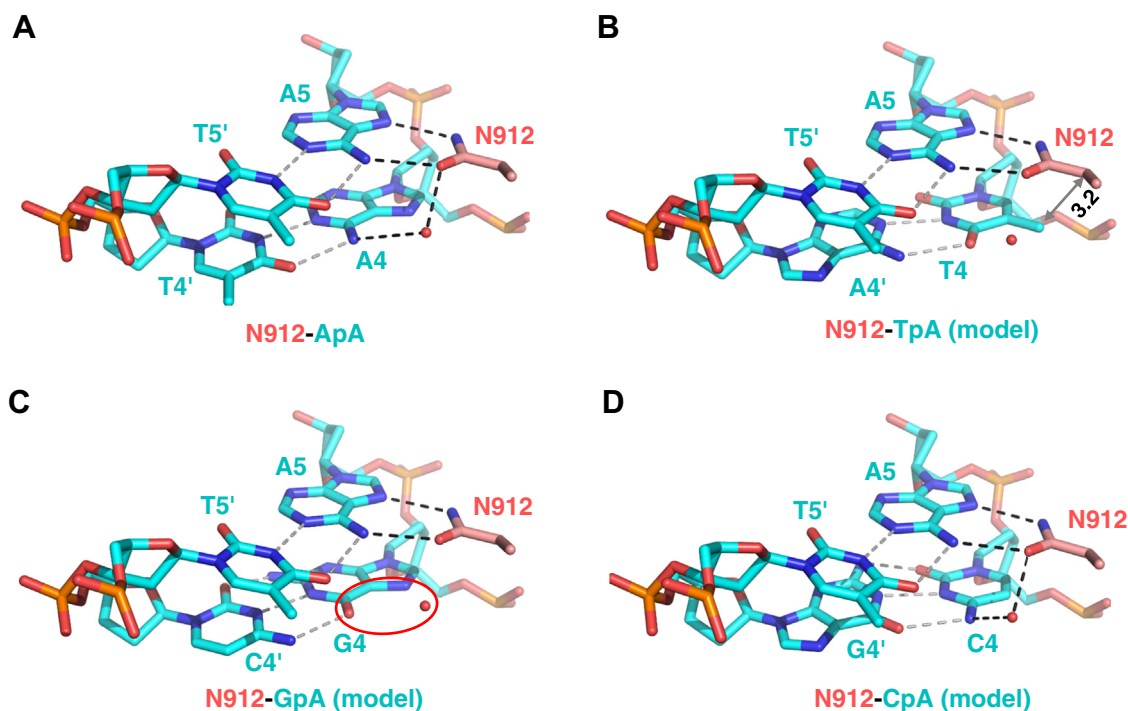


Figure 3. SALL4^{ZFC4} Asn912 disfavored TpA and GpA dinucleotides. A, in the DNA-bound SALL4^{ZFC4} structure, A5 is recognized by Asn912, which further stacks with upstream A4. B, in the modeled structure, the A4 substituted by T4 leads to potential steric clash between Asn912 side chain and the methyl group of T4. C, the A4 substituted by G4 disrupts the water-mediated hydrogen bond. D, the A4 substituted by C4 maintains the Asn912-mediated base specific interactions. SALL, Spalt-like transcription factor; ZFC, zinc finger cluster.

2.5 Å structure of SALL4^{ZFC4} with a 16-mer dsDNA containing two ATA motifs (Table S2). In the structure, the two SALL4^{ZFC4} molecules recognize ⁸AATA¹¹ and ⁵TATA⁸ within the 16-mer dsDNA, respectively, to form the complex in a 2:1 M ratio (protein: DNA) (Fig. S2, A–E). The sequence-specific recognition mode is the same as observed in the 12-mer DNA complex (Fig. S2, B–E). Consistently, the ITC-binding assay also demonstrates that SALL4^{ZFC4} binds to the 16-mer dsDNA with two K_Ds in a range of 13 to 16 μM. While N912Q mutant binds to the 16-mer DNA with K_Ds similar to those of the WT (K_Ds: 16–19 μM vs. 13–16 μM), R905A/K910A/K914A mutant displays no binding towards the 16-mer dsDNA (Table S1). Thus, we conclude that SALL4^{ZFC4} prefers AATA over other motifs within AT-rich dsDNA even within the context of multimeric binding.

Structure of SALL3^{ZFC4} with the 12-mer AT-rich dsDNA

To understand whether above DNA recognition mode also applies for other SALL members, we determined the crystal structure of SALL3^{ZFC4} with the same 12-mer AT-rich dsDNA

in a 2.50 Å resolution (Table S2). There is only one SALL3^{ZFC4}-DNA complex in an asymmetric unit. Similar to that of SALL4^{ZFC4} with 12-mer dsDNA, SALL3^{ZFC4} binds to the central major groove of the 12-mer DNA *via* its extensive positive charged surface (Fig. 4A). The DNA-bound SALL3^{ZFC4} structure is superimposed well with the two SALL4 complexes (Fig. 4B), with the rmsd in a range of 0.53 to 0.64 Å over 681 atoms, suggesting the conserved architecture of SALL complexes.

Extensive hydrogen bonding and van der Waals interactions were found between SALL3^{ZFC4} and DNA. SALL3 Asn1155, the counterpart of SALL4 Asn912, forms two base-specific hydrogen bonds with A5; Ile1130 and Thr1152 of SALL3, the counterparts of SALL4 Ile887 and Thr909, respectively, make van der Waals interaction with T6, which forms a water-mediated hydrogen bond with Thr1152; Gly1154 makes additional van der Waals interaction with T7' (Fig. 4C). Overall, ⁴AATA⁷ recognition by SALL3^{ZFC4} is the same as that by SALL4^{ZFC4}. In addition, the electrostatic interactions between DNA backbone phosphates and SALL3 residues, including Ser1124, Ser1126, His1131, Lys1139,

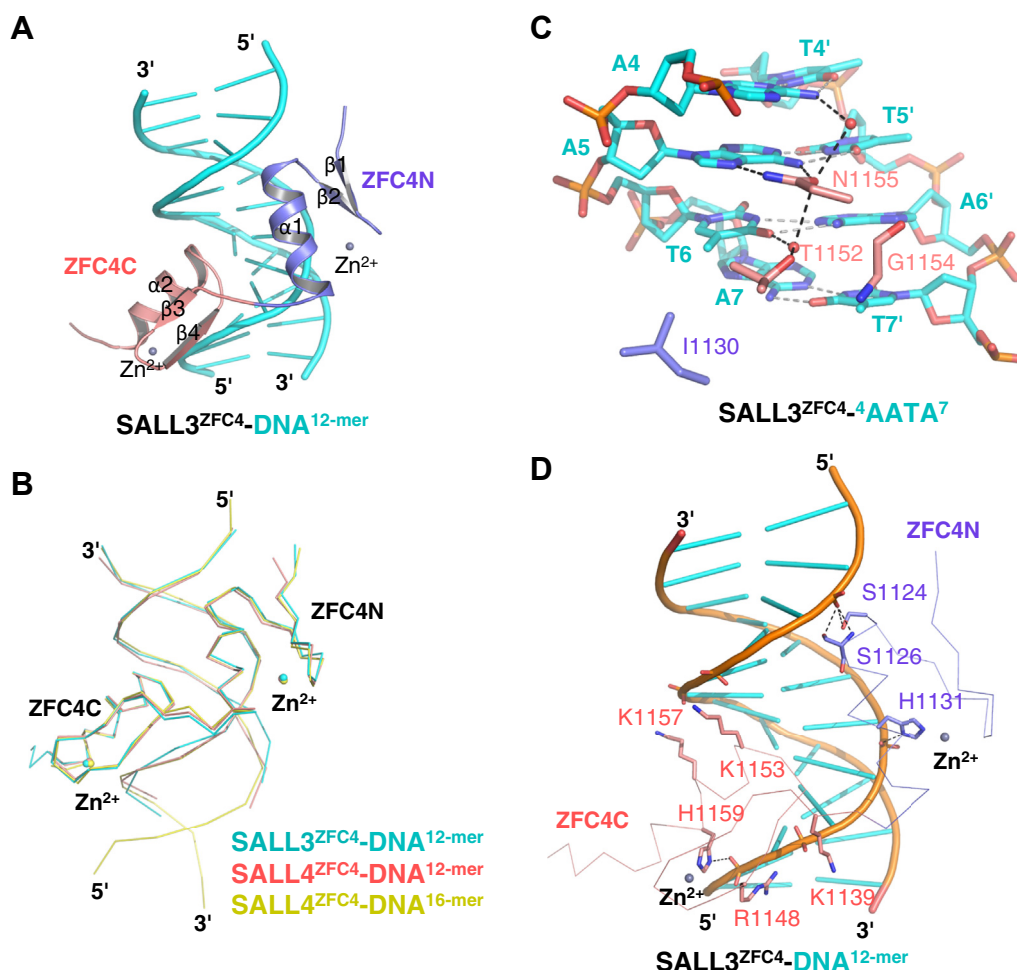


Figure 4. SALL3^{ZFC4} specifically recognizes AT-rich dsDNA. A, structure of SALL3^{ZFC4} bound with the 12-mer dsDNA (5'-GGTAATATTTCC-3'). The DNA is shown in cyan cartoon, while the two zinc fingers of SALL3^{ZFC4} are shown in purple and red cartoon, respectively. B, superposition of the structures of SALL3^{ZFC4} with the 12-mer dsDNA (cyan ribbon), SALL4^{ZFC4} with the 12-mer dsDNA (red ribbon), and SALL4^{ZFC4} with the 16-mer dsDNA (yellow ribbon). C, base specific interactions between SALL3^{ZFC4} and central ⁴AAT⁶, which are colored the same as in Figure 1E. D, interactions between SALL3^{ZFC4} and DNA backbone. Protein and DNA are shown in ribbon and cartoon, respectively. SALL3 residues are colored the same as in Figure 1E. SALL, Spalt-like transcription factor; ZFC, zinc finger cluster.

Structures of SALL4 bound with AT-rich DNAs

Arg1148, Lys1153, Lys1157, and His1159, are also conserved in the SALL4 complex (Fig. 4D). Given that the base-specific binding residues of SALL3 are conserved in SALL1 but not in SALL2, we reason that the preference for AATA-containing dsDNAs is conserved in ZFC4 domains of SALL1, SALL3, and SALL4.

SALL4^{ZFC1} recognizes AT-rich DNAs

Sequence alignment of SALL4 ZFC1 and ZFC4 shows that all SALL4^{ZFC4} residues involved in the recognition of AATA motif are conserved in SALL4^{ZFC1} except Ala882, which is replaced by an Asp (Asp394) in SALL4^{ZFC1} (Fig. S3A). We examined the DNA binding of SALL4^{ZFC1} (aa 378–453) by ITC. Binding data show that SALL4^{ZFC1} binds to the 12-mer AT-rich DNA with a K_D of 24 μ M and binds to the 16-mer AT-rich DNA with K_D s in a range of 17 to 21 μ M (Table S1), weaker than that for SALL4^{ZFC4}.

We further solved the structure of SALL4^{ZFC1} with the 16-mer AT-rich DNA at 2.72 Å resolution (Table S2). There are three dsDNAs and six SALL4^{ZFC1} molecules in one asymmetric unit, with one dsDNA bound with two SALL4^{ZFC1} molecules (Fig. 5A). In the structure, all six SALL4^{ZFC1}

recognizes A-T base pair (A9-T9' or A7'-T7) via Asn424 (Fig. 5, B–G). Asn424 of molecules A, D, and E recognizes A9 in the context of ApA (Fig. 5, B, E, and F), whereas Asn424 of molecules B, C, and F interacts with A7' in the context of TpA (Fig. 5, C, D, and G). The lengths of hydrogen bonds between Asn424 and A7' are in a range of 3.1 to 3.6 Å, longer than those observed between Asn424 and A9 (2.7–3.1 Å) (Fig. 5, B–G), suggesting the weaker hydrogen bonds in the context of TpA.

Next, we superimposed the structure of DNA-bound structure of SALL4^{ZFC1} with that of SALL4^{ZFC4} and found that Ala882 of ZFC4 is spatially proximal to the phosphate group of A14' due to the hydrogen bond between the DNA backbone and the main chain amide of Ser883 (Fig. S3A). In contrast, SALL4^{ZFC1} Asp394, the counterpart of SALL4^{ZFC4} Ala882, leads to charge repulsion with the DNA backbone phosphate, which would impair the hydrogen bond between A14' and SALL4^{ZFC1} Ser395, the counterpart of SALL4^{ZFC4} Ser883 (Fig. S3B). Consistent with the structural analysis, we found that A882D of SALL4^{ZFC4} reduced the DNA-binding affinity by > 6.5-fold (K_D s: 47 μ M vs. 6.9 μ M) (Fig. S3C). Collectively, our structural data, complemented by mutagenesis and binding experiments, reveals that SALL4^{ZFC1} also

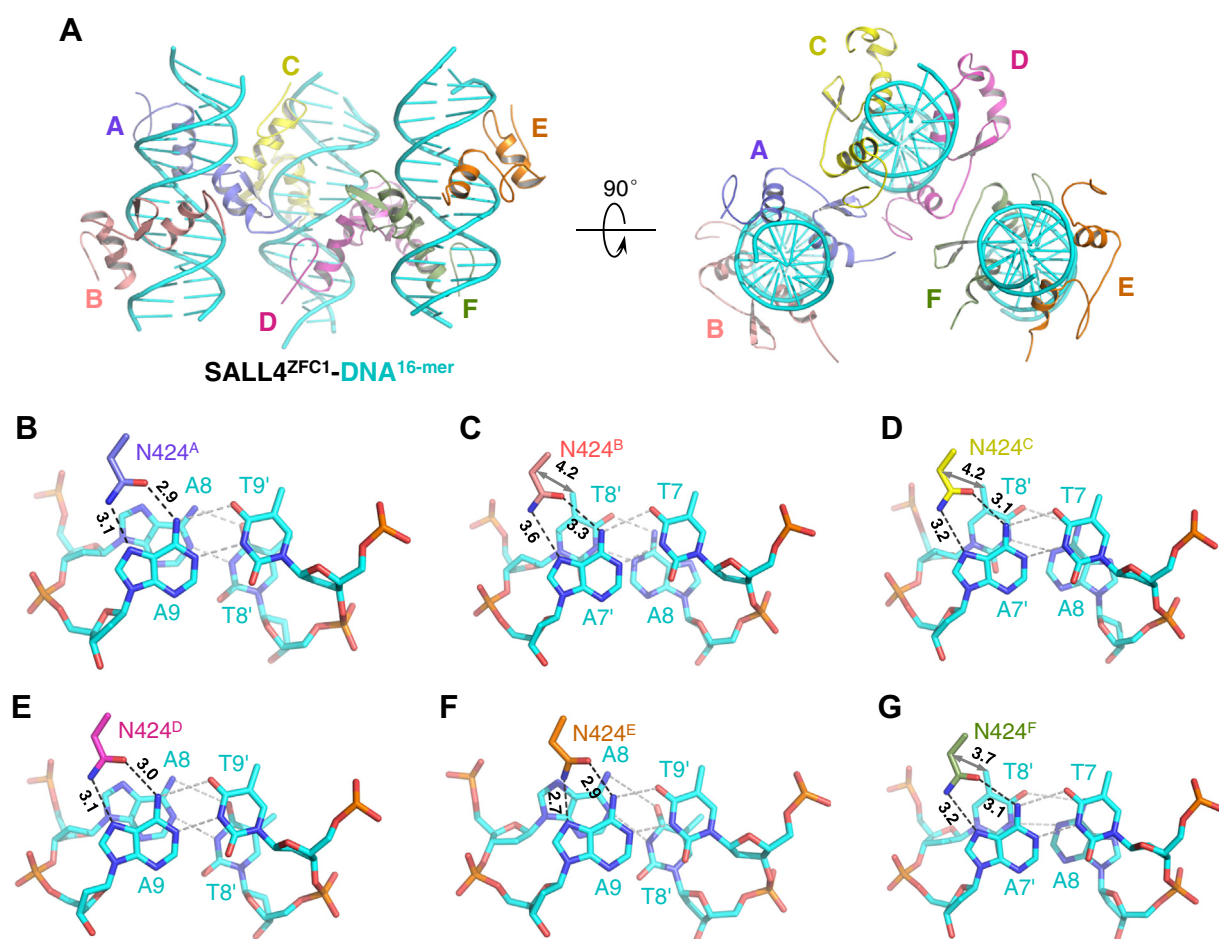


Figure 5. SALL4^{ZFC1} selectively binds to AT-rich dsDNA. A, structure of SALL4^{ZFC1} bound with the 16-mer dsDNA (5'-GGAATATAATATTCC-3'). Three dsDNAs are shown in cyan cartoon, while six SALL4^{ZFC1} molecules are shown in cartoon with different colors. B–G, all six SALL4^{ZFC1} molecules recognize central adenosine via Asn424. In (B, E, and F), Asn424 recognizes A9 in the context of ApA; In (C, D, and G), Asn424 recognizes A7' in the context of TpA. Asn424 is shown in sticks and all DNAs are shown in cyan sticks. SALL, Spalt-like transcription factor.

specifically recognizes AT-rich DNA, albeit with weaker affinity.

Targeting of SALL4 at AT-rich sites inhibits aberrant expression of differentiation prompting genes

In mouse ESCs, binding of SALL4 to AT-rich putative enhancer sequences prevents expression of differentiation promoting genes (14). To test whether loss of DNA binding in the SALL4^{ZFC4} Asn912 mutation has a biological significance, we generated *Sall4*^{-/-} mouse ESCs from *Sall4*^{flox} ESCs by infecting adenovirus-EGFP-Cre, which does not integrate into the genome. Then, we infected *Sall4*^{-/-} ESCs with lentivirus carrying either WT mouse *Sall4* or mouse *Sall4* N922D mutant (human SALL4 Asn912 corresponds to mouse SALL4 Asn922). First, we examined *Sall4* expression levels by qRT-PCR. Expression of transgene WT *Sall4* and *Sall4* N922D is approximately 1.8 and 0.9 fold, respectively, compared to *Sall4* expression in control ESCs (Fig. 6A). Then, we examined expression of several neural differentiation genes to which

SALL4 is enriched. We also performed SALL4 CUT&RUN experiments in order to detect SALL4 enrichment. De novo motif analysis of SALL4-enriched sequences showed AT-rich motifs (Fig. 6B), which is consistent with a recent SALL4 ChIP-seq result in mouse ESCs (14).

CUT&RUN experiments also showed enrichment of SALL4 near the *Sox1* gene (Fig. 6D). As previously shown (20), *Sox1* expression is elevated in *Sall4*^{-/-} ESCs, compared to the control (Fig. 6C). Both WT *Sall4* and *Sall4* N922D, introduced into *Sall4*^{-/-} ESCs, prevented aberrant expression of *Sox1* (Fig. 6C). SALL4 was also enriched in the region where *Irx3*, *Irx5*, and *Irx6* are closely located on chromosome 8 (Fig. 6H). Expression of *Irx3* was elevated in *Sall4*^{-/-} ESCs. The WT *Sall4* transgene repressed aberrant expression of *Irx3*, but *Sall4* N922D failed to repress *Irx3* expression (Fig. 6E). Similarly, WT not N922D *Sall4* repressed aberrant expression of *Irx5* (Fig. 6F), and a similar trend was observed for *Irx6* (Fig. 6G). We found that both peaks near the *Sox1* gene and 15 out of 17 peaks at the *Irx3*-*Irx5*-*Irx6* region contain the AATA sequence, which was found in 55,995 of the 64,249 (87%) of the genome-wide

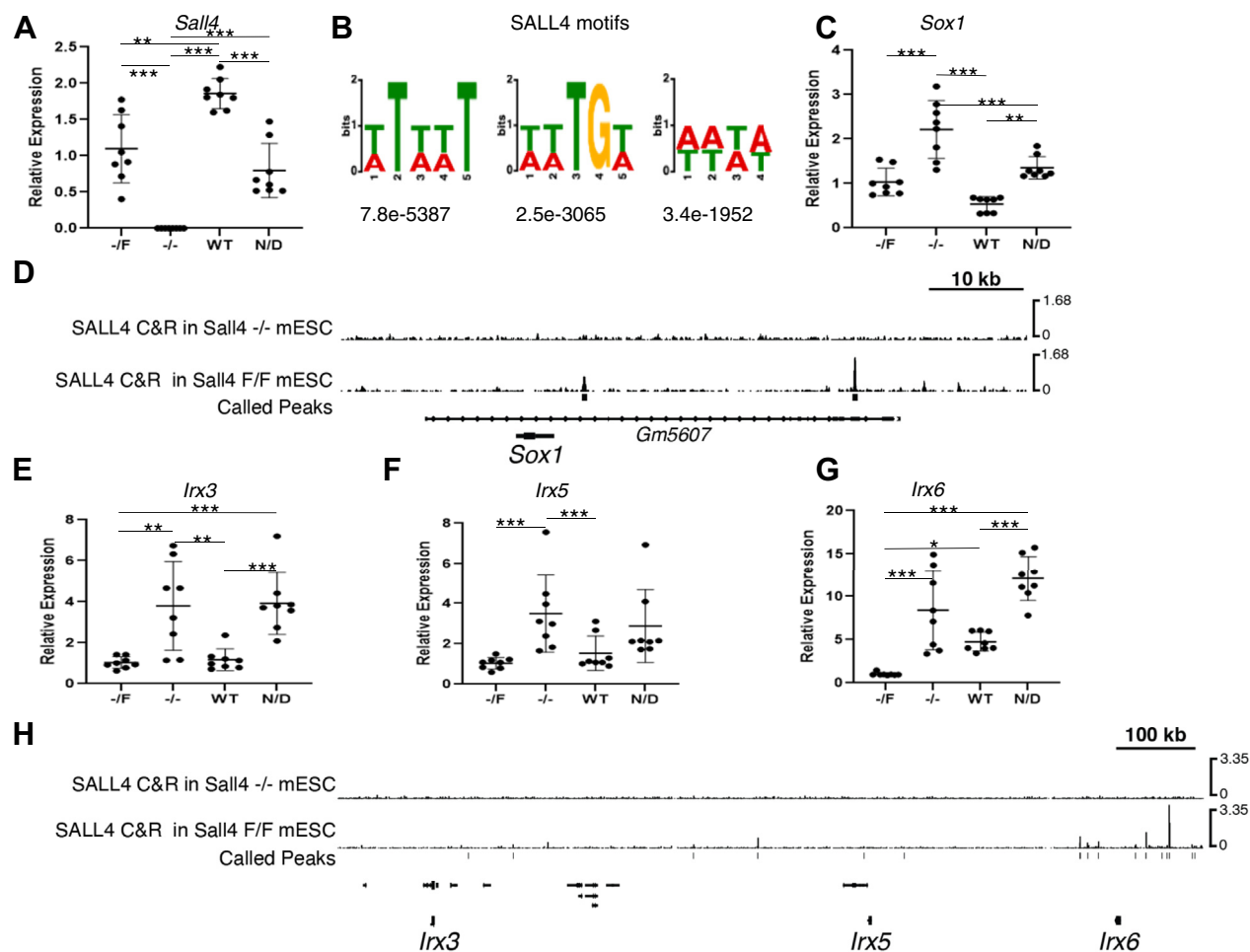


Figure 6. *mSall4* N922D mutant partially rescues aberrant gene expression in *Sall4*^{-/-} mouse ESCs. A, C, E-G, graphs showing relative expression levels of *mSall4* (A), *Sox1* (C), *Irx3* (E), *Irx5* (F), and *Irx6* (G) in *Sall4*^{flox} cells, *Sall4*^{-/-} cells, *Sall4*^{-/-} cells with WT *mSall4* expression (WT), and *Sall4*^{-/-} cells with *Sall4* N922D (ND) mutant expression. **p* < 0.05, ***p* < 0.01, ****p* < 0.001 by One-Way ANOVA with post-hoc Tukey HSD test. Each replicate is shown as a dot, and the average plus/minus SD is shown. B, top three motifs obtained by *de novo* motif analysis of SALL4-enriched sequences by CUT&RUN in mouse ESCs with *p* values shown under the motifs. D and H, SALL4 CUT&RUN tracks of the *Sox1* and *Irx3*-*Irx5*-*Irx6* regions in *Sall4*^{-/-} and *Sall4*^{flox/flox} (F/F) mouse ESCs. In (H), genes other than *Irx3*-*Irx5*-*Irx6* are not labeled for the simplicity. Mouse SALL4 Asn922 corresponds to human SALL4 Asn912. ESC, embryonic stem cell; SALL4, Spalt-like transcription factor.

Structures of SALL4 bound with AT-rich DNAs

SALL4-binding peaks. Inhibition of aberrant expression of *Sox1* and *Irx* genes is consistent with the notion that ZFC4-dependent DNA binding of SALL4 contributes to the repression of these gene expression. In addition, repression of *Sox1* by *Sall4* N992D might be associated with the binding of SALL4 N922D to the AT-rich region *via* its ZFC1 domain, consistent with our structural and biochemical data.

Discussion

More than 700 Zinc fingers proteins in human genome belong to the C2H2-type, and ~400 of them were annotated as TFs (22, 23). Uncovering the preferred motif of zinc finger TFs is important for understanding their roles in orchestrating spatiotemporal gene transcription. SALL family members are a subfamily of zinc finger proteins playing important roles in cell development and differentiation. Dysfunctional SALL proteins are associated with different types of cancers. In this study, we uncovered the conserved AATA-rich DNA recognition mode by SALL family members through presenting several structures of SALL proteins with respective DNA ligands. SALL proteins utilize a conserved Asn, such as Asn912 of SALL4 or Asn1155 of SALL3, to recognize the adenosine of A-T base pair, while hydrophobic residues of SALL proteins interact with the methyl moiety of the downstream thymine. In addition, the adenosine-binding Asn favors an adenosine at the upstream position. Overall, these base-specific interactions confer SALL proteins the ability to interpret AT-rich DNAs.

Given that SALL4^{ZFC1} binds to AT-rich DNAs weaker than SALL4^{ZFC4}, it might be insufficient to maintain the occupancy of SALL4 after the deletion of ZFC4, consistent with previous report that mutation or deletion of SALL4^{ZFC4} impairs the targeting of SALL4 at genome sites (14).

Comparison of SALL4-DNA structure with those of other AT-rich DNA complexes

It has been reported that transcriptional repressor MogR specifically recognizes AT-rich DNAs (24), which prompts us to compare it with the SALL4^{ZFC4} complex. The DNA recognition modes in two complexes are quite different (Fig. S4). MogR specifically recognizes the AT-rich motif but in a manner distinct from that between SALL4 and DNA ligands (Fig. S4). MogR binds to dsDNA as a dimer with the central AAAA tetranucleotide contacting both protomers. Arg140 of protomer A (Arg140^A) inserts into the narrow minor groove of AAAA tetranucleotide by forming two hydrogen bonds with the T6-A6' base pair. The other base-specific interactions are conferred by protomer B. Asn118^B of protomer B forms one hydrogen bond with A6'; Ser114^B and Gln117^B form water-mediated hydrogen bonds with A5' and T4, respectively; Val94^B and Tyr121^B make hydrophobic interactions with T3 and T4 (24). The AT-rich dsDNA recognition by MogR is likely minor-groove-specific (Fig. S4B), distinct from the major-groove-specific binding observed in the SALL4 complexes, which is mediated by the Asn-Adenosine pair (Fig. S4A).

Disease-associated mutations

Many mutations or deletions in SALL4 were known to result in Okhiro Syndrome. Only very few single mutations within ZFC4, including H888R, is reported to be associated with Okhiro Syndrome. Based on the Catalogue of Somatic Mutations in Cancer database (<https://cancer.sanger.ac.uk/cosmic>), identified single mutations in SALL4^{ZFC1} and SALL4^{ZFC4} likely have impact on protein stability and/or DNA binding affinity, including S396F and R431Q of SALL4^{ZFC1} and H888Q, R905Q, K914N, and H916Y of SALL4^{ZFC4} (Fig. S5, A and B). S396F disrupts the intramolecular hydrogen bond and might destabilize the protein, while R431Q of SALL4^{ZFC1} weakens the interaction with DNA backbone; R905Q and K914N would weaken the binding of SALL4^{ZFC4} with DNA backbone, whereas H888Q and H916Y not only disrupt the binding to Zn²⁺ but also abolish the hydrogen bond with DNA backbone. Consistently, our ITC assays show that while R905Q and K914N reduced the 12-mer dsDNA-binding affinity by ~14-fold and ~6-fold, respectively, neither H888Q nor H916Y displays detectable DNA-binding affinity (Fig. S5, C-F and Table S1). These disease-associated mutations suggest that impaired DNA-binding affinities of SALL4^{ZFC4} mutants are likely associated with human cancers.

It has been reported that the N-terminal 12 amino acid stretch of SALL4 interacts with the nucleosome remodeling deacetylase complex that creates repressive chromatin structure in ESCs (25, 26). Here our structural study illustrates module-specific roles of SALL4 in target sequence recognition by ZFC and recruiting nucleosome remodeling deacetylase. In this way, our study not only uncovers the conserved DNA recognition mode by SALL family members but also provides insights into a better understanding how SALL4 mutations result in human cancers *via* altering the expression profile of key regulators such as *Sox1*.

In summary, our study not only provides mechanistic insight into the AT-rich DNA recognition by SALL4^{ZFC4} and SALL4^{ZFC1}, but also uncovers that the binding of SALL4 at specific AT-rich genomic DNA regions influences cell differentiation and cell fate *in vivo*.

Experimental procedures

Cloning, protein expression, and purification

The sequence encoding SALL4^{ZFC4} (residues 856–930) was amplified by PCR from a complementary cDNA library; sequences encoding human SALL4^{ZFC1} (aa 378–453) and SALL3^{ZFC4} (aa 1102–1166) were synthesized by Genscript (Nanjing); the sequence encoding human SALL4⁵⁴⁸⁻¹⁰²⁹, which spans ZFC2 and ZFC4, was synthesized by Sangon Biotech (Shanghai). All of them were cloned into pET28-MHL vector, and the cloned plasmid was transformed into *Escherichia coli* BL21 (DE3). Cells were grown in LB medium at 37 °C until the A₆₀₀ reached ~0.8. The recombinant protein was overexpressed at 16 °C for 18h after induction by 0.2 mM (final concentration) IPTG and 40 μM ZnCl₂ (final concentration). Cells were harvested by centrifuging at 3600g, 4 °C for 15 min and pellets were resuspended in a buffer containing 20 mM

Tris-HCl, pH 7.5, and 400 mM NaCl. Lysates were centrifuged at 10,000g, 4 °C for 30 min and supernatants were collected.

Recombinant SALL4^{ZFC4} was purified by Ni-NTA column (GE healthcare) and eluted by 20 mM Tris-HCl, pH 7.5, 400 mM NaCl, and 500 mM imidazole. N-terminal poly-histidine tags (His-tags) of the recombinant protein were cleaved by Tobacco etch virus protease and dialyzed overnight with the buffer containing 20 mM Tris-HCl, pH7.5, and 150 mM NaCl. SALL4^{ZFC4} was further purified by Superdex 75 gel filtration (GE Healthcare) and HitrapTM S HP column (GE healthcare). The purified protein was concentrated to 8 mg/ml in the buffer containing 20 mM Tris-HCl, pH 7.5, and 150 mM NaCl and was stored at -80 °C before further use.

Expression and purification of SALL4^{ZFC1} and SALL3^{ZFC4} were performed in the same way as for SALL4^{ZFC4}. Site-specific mutations were carried out by using two reverse and complementary primers containing mutated codons. Primer sequences used for cloning mutants are listed in Table S3. All mutants were purified in the same way as for WT proteins.

Isothermal titration calorimetry

ITC experiments were performed on a MicroCal PEAQ-ITC calorimeter (Malvern Panalytical) at 25 °C by titrating 2 µl of protein (1-2 mM) into cell containing 40 µM double strand DNA, with a spacing time of 120 s and a reference power of 5 µCal s⁻¹. The ITC binding assays were performed in a buffer containing 20 mM Tris-HCl, pH 7.5, and 150 mM NaCl. Control experiments were performed by titrating proteins (1-2 mM) into the buffer only, which were subtracted during analysis. Binding isotherms were plotted, analyzed, and fitted by MicroCal PEAQ-ITC Analysis software (Malvern Panalytical). The dissociation constants (K_Ds) were determined from a minimum of two experiments (mean ± SD). Sequences of dsDNAs used for ITC are listed in Table S4.

The SALL4^{ZFC4} N912Q mutant is less stable in 20 mM Tris-HCl, 150 mM NaCl, pH 7.5, so its ITC experiments were carried out by titrating 2 µl of dsDNA (0.7 mM) into cell containing 40 µM protein. Representative ITC-binding curves are shown in Fig. S6.

Crystallization, data collection, and structure determination

All crystals were grown using the sitting drop vapor diffusion method at 18 °C. Before crystallization, protein is mixed with respective dsDNA ligand at a ratio of 1:1. Crystal of SALL4^{ZFC4} in complex with the 12-mer dsDNA (5'-GGTAATATTTCC-3') was obtained by mixing 1.0 µl of complex with 1.0 µl of well solution containing 0.1 M BIS-TRIS, pH 6.5, and 25% (w/v) PEG 3350. Crystal of SALL4^{ZFC4} in complex with the 16-mer dsDNA (5'-GGAA-TATAATATTTCC-3') was obtained by mixing 1.0 µl of complex with 1.0 µl of well solution containing 0.1 M BIS-TRIS, pH 5.5, 0.2 M sodium chloride, and 25% PEG 3350. Crystal of SALL4^{ZFC3} in complex with the 12-mer dsDNA (5'-GGTAATATTTCC-3') was obtained by mixing 1.0 µl of complex with 1.0 µl of well solution containing 0.1 M MES monohydrate, pH 6.5, 0.2 M ammonium sulfate, and 30% w/v

PEG monomethyl ether 5000. Crystal of SALL4^{ZFC1} in complex with the 16-mer dsDNA (5'-GGAAATATAATATTTCC-3') was obtained by mixing 1.0 µl of complex with 1.0 µl of well solution containing 0.1 M HEPES, pH 6.5, 10% PEG 6000, and 5% (v/v) 2-methyl-2,4-pentanediol. Before flash-freezing crystals in liquid nitrogen, crystals were soaked in a cryoprotectant consisting of 85% reservoir solution plus 15% glycerol.

The diffraction data were collected on beam line BL17B and BL18U1 at Shanghai Synchrotron Facility and processed with HKL2000/3000 (27, 28) or XDS software (29). Although the dataset of SALL4^{ZFC4} with 12-mer DNA was collected at 0.978560 Å, zinc single-wavelength anomalous dispersion phasing was successful owing to the good signal strength. The structure of SALL4^{ZFC4} was solved by CRANK2 (30), with 6 Zn²⁺ versus ~1200 atoms. Then, the DNA was built manually by COOT (31), and the complex model was further refined by Phenix (32). The other complexes were solved by molecular replacement using Phaser (33) with previously solved SALL4^{ZFC4} complex as the search model. Then, the models were built and refined manually by COOT (31) and were further refined by Phenix (32). The statistic details about data collection and structure refinement were summarized in Table S2.

Mouse ESC culture

Sall4^{fllox} mouse ESCs were previously described (34). Cells are maintained in the 2i media (35). To generate Sall4^{-/-} cells, Sall4^{fllox} ESCs were suspended by trypsinization and neutralization, then, infected with adenovirus EGFP-Cre (36). Independent clones were isolated, expanded, and the Sall4^{-/-} genotype was confirmed by genomic PCR as previously described (34).

WT mouse Sall4 was cloned in the pLV-EF1a-IRES-Puro vector (37). The Sall4 N922D mutant was generated by site-directed mutagenesis using Q5 High-Fidelity DNA Polymerase (New England Biolabs) and In-Fusion Snap Assembly (Takara Bio USA) following the manufacturer's instructions. Lentiviruses were produced according to a standard procedure (38) and were concentrated using Lenti-X Concentrator (Takara Bio USA). Approximately, 1x 10⁵ Sall4^{-/-} mESCs were infected with lentivirus carrying WT or mutant Sall4 and selected by 2 µg/ml puromycin. Selected cells were expanded and used for experiments.

For qRT-PCR, total RNA was isolated using the Direct-zol RNA MicroPrep kit (Zymo Research), and complementary DNA was synthesized using iScript cDNA synthesis kit (Bio-Rad) according to the manufacturers' instructions. Quantitative PCR was performed using SYBR green master mix (ThermoFisher) and primers in Table S5.

CUT&RUN experiments

CUT & RUN (39) was performed essentially as described in the online protocol (dx.doi.org/10.17504/protocols.io.zcpcf2vn) using Sall4^{fllox/fllox} or Sall4^{-/-} mouse ESCs (10⁵ cells per reaction) cultured in the 2i + LIF media (20). Anti-SALL4 antibody (SC-101147 (EE-30)) or normal rabbit IgG (SC-2025) were each used at a 1:300 dilution. EDTA was excluded from all buffers prior to MNase inactivation to avoid Zn²⁺ chelation. Cell permeabilization and all subsequent steps were performed

Structures of SALL4 bound with AT-rich DNAs

using buffers containing 0.02% digitonin. Recovered DNA fragments were end-repaired, A-tailed, and ligated with xGen adapters (10005974, Integrated DNA Technologies) using the Kapa Hyper Prep Kit (07962312001, Roche) and barcoded during amplification using Kapa HotStart Readymix (7958927001, Roche). Libraries were sequenced using a 2 × 150 paired-end configuration on a HiSeq 4000 (Genewiz). Reads were trimmed using TrimGalore (0.6.0) and CutAdapt (1.18) and read quality was assessed with Fastqc (0.11.8). Trimmed reads were mapped with BWA MEM (0.7.17) using the mouse genome (GRCm38) as reference. Peaks were identified using MACS (2.1.1.20160309) using the $-call-summits -g\ mm$ parameters. Peak lists from each replicate were merged using R (4.1.2) to find high confidence peaks present in both replicates. The 500 bp of sequence flanking the summit of each peak was used for de novo motif analysis with MemeChip (v5.0.1) using $-order\ 2 -meme-p\ 2 -meme-nmotifs\ 10 -psp-gen$ parameters and the DNase accessible regions from ENCODE dataset ENCFF782QYA for the background model.

Data availability

The coordinates and structure factors files for the structures of SALL4^{ZFC4} with 12-mer dsDNA, SALL4^{ZFC4} with 16-mer dsDNA, SALL3^{ZFC4} with 12-mer dsDNA, and SALL4^{ZFC1} with 16-mer dsDNA were deposited into Protein Data Bank under accession codes 7Y3I, 7Y3K, 7Y3L, and 7Y3M, respectively. CUT&RUN data were deposited in the Gene Expression Omnibus with the accession code GSE203303 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE203303>).

Supporting information—This article contains supporting information.

Acknowledgments—We thank the staff from the BL17B/BL18U1/BL19U1/BL19U2/BL01B beamline45 of National Facility for Protein Science in Shanghai at Shanghai Synchrotron Radiation Facility for assistance during data collection. We also thank Dr Brian Hendrich and Dr Ryuichi Nishinakamura for providing mouse ESCs.

Author contributions—W. R. and C. X. methodology; W. R., T. K., and C. X. investigation; W. R., T. K., X. W., Q. G., M. D. G., S. Z., M. M., H. K., D. C., J. Z., Z. Z., X. Y., Y. K., and C. X. formal analysis; W. R., Y. K., and C. X. writing—original draft; W. R., T. K., X. W., Q. G., M. D. G., S. Z., M. M., H. K., D. C., J. Z., Z. Z., X. Y., Y. K., and C. X. writing—review and editing; Y. K. and C. X. supervision.

Funding and additional information—C. X. is supported by the National Natural Science Foundation of China (grant nos. 22137007 and 92053107), “the Fundamental Research Funds for the Central Universities”, the Major/Innovative Program of the Development Foundation of the Hefei Center for Physical Science and Technology (2021HSC-CIP014), and “the Thousand Young Talent program”. Y. K. is supported by a grant from National Institutes of Health of USA (R01AR064195) and Grant-in-Aid of Artistry, Research and Scholarship of the University of Minnesota (#378836).

Conflict of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: ESC, embryonic stem cell; ITC, isothermal titration calorimetry; SALL, Spalt-like transcription factor; TFs, transcription factors; ZFC4, zinc finger cluster 4.

References

1. Spitz, F., and Furlong, E. E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626
2. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., *et al.* (2018) The human transcription factors. *Cell* **172**, 650–665
3. Johnson, P. F., and McKnight, S. L. (1989) Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.* **58**, 799–839
4. Long, H. K., Blackledge, N. P., and Klose, R. J. (2013) ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem. Soc. Trans.* **41**, 727–740
5. Mitchell, P. J., and Tjian, R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378
6. Vazquez, M. E., Caamano, A. M., and Mascarenas, J. L. (2003) From transcription factors to designed sequence-specific DNA-binding peptides. *Chem. Soc. Rev.* **32**, 338–349
7. Deaton, A. M., and Bird, A. (2011) CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022
8. Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., *et al.* (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**, 1082–1086
9. Xu, C., Bian, C., Lam, R., Dong, A., and Min, J. (2011) The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nat. Commun.* **2**, 227
10. Lifton, R. P., Goldberg, M. L., Karp, R. W., and Hogness, D. S. (1978) The organization of the histone genes in *Drosophila melanogaster*: functional and evolutionary implications. *Cold Spring Harb Symp. Quant. Biol.* **42**, 1047–1051
11. Smale, S. T., and Kadonaga, J. T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479
12. Gordon, B. R., Li, Y., Cote, A., Weirauch, M. T., Ding, P., Hughes, T. R., *et al.* (2011) Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 10690–10695
13. Lorch, Y., Maier-Davis, B., and Kornberg, R. D. (2014) Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes Dev.* **28**, 2492–2497
14. Pantier, R., Chhatbar, K., Quante, T., Skourti-Stathaki, K., Cholewa-Waclaw, J., Alston, G., *et al.* (2021) SALL4 controls cell fate in response to DNA base composition. *Mol. Cell* **81**, 845–858.e848
15. Kong, N. R., Bassal, M. A., Tan, H. K., Kurland, J. V., Yong, K. J., Young, J. J., *et al.* (2021) Zinc finger protein SALL4 functions through an AT-rich motif to regulate gene expression. *Cell Rep.* **34**, 108574
16. Alvarez, C., Quiroz, A., Benitez-Riquelme, D., Riffo, E., Castro, A. F., and Pincheira, R. (2021) SALL proteins; common and antagonistic roles in cancer. *Cancers (Basel)* **13**, 6292
17. Kohlhase, J., Schubert, L., Liebers, M., Rauch, A., Becker, K., Mohammed, S. N., *et al.* (2003) Mutations at the SALL4 locus on chromosome 20 result in a range of clinically overlapping phenotypes, including Okhiro syndrome, Holt-Oram syndrome, acro-renal-ocular syndrome, and patients previously reported to represent thalidomide embryopathy. *J. Med. Genet.* **40**, 473–478
18. Kohlhase, J., Chitayat, D., Kotzot, D., Ceylaner, S., Froster, U. G., Fuchs, S., *et al.* (2005) SALL4 mutations in Okhiro syndrome (Duane-radial ray syndrome), acro-renal-ocular syndrome, and related disorders. *Hum. Mutat.* **26**, 176–183

19. Yang, L., Liu, L., Gao, H., Pinnamaneni, J. P., Sanagasetti, D., Singh, V. P., *et al.* (2017) The stem cell factor SALL4 is an essential transcriptional regulator in mixed lineage leukemia-rearranged leukemogenesis. *J. Hematol. Oncol.* **10**, 159
20. Miller, A., Ralser, M., Kloet, S. L., Loos, R., Nishinakamura, R., Bertone, P., *et al.* (2016) Sall4 controls differentiation of pluripotent cells independently of the Nucleosome Remodelling and Deacetylation (NuRD) complex. *Development* **143**, 3074–3084
21. Xu, Y., Xu, C., Kato, A., Tempel, W., Abreu, J. G., Bian, C., *et al.* (2012) Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for *Xenopus* eye and neural development. *Cell* **151**, 1200–1213
22. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921
23. Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., *et al.* (2017) Zinc-finger proteins in health and disease. *Cell Death Discov.* **3**, 17071
24. Shen, A., Higgins, D. E., and Panne, D. (2009) Recognition of AT-rich DNA binding sites by the MogR repressor. *Structure* **17**, 769–777
25. Liu, B. H., Jobichen, C., Chia, C. S. B., Chan, T. H. M., Tang, J. P., Chung, T. X. Y., *et al.* (2018) Targeting cancer addiction for SALL4 by shifting its transcriptome with a pharmacologic peptide. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7119–E7128
26. Hainer, S. J., and Fazzio, T. G. (2015) Regulation of nucleosome architecture and factor binding revealed by nuclease footprinting of the ESC genome. *Cell Rep.* **13**, 61–69
27. Otwinowski, Z., and Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Met. Enzymol.* **276**, 307–326
28. Minor, W., Cymborowski, M., Otwinowski, Z., and Chruszcz, M. (2006) HKL-3000: The integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 859–866
29. Kabsch, W. (2010) Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132
30. Pannu, N. S., Waterreus, W. J., Skubak, P., Sikharulidze, I., Abrahams, J. P., and de Graaff, R. A. (2011) Recent advances in the CRANK software suite for experimental phasing. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 331–337
31. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132
32. Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., *et al.* (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1948–1954
33. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674
34. Sakaki-Yumoto, M., Kobayashi, C., Sato, A., Fujimura, S., Matsumoto, Y., Takasato, M., *et al.* (2006) The murine homolog of SALL4, a causative gene in Okhiro syndrome, is essential for embryonic stem cell proliferation, and cooperates with Sall1 in anorectal, heart, brain and kidney development. *Development* **133**, 3005–3013
35. Mulas, C., Kalkan, T., von Meyenn, F., Leitch, H. G., Nichols, J., and Smith, A. (2019) Defined conditions for propagation and manipulation of mouse embryonic stem cells. *Development* **146**, dev173146
36. Taniguchi, N., Carames, B., Kawakami, Y., Amendt, B. A., Komiya, S., and Lotz, M. (2009) Chromatin protein HMGB2 regulates articular cartilage surface maintenance via beta-catenin pathway. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 16817–16822
37. Hayer, A., Shao, L., Chung, M., Joubert, L. M., Yang, H. W., Tsai, F. C., *et al.* (2016) Engulfed cadherin fingers are polarized junctional structures between collectively migrating endothelial cells. *Nat. Cell Biol.* **18**, 1311–1323
38. Binder, Z. A., Thorne, A. H., Bakas, S., Wileyto, E. P., Bilello, M., Akbari, H., *et al.* (2018) Epidermal growth factor receptor extracellular domain mutations in glioblastoma present opportunities for clinical imaging and therapeutic development. *Cancer Cell* **34**, 163–177.e167
39. Skene, P. J., and Henikoff, S. (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, e21856