Structural analysis of SALL4 zinc-finger domain reveals a link between AT-rich DNA binding and Okihiro syndrome

James A. Watson[1§], Raphaël Pantier[1§], Uma Jayachandran[1], Kashyap Chhatbar[1], Beatrice Alexander-Howden[1], Valdeko Kruusvee[1,2], Michal Prendecki[1,3], Adrian Bird[1], Atlanta G. Cook[1]*

1. Wellcome Centre for Cell Biology, Max Born Crescent, Edinburgh EH9 3BF, United Kingdom
2. Current address: University of Copenhagen, Copenhagen Plant Science Centre, Plant Biochemistry, Thorvaldsensvej 40, 1871 Frederiksberg C, Copenhagen, Denmark
3. Current address: R&D Center in Poznań, Medicofarma Biotech S.A., Wielkopolska Center of Advanced Technologies, 10 Uniwersytetu Poznańskiego St., 61-614 Poznan, Poland

§ These authors contributed equally to the work
* to whom correspondence should be addressed atlanta.cook@ed.ac.uk

Abstract

Spalt-like 4 (SALL4) maintains vertebrate embryonic stem cell identity and is required for the development of multiple organs, including limbs. Mutations in SALL4 are associated with Okihiro syndrome and SALL4 is also a known target of thalidomide. SALL4 protein has a distinct preference for AT-rich sequences, recognised by a pair of zinc fingers at the C-terminus. However, unlike many characterised zinc finger proteins, SALL4 shows flexible recognition with many different combinations of AT-rich sequences being targeted. SALL4 interacts with the NuRD corepressor complex which potentially mediates repression of AT-rich genes. We present a crystal structure of SALL4 C-terminal zinc fingers with an AT-rich DNA sequence, which shows that SALL4 uses small hydrophobic and polar side chains to provide flexible recognition in the major groove. Missense mutations reported in patients that lie within the C-terminal zinc fingers reduced overall binding to DNA but not the preference for AT-rich sequences. Furthermore, these mutations altered association of SALL4 with AT-rich genomic sites, providing evidence that these mutations are likely pathogenic.

Introduction

Embryonic stem cells (ESCs) balance pluripotency with a development and differentiation program to generate distinct tissues within an organised body plan. Proteins involved in development are typically expressed transiently, at specific embryonic locations, and are absent from adult tissues or restricted to specific tissue progenitor cells. SALL4 is a protein of this type which is expressed both in ESCs and in later lineages during embryogenesis and plays critical roles in the development of various organs (Sweetman and Munsterberg 2006). It is one of four Spalt-like C2H2 zinc finger DNA binding proteins in mouse and human. SALL4 deficiency leads to peri-implantation lethality in mice (Sakaki-Yumoto, Kobayashi et al. 2006) and increased neuronal differentiation potential in mouse embryonic stem cells (ESCs) (Miller, Ralser et al. 2016), indicating that SALL4 helps to maintain stem cell identity. Heterozygous *SALL4* mutation in mice causes defects in multiple organs including the nervous system, limbs, kidneys, heart and anorectal tract (Koshiba-Takeuchi, Takeuchi et al. 2006, Sakaki-Yumoto, Kobayashi et al. 2006). Consistent with the phenotypes of SALL4 haploinsufficiency in mice, patients with Okihiro syndrome, an autosomal dominant disorder caused by mutations in SALL4, also present a range of symptoms including limb defects, eye anomalies (Duane syndrome), vertebral malformations, hearing loss, kidney defects, heart anomalies and anal stenosis (Al-Baradie, Yamada et al. 2002, Kohlhase, Heinrich et al. 2002). Some Okihiro syndrome patients have a presentation similar to thalidomide embryopathy (Kohlhase, Schubert et al. 2003); consistent with this, SALL4 is a cellular target of thalidomide, which facilitates binding of SALL4 to the $CLR4^{CRBN}$ E3 ubiquitin ligase that ubiquitylates SALL4 and leads to its destruction (Donovan, An et al. 2018, Matyskiela, Couto et al. 2018, Matyskiela, Clayton et al. 2020).

SALL4 contains 7 zinc fingers (Znfs) arranged in three clusters (Fig. 1a). Sequence comparisons suggest that these are closely related to zinc finger clusters (ZFCs) 1, 2 and 4 of SALL1 (Sweetman and Munsterberg 2006). ZFC4 of SALL4 is required for its localization to mouse heterochromatin (Sakaki-Yumoto, Kobayashi et al. 2006) and has a strong preference for a range of AT-rich DNA sequence motifs (Kong, Bassal et al. 2021, Pantier, Chhatbar et al. 2021). The molecular basis of this broad specificity is unknown, but there is evidence that it is essential for the ability of SALL4 to maintain stemness in ESCs by sensing differences in sequence composition in the genome (Pantier, Chhatbar et al. 2021). Importantly, discrete mutation of ZFC4 leads to precocious ESC differentiation and embryonic lethality, phenocopying complete loss of SALL4 (Sakaki-Yumoto, Kobayashi et al. 2006). To gain insight into how SALL4 selects AT-rich sequences, and the likely effect of missense mutations on DNA binding, we undertook a structural, biochemical and cell-based analysis of the ZFC4. We solved the X-ray crystal structure of SALL4 ZFC4 with an AT-rich sequence motif to gain

insight into this broad sequence specificity. We also characterised two patient missense mutations that are likely to be deleterious and causative of Okihiro syndrome. We show that these mutations reduce SALL4 ZFC4 binding to AT-rich DNA, yet the proteins retain preference for AT-rich sequences. In cells, full length mutant proteins fail to localise to heterochromatin. These results confirm that SALL4 binding to AT-rich sequences is fundamental to its *in vivo* function and that disruptions to this interaction contribute to disease presentation.

Results

ZFC4 domain is depleted of population missense variants

According to the gnomAD database (Karczewski, Francioli et al. 2020) the loss-of-function observed/expected upper bound fraction (LOEUF) indicates that SALL4 is depleted of inactivating variants and under purifying selection (LOEUF=0.101). This is consistent with the finding that SALL4 haploinsufficiency is responsible for an autosomal dominant disorder. To further understand the contribution of different SALL4 domains to function, we extracted population missense mutations and calculated an overall missense depletion score for SALL4 protein (Vp) of 0.38 (Deak and Cook 2022) (Fig. 1a). We then considered individual domains of SALL4 and calculated missense depletion relative to the whole protein (VdVp ratio), where a score of ≥1 would indicate that a single domain is not depleted of missense variants compared to the full protein sequence (Deak and Cook 2022). Three regions were observed to be comparatively depleted of population missense mutations: the N-terminal NuRD binding motif (Lauberth and Rauchman 2006) (VdVp=0.22); a glutamine-rich (Q-rich) sequence that has been reported to participate in SALL protein homo- and heterodimer formation (VdVp=0.50) (Sweetman, Smith et al. 2003); and ZFC4, which is essential for SALL4 function in mice (VdVp=0.47) (Pantier, Chhatbar et al. 2021)(Fig. 1a). Missense depletion of these regions indicates that they are likely to contribute to the essential functions of SALL4. The two other zinc finger domain regions, ZFC1 and ZFC2 are less depleted (VdVp=0.78 and 0.89 respectively). Indeed, the gnomAD database, which excludes individuals with severe pathological symptoms compared with the general population, reveals mutations in ZFC1 and ZFC2 (residues 382-432 and 566-648, respectively) that alter the cysteine and histidine residues that are essential for zinc finger integrity (C387Y: Znf1; C412S: Znf2; H644D: Znf5; H644L: Znf5). The absence of both these domains in a shorter splice variant of SALL4 is also consistent with a specialised role for these zinc finger clusters in SALL4 function.

A number of likely pathogenic mutations have been reported for SALL4 in Okihiro syndrome patients (Kohlhase, Heinrich et al. 2002, Kohlhase, Schubert et al. 2003, Borozdin, Boehm et al. 2004, Borozdin, Wright et al. 2004, Kohlhase, Chitayat et al. 2005, Diehl, Mu et al. 2015). We searched the literature and the ClinVar database for missense variants affecting SALL4 as these can inform on functional regions within SALL4 (Landrum, Lee et al. 2018). Of 41 listed variants of "uncertain significance", 29 are present in gnomAD, which excludes pathogenic mutations, and are therefore unlikely to cause disease. The only missense variant with clear evidence for pathogenicity is H888R, which is within ZFC4 (Miertus, Borozdin et al. 2006). We noted two other missense mutations mapping to ZFC4 (R890W and G911D), both of which are absent in gnomAD and may therefore be pathogenic (Diehl, Mu et al. 2015). As would be expected for pathogenic mutations, these three missense variants are found in a region of ZFC4 that is depleted of population variants (Fig. 1a).

SALL4 ZFC4 binds to AT-rich sequences using polar interactions

To gain insight into SALL4 recognition of DNA, a construct of mouse SALL4 ZFC4 (residues 870-940) was co-crystallised in the presence of a palindromic AT-rich DNA sequence (Fig. 1b). This sequence was based on a motif, ATATT that was most enriched by SALL4 on systematic evolution of ligands by exponential enrichment (SELEX) (Pantier, Chhatbar et al. 2021). Long, needle-like crystals were

grown that diffracted to 2.76 Å, with high anisotropy in the diffraction pattern and *P*1 symmetry. After data reduction, a theoretical model of B-form DNA was used to search for a molecular replacement solution and four molecules of dsDNA were fitted into the asymmetric unit of the crystal. Subsequently, individual zinc fingers were found, using iterative searches with a model based on PRDM9 (Patel, Zhang et al. 2017), to complete the asymmetric unit with four copies of ZFC4. Although the stoichiometry of the asymmetric unit is 1:1 for ZFC4 to dsDNA, the ZFC4 chains are not evenly distributed among the DNA molecules, with one copy of the dsDNA lacking any associated protein and one dsDNA binding simultaneously to two ZFC4 chains. The structures were completed through iterative model building and refinement and have good stereochemistry and final R_{work}/R_{free} values of 24.7% and 25.4%, respectively (Table S1). All DNA bases are visible in the map. For all ZFC4 chains, residues 880 to 930 were visible, with chain L extending from 878 to 933. We base our description on this chain (Fig. 1c). Root mean square deviation (r.m.s.d.) values for Cα superposition of each of the protein chains ranged from 0.57-0.78 Å, indicating a high level of similarity between all four copies in the asymmetric unit (Fig. S1a). Comparison of the refined dsDNA structure with ideal B-form DNA showed that ZFC4-bound DNA has a compressed minor groove and a slightly expanded major groove (Fig. S1b)

We measured SAXS scattering curves for ZFC4 alone, dsDNA alone and ZFC4 DNA complexes as they eluted from size exclusion chromatography (Fig. S1c, Table S2). Scattering curves and maximum dimensions of ZFC4, DNA and the complex were highly consistent with models and measurements from the crystal structure (Fig. S1d,e). A bead model calculated from real space analysis of the ZFC4-DNA complex was consistent with a primarily 1:1 protein:DNA stoichiometry in solution (Fig. 1d). Kratky analysis of these data indicate that ZFC4 is highly dynamic in solution and becomes ordered on binding to dsDNA (Fig. S1f).

Overall, the structure of ZFC4 bound to dsDNA resembles that of other C2H2 zinc finger pairs bound to DNA (Wolfe, Nekludova et al. 2000)(Fig. 1c). The helix of each zinc finger probes the major groove of the DNA (Fig. 1c-d, Fig. 2). The orientation of the Znf6 to Znf7 is similar to that of zinc finger pairs in Zif268 (Elrod-Erickson, Rould et al. 1996), indicating that SALL4 ZFC4 belongs to a mode I binding orientation (Garton, Najafabadi et al. 2015). We use a common numbering scheme for DNA interacting residues where position 1 is the first residue of the helix and position 7 is the first histidine side chain that interacts with the zinc ion (Wolfe, Nekludova et al. 2000)(Fig. 2a). Mode I orientations are promoted by interactions between the residue in position 9 of the first zinc finger with residue in position -2 of the second zinc finger. In ZFC4, these residues are R900 (R890 in human) and T918 (T908 in human) respectively (Fig. S2a); mode I zinc finger pairs typically have arginine and serine residues at these positions but many sequence pairs can be accommodated (Garton, Najafabadi et al. 2015).

Small polar side chains allow ZFC4 to recognise AT-rich sequences

SALL4 differs from other zinc finger proteins in that it recognises a wide variety of AT-rich sequences, rather than a fixed DNA sequence. Binding affinity is relatively low, in the micromolar range (Kong, Bassal et al. 2021). Of note, residues that make up the SALL4 binding interface are predominantly small and polar or hydrophobic (Fig. 2b). Specificity in zinc finger proteins is typically conferred by interactions of residues at positions +2, +3 and +6 with bases on the DNA strand that runs 3'->5', with additional contributions from the residue in the -1 position, that can interact with bases on the forward strand. Water molecules typically contribute to base recognition but, given the limited resolution of our structure, we were not able to fit structured water at the interface. In our co-crystal structure of SALL4 and DNA, the -1 and +2 positions of Znf6 (S891 and S893) are <4Å from N7 and O6 of G1 on strand A, suggesting that these residues may anchor the protein at the beginning of the palindromic sequence through polar interactions (Fig. 2b,c). Were the sequence of strand A to

start with an adenine base, similar interactions could be made with N6 and N7. The residue at position +3 (A894) does not make direct contact to DNA but allows a close approach of the -2 position (S890), which is 3.9Å from the methyl group of T11 on strand B (Fig. 2b,c). If this base were an adenine, a hydrogen bond could be formed with N7, suggesting sequence flexibility at this site.

A10, the next base along strand B, does not interact with SALL4. However, there is a close contact of SALL4 with its base-pair partner, T3 on strand A. The Cα atom of G921, which is at the +2 position in Znf7, is 3.5Å from the methyl group of T3, suggesting that the presence of a small residue is required for the close approach to this base. T9 is the next base on strand B to be directly recognised, interacting with both the +6 residue (I897) of Znf6 and the -1 residue for Znf7 (T919) (Fig. 2b,d). T919 forms a hydrogen bond with O4 of T9. This suggests that Znf6 provides a direct readout for at least one thymine base through hydrophobic and polar interactions. The next base, A8, is also directly read out by SALL4, via a bidentate hydrogen bond with N922, the +3 position of Znf7. Minor adjustments in position of N922 could allow a hydrogen bond to form with A7, the following base on strand B (Fig. 2b,d; Fig. S2b). T6 then interacts with small hydrophobic residue, V925, at the +6 position of Znf7 (Fig. 2b,e). This interaction is similar to the interaction observed between T9 and I897, suggesting that the small hydrophobic side chain provides a good environment for the methyl group of a thymine base. This series of interactions suggest a preference for a core 5'-TAT-3' sequence along the A strand (equating to A10-T9-A8 on the B strand), but that alternative interactions with AT and TA base pairs could be accommodated before, within and after this core sequence. The combination of small polar and hydrophobic residues provides an interface where the methyl groups of T bases are accommodated but that can also allow for alternative base interactions.

Previous studies by Garton and colleagues noted that sequence preferences for individual bases is influenced by the relative orientation, or binding mode, of pairs of zinc fingers (Garton, Najafabadi et al. 2015). In a large-scale analysis of different possible sequence preferences, they observed that when position 6 is occupied by valine, an A base is typically specified. This fits our observation of V925 interacting with T6, to specify A7 on the forward strand, and I897 interacting with T9, specifying A4 on the forward strand. This study also indicated that when position +2 is occupied by alanine or serine residues, A or A/T preferences are likely to be observed. SALL4 has S893 and G921 at these positions in Znf6 and Znf7 respectively. In contrast, asparagine at position +3 is normally associated with a C base, whereas we see direct interaction of the +3 residue N922 with A7, specifying a T on the complementary strand.

The sequence of SALL4 ZFC4 is conserved across vertebrates and all residues that interact with DNA are identical across species (Fig. S2c). Furthermore, the ZFC4 sequence is highly conserved with equivalent sequences in SALL1 and SALL3 across the same group of organisms and with Drosophila Salr (Fig. S2c). Only one residue differs between the ZFC4 domains of SALL4 and SALL1, A892 (mouse), which points away from the DNA binding site. This suggests that SALL1 and SALL3 have an identical DNA binding specificity to SALL4.

Patient mutations likely interfere with DNA binding

Three patient mutations that affect conserved residues of ZFC4 were modelled into the structure to assess their likely impact on SALL4 function (Fig. 1a). H888R (mouse equivalent is H898R), is the only established pathogenic missense mutation (Miertus, Borozdin et al. 2006). While this mutation was proposed to enhance DNA binding, we conclude that this change alters a histidine ligand of the Znf6 zinc ion and so is highly likely to disrupt the fold of Znf6, preventing DNA binding (Fig. 3a). R890W (R900W in mouse) is noted in ClinVar (VCV000850032.2), with uncertain significance. Extending from the helix of Znf6, R900 forms bridging interactions with Znf7 via a backbone interaction with T918

(Fig. S2a). Furthermore, it forms a closely packed network of interactions with residues of the loop connecting Znf6 and Znf7, including residues E905 and P907. Residues equivalent to R900 (position 9 on Znf6) in other zinc finger proteins play an important role in defining the relative orientation of one zinc finger with respect to the next by interacting with the conserved TGEKP sequence that connects Znf6 and Znf7. Mutation of the TGEKP connector sequence typically affects DNA binding affinity (Wolfe, Nekludova et al. 2000). Mutation of R900 to tryptophan is likely to disrupt the network of close contacts between zinc fingers and could alter the orientation of Znf7 with respect to Znf6 (Fig. 3a). Given that the angle between the domains impacts on its ability to bind DNA, this mutation is likely to reduce binding to DNA. A third mutation, G911D (G921D in mouse), places a larger, negatively-charged side chain at the beginning of the Znf7 helix. G921 mediates close contacts to the major groove (Fig. 2e). An aspartate side chain at this position is likely to clash with the DNA bases, potentially altering the overall angle with which Znf7 binds DNA. This mutation is also likely to be pathogenic (Diehl et al, 2015) (Fig. 3b).

Patient mutations in SALL4 ZFC4 disrupt dsDNA binding *in vitro* and in cells

To test whether uncharacterised patient mutations do indeed alter DNA binding, we purified ZFC4 fragments with mutations R900W and G921D (Fig 4a, Fig. S3a,b). Given that H898R is likely to disrupt the fold of the protein, we did not pursue characterisation of this mutation *in vitro*. Electrophoretic mobility shift assays (EMSA) of these proteins showed that both point mutations substantially reduce binding to an AT-rich DNA probe that otherwise binds efficiently to wild type protein (Fig. 4b).

To assess the impact of ZFC4 mutation on DNA binding in cells, full-length mouse SALL4 cDNA carrying the wild-type sequence or patient missense mutations (H898R, R900W, G921D) were cloned into a mammalian expression vector (Chambers, Colby et al. 2003) (Fig. S4a). Mouse embryonic fibroblasts (NIH 3T3 cells) were chosen for transfection as they lack expression of endogenous SALL4 and SALL1 and present large nuclear foci with intense DAPI signal, corresponding to AT-rich pericentric heterochromatin (Fig. S4b). Strikingly, all mutant proteins showed a diffuse nuclear signal while SALL4 wild-type co-localised with DAPI bright spots (Fig. 4c; Fig. S4c). This observation, along with EMSA data, demonstrates that mutating single residues within ZFC4 is sufficient to disrupt SALL4 binding to AT-rich DNA. Of note, R900W and G921D show similar effects to H898R, indicating that both of these point mutations have an impact on binding equivalent to disrupting the protein fold.

The observations above could either be explained by an overall loss of DNA binding affinity or by a loss of specificity for AT-rich sequences. To investigate whether mutations induced a change in sequence specificity, we performed systematic evolution of ligands by exponential enrichment (SELEX) coupled with high-throughput sequencing (HT-SELEX) (Jolma, Kivioja et al. 2010, Pantier, Chhatbar et al. 2021)(Pantier, Chhatbar et al. 2022, accepted) (Fig. 5a). ZFC4 wild-type and mutant (R900W, G921D) proteins were purified and submitted to HT-SELEX, together with a negative control (no protein) to account for PCR bias during the protocol. All possible 6-mer motifs were divided into different categories depending on their proportion of A/T nucleotides. Their relative enrichment was compared across samples at cycle 1, 3 and 6 of HT-SELEX (Fig. 5b; Supplementary data file 1). This analysis revealed that ZFC4 wild-type and both mutant proteins preferentially bind to a large number of AT-rich motifs. However, the level of enrichment was much higher for ZFC4 wild-type compared to R900W and G921D proteins (Fig. 5b). This observation indicates that ZFC4 mutants present decreased DNA binding affinity, in agreement with EMSA data (Fig. 4b). The majority of enriched DNA motifs by HT-SELEX were shared between ZFC4 wild-type and mutants, indicating conserved sequence specificity (Fig. 5c). As expected, the top motifs were exclusively composed of A and T nucleotides (Fig. S5a). Interestingly, the enrichment of DNA motifs correlated better with the

total number of A/T nucleotides within a 6bp motif rather than the number of consecutive A/T nucleotides (Fig. S5b). This indicates that A/T base composition is a critical parameter for DNA binding, and that SALL4 ZFC4 can tolerate the presence of a single G or C nucleotide within its binding site. Overall, mutations in ZFC4 (R900W, G921D) dramatically reduced DNA binding without affecting preference for AT-rich motifs.

Discussion

SALL4 is an unusual example of a zinc finger protein that has an expanded specificity for a range of AT-rich sequences. Our structure of SALL4 with an AT-rich DNA sequence shows that SALL4 ZFC4 makes close contacts to bases in the major groove primarily mediated by small hydrophobic or polar side chains that allow hydrogen bonding interactions. Two small aliphatic residues, I897 and V925 provide hydrophobic surfaces that interact with methyl groups at C5 on thymine that point into the major groove. These hydrophobic surfaces likely provide some specificity of thymine bases within SALL4-associated sequences. The small size and non-charged nature of DNA binding residues in SALL4 allow a close association of the zinc fingers to the major groove with a concomitant narrowing of the minor groove.

Previously, we showed that a double point mutant of SALL4 (T919D, N922A) had a cellular phenotype equivalent to deletion of ZFC4 (Pantier, Chhatbar et al. 2021) (Fig. S2c). Our structure reveals that these two residues indeed play important roles in DNA recognition, as predicted (Fig. 2d). Our SELEX data indicate that more than one G/C base pair is not well tolerated within SALL4 binding sites (Fig. S5b). It is possible that G/C base pairs are selected against because they are more polar than A/T base pairs or that A/T base pairs permit more compression of the minor groove.

Residues that contact DNA in our structure are highly conserved among SALL4 proteins (Fig. S2c, (Pantier, Chhatbar et al. 2021)). The highly similar sequences and expression profiles of SALL4 and SALL1 suggest some functional redundancy. Like SALL4, SALL1 protein is expressed in ESCs, is targeted to heterochromatin and forms homo- and heterodimers with SALL4 (Yamashita, Sato et al. 2007, Rao, Zhen et al. 2010). Indeed, genetic deletion of both *Sall4* and *Sall1* results in stronger phenotypes than either single mutation both in ESCs and mice (Sakaki-Yumoto, Kobayashi et al. 2006, Miller, Ralser et al. 2016). Okihiro syndrome has an overlap in presentation with Townes-Brocks syndrome, which is caused by mutations in the *SALL1* gene (Kohlhase, Heinrich et al. 2002), further indicating that these two proteins have overlapping functions.

The majority of patient mutations described for SALL4 are nonsense or insertion/deletion mutations that are likely to cause loss of function of the gene, with consequent haploinsufficiency. The effects of missense mutations are less clear. We noted two less-characterised patient missense mutations, along with H888R, that map to ZFC4 in a region highly depleted of population missense variants. Our previous work has established that specific disruption of ZFC4 in an otherwise intact SALL4 protein leads to embryonic lethality in mice, demonstrating the importance of this DNA binding domain (Pantier, Chhatbar et al. 2021). Our biochemical and cellular characterisation of SALL4 ZFC4 missense mutations showed disrupted DNA binding *in vitro* and in cells. While H888R was already linked with Okihiro syndrome (Miertus, Borozdin et al. 2006), our study provides experimental evidence that G911D (Diehl, Mu et al. 2015) and R890W (ClinVar, VCV000850032.2) are also likely to be disease-causing mutations.

Our HT-SELEX analysis on 6-mer motifs (based on coverage of the major groove by ZFC4 in the crystal structure) are similar to our previous study on 5bp motifs (Pantier, Chhatbar et al. 2021).

SALL4 ZFC4 binds to a wide range of AT-rich DNA motifs, potentially allowing the protein to "read" DNA base composition. Interestingly, while the point mutations reduce binding to DNA, the proteins still retain AT-rich specificity. In the case of G921D, this is likely to be because only Znf7 is affected by the mutation and some specificity will be retained from Znf6. In the case of R900W, the prediction is that the orientation between the zinc fingers is likely to be altered. However, each individual zinc finger is still likely to be able to interact with DNA. This suggests that the loss of affinity is likely to be because the two zinc fingers cannot optimally interact with DNA at the same time.

Overall, our structural, biochemical and cell-based data show that ZFC4 presents a highly conserved binding interface with DNA. The hydrophobic and polar residues that make up this interface likely provide a flexible interface that allows optimal interaction with methyl groups from thymine residues. Patient missense mutations that alter DNA binding have a major impact on SALL4 localisation in cells even though a preference for AT-rich sequences is retained. This suggests that the DNA binding affinity of SALL4 plays an important role in determining protein localisation and transcriptional silencing in cells.

Materials and Methods

Primary sequence analysis

Domain boundaries of SALL4 were identified based on UniProt annotations and previous sequence analyses (Pantier, Chhatbar et al. 2021). Missense mutations from gnomAD were processed using 1D-to-3D and VdVp_calculator scripts (Deak and Cook 2022) and plotted using Plot Protein (Turner 2013).

SALL4 ZFC4 cloning and purification

Mouse SALL4 (Q8BX22-1) coding sequence (encompassing codons of residues 870-940) was cloned into a pET-based expression vector as a hexahistidine-glutathione S-transferase (GST) tagged fusion protein. Point mutants were introduced using whole plasmid amplification with Pfu Ultra II (Agilent, 600670-61) and complementary primers, followed by DpnI digestion, transformation, plasmid preparation and sequencing. These constructs were expressed in BL21 (*DE3*) cells and induced overnight at 20°C with 1 mM IPTG. Cells were lysed using a cell disruptor (Constant Systems) in a buffer containing 20 mM Tris.HCl pH 7.5, 200 mM NaCl, 0.5 mM β-mercaptoethanol with protease inhibitor cocktail (Roche) and DNase I (Sigma). The clarified lysate was allowed to bind in batch to GSH resin (Cytiva) and eluted using lysis buffer containing 20 mM reduced glutathione. The GST tag from the eluted proteins were cleaved using rhinovirus 3C protease during dialysis (20mM Tris.HCl pH 7.5, 50 mM NaCl, 0.5 mM β-mercaptoethanol). The cleaved proteins were then purified on a 6mL Resource S (Cytiva) ion exchange column and the proteins were eluted using a salt gradient ranging from 50-1000 mM NaCl. The eluted proteins were then further purified by size exclusion chromatography (Superdex S75, Cytiva) in 20 mM Tris.HCl pH 7.5, 200 mM NaCl.

For expression in mammalian cells, mouse SALL4 coding sequence was subcloned into pPYCAG expression plasmids carrying a constitutive CAG promoter (Chambers, Colby et al. 2003). Equivalent ZFC4 patient mutations were introduced by subcloning mutations from expression plasmids and incorporation using Gibson assembly (NEBuilder HiFi E2621S, NEB). Plasmids are available upon request.

Crystallisation and structure solution

An equimolar mixture of SALL4 protein with palindromic oligonucleotide (5' GATATTAATATC 3') was set up (18 nmol + 18 nmol), giving a final protein concentration of 1.9 mg/ml. The complex was crystallised in 50 mM MES pH 6.0, 20 % PEG 3350, 60 mM $MgCl_2$. Cryoprotectant solution was made by supplementing well buffer with 30% glycerol and added to the drops prior to harvesting and flash cooling crystals in liquid nitrogen. Data were collected at Diamond Light Source beamline i04. Data were reduced using AUTOPROC with anisotropy correction done by STARANISO (Vonrhein, Flensburg et al. 2011, Tickle and Vonrhein 2018). Molecular replacement was carried out using calculated models of B form DNA (COOT (Emsley and Cowtan 2004, Emsley, Lohkamp et al. 2010)) in PHASER (McCoy, Grosse-Kunstleve et al. 2007), followed by a search model based on PDBid 5v3g (Patel, Zhang et al. 2017) and prepared using CHAINSAW (Stein 2008). The structure was refined in PHENIX with rebuilding in COOT (Emsley and Cowtan 2004, Emsley, Lohkamp et al. 2010). Validation was carried out using MOLPROBITY (Chen, Arendall et al. 2010) and figures were generated using PyMOL (Schrodinger 2015).

Electrophoretic Mobility Shift Assay

Purified SALL4 wild-type and mutant untagged proteins were incubated with increasing concentrations of 0.5 µM DY681-labelled dsDNA in a buffer containing 20 mM HEPES, pH 7.5, 150 mM potassium acetate and 5 mM magnesium acetate. The total reaction volume of 10 µl was incubated on ice for 30 minutes, after which 2 µl of native loading buffer (0.25% bromophenol blue and 50% glycerol) was added. 4 µl of the reaction was loaded on an 8% native polyacrylamide gel and separated at 2W at 4°C in TBE buffer. After an hour, the gel was scanned at 700nm using LICOR-Odyssey infrared scanner. The images were converted to greyscale using LICOR Image Studio Software.

SAXS

SEC-SAXS experiments were performed at Diamond Light Source on the B21 beamline. Samples at 5-7 mg/ml were injected onto a Superdex S200 Increase 3.2/300 size-exclusion chromatography column in 20 mM Tris, pH 7.5, 200 mM NaCl at 0.1 ml/min. SAXS data were recorded using a 3 s exposure. The ATSAS 3.0.5 suite of software was used for processing data (Manalastas-Cantos, Konarev et al. 2021). CHROMIXS was used for frame selection and sample-solvent subtraction (Panjkovich and Svergun 2018). Guinier and distance distribution analyses were carried out using PRIMUS (Konarev, Volkov et al. 2003). *Ab initio* bead models were generated with DAMMIF launched from within PRIMUS (Franke and Svergun 2009). 15 Å density maps were generated from each bead model and the corresponding crystal structures docked into this density using ChimeraX (Pettersen, Goddard et al. 2021). Additional residues were modelled onto the crystal structure using COOT to match the whole complex used in the SAXS experiment. These models were also fitted to the experimental SAXS data using CRYSOL (launched from PRIMUS) (Svergun, Barberato et al. 1995).

Cell culture

Mouse embryonic stem cells (ESCs) (Hooper, Hardy et al. 1987) were grown in Glasgow Minimum Essential Medium (GMEM; Thermo Fisher Scientific cat. 11710035) supplemented with 15% fetal bovine serum (batch tested), 1x L-glutamine (Thermo Fisher Scientific cat. 25030024), 1x MEM non-essential amino acids (Thermo Fisher Scientific cat. 11140035), 1mM sodium pyruvate (Thermo Fisher Scientific cat. 11360039), 0.1mM 2-mercaptoethanol (Thermo Fisher Scientific cat. 31350010) and 100U/ml leukemia inhibitory factor (LIF, batch tested). NIH 3T3 mouse fibroblasts (ECACC, 93061524) were grown in Dulbecco's Modified Eagle Medium (DMEM, Thermo Fisher Scientific cat. 41966) supplemented with 10% fetal bovine serum. All cell lines were incubated in gelatin-coated dishes at 37°C and 5% CO_2.

For immunofluorescence, 1.2 x10^4 cells were seeded in gelatinised chambered coverslips (Ibidi cat. 80286). Cells were transfected with 2 µg of SALL4 expression plasmid (pPYCAG-Sall4 WT/H898R/R900W/G921D) using the Lipofectamine 3000 reagent (Thermo Fisher Scientific cat. L3000008) and following manufacturer's instructions.

Immunofluorescence

One day after transfection, cells were washed with PBS and fixed for 10 min at room temperature with a 4 % (w/v) paraformaldehyde solution. After fixation, cells were washed with PBS and permeabilised for 10min at room temperature in PBS supplemented with 0.3 % (v/v) Triton X-100. Samples were incubated for 2 h at room temperature in blocking buffer: PBS supplemented with 0.1 % (v/v) Triton X-100, 1 % (w/v) Bovine serum albumin and 3 % (v/v) goat serum (Merck Life Science cat. G9023). Following blocking, samples were incubated overnight at 4°C (with gentle mixing) with primary antibodies diluted at the appropriate concentration in blocking buffer (Table S3). After 4x washes in PBS supplemented with 0.1 % (v/v) Triton X-100, samples were incubated for 2 h at room

temperature (in the dark) with secondary antibodies conjugated with Alexa Fluor Plus dyes (Thermo Fisher Scientific cat. A32723 or cat. A32733) diluted 1:500 in blocking buffer. Cells were washed 4x times with PBS supplemented with 0.1 % (v/v) Triton X-100. DNA was stained with 4',6-diamidino-2-phenylindole (DAPI) for 5 min at room temperature, and cells were washed a final time with PBS. Samples were mounted on coverslips using the ProLong glass mounting medium (Thermo Fisher Scientific cat. P36980), following manufacturer's instructions. Samples were imaged using the Zeiss LSM 880 microscope with Airyscan. Images were analysed and processed using the software Fiji. For each transfection experiment, all SALL4 positive-cells were counted and categorised according to their nuclear expression pattern (foci or diffuse signal).

HT-SELEX

SELEX coupled with high-throughput sequencing (HT-SELEX) was performed as previously described (Pantier, Chhatbar et al. 2021)(Pantier, Chhatbar et al, 2022 accepted), in triplicate experiments. Oligonucleotides were ordered from Integrated DNA Technologies (IDT). Throughout the protocol, SELEX libraries were amplified by PCR using the high-fidelity Phusion DNA polymerase (NEB cat. M0530L) and purified using the MinElute PCR Purification Kit (Qiagen cat. 28004). Purified, recombinant SALL4 ZFC4 WT, R900W and G921D (residues 870-940) were used in SELEX reactions. SELEX libraries (1.5 µg for the first cycle, 200 ng for subsequent cycles) were mixed with 1 µg of recombinant ZFC4 wild-type or mutant proteins in 100µl of SELEX buffer (50 mM NaCl, 1 mM $MgCl_2$, 0.5 mM EDTA, 10 mM Tris.HCl pH7.5, 4 % glycerol) freshly supplemented with 5µg/ml Poly(dI-dC) (Merck Life Science cat. P4929) and 0.5 mM DTT. A negative control experiment (without addition of proteins) was also performed to control for technical bias during the SELEX protocol. Following a 10 min incubation at room temperature, 50 µl of Ni^{2+} Sepharose 6 Fast Flow beads (Cytiva cat. 17531806), previously equilibrated in SELEX buffer, was added to each sample to capture protein-DNA complexes. Following a 20 min incubation at room temperature, beads were washed 5 times with 1ml of SELEX buffer to remove unbound oligonucleotides. After the final wash, beads were resuspended in 100 µl H_2O and used directly for PCR amplification. For each SELEX sample, optimal PCR conditions were empirically determined by running several times the same PCR reaction with increasing numbers of cycles. Amplified and purified SELEX libraries were used as input for subsequent rounds of SELEX, up to 6x cycles. To generate samples for high-throughput sequencing, SELEX libraries were amplified using primers containing Illumina adapters and unique indexes. HT-SELEX libraries were pooled in equimolar amounts and contaminating primers were eliminated by performing a clean-up with KAPA Pure beads (Roche cat. 07983271001), using a 3x beads-to-sample ratio. The HT-SELEX library pool was submitted to high-throughput sequencing using the Illumina MiSeq platform (EMBL GeneCore facility, Germany).

HT-SELEX Analysis

All possible canonical k-mer sequences (k=6) were searched individually in SELEX libraries at different cycles using eme_selex (Pantier, Chhatbar et al, 2022, accepted). A canonical sequence of a k-mer pair is the lexicographically smaller of the two reverse complementary sequences. For every k-mer, the number of reads containing the k-mer is normalised by the total number of reads in the library to generate a fraction. To quantify the abundance of the k-mer, fold change of fraction at higher SELEX cycle(s) vs fraction at initial random library (cycle 0) is calculated. This fold change (vs cycle 0) is visualised for k-mers grouped according to the total number of A/Ts and consecutive number of A/Ts. Top 100 abundant canonical k-mers from ZFC4 wild-type and mutant HT-SELEX experiments at SELEX cycle 6 are used to visualise the overlap using Venn diagram. Top 9 abundant k-mers from ZFC4 wild-type SELEX library at cycle 6 are searched allowing one mismatch and a position frequency matrix (PFM) is generated. Subsequently, the PFM is used to visualise the motif logos. Raw and processed HT-SELEX data is deposited in the ArrayExpress database at EMBL-EBI

(www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-11519. Source code to reproduce the analysis is available at https://eme-selex.readthedocs.io.

Acknowledgements

We thank David Kelly and the COIL facility for microscopy support, Vladimir Benes (EMBL GeneCore facility, Germany) for support with high-throughput sequencing. We thank Diamond Light Source i04 and B21 staff for assistance with data collection. Coordinates are deposited in the PDB with code 8A4I; SASDB accession codes are pending. This work utilised the Edinburgh Protein Production Facility (EPPF) and the Centre Optical Instrumentation Laboratory funded by Wellcome Core Grants 092076 and 203149 to the Centre for Cell Biology. AGC is supported by a Wellcome Senior Fellowship (200898). AB holds a Wellcome Investigator Award (107930) and a European Research Council Advanced grant (EC 694295 *Gen-Epix*) and is a member of the Simons Initiative for the Developing Brain, University of Edinburgh.

Author Contributions

JAW found and optimised crystal conditions for the palindromic DNA partner, carried out primary sequence analysis and SAXS measurements/analysis.
RP performed HT-SELEX on ZFC4 constructs and carried out localisation studies of SALL4 mutants in cells.
KC analysed HT-SELEX data.
UJ prepared mutants and carried out binding assays.
BA-H carried out transfections and immunofluorescence.
VK developed purification protocols for different SALL4 constructs and found initial crystals for SALL4-DNA complexes.
MP developed initial purification protocols for SALL4/DNA complexes.
AGC solved, built and refined the structure.
AGC, RP and AB conceived the project and co-wrote the manuscript with input from all authors.

Competing Interests statement

The authors declare no competing interests.

Figure Legends

Figure 1 SALL4 ZFC4 in complex with DNA. (a) Domain overview of human SALL4 protein. Missense variants from gnomAD population data (blue) are placed relative to the sequence. Above are pathogenic missense mutations (red) and variants of unknown significance (orange) relevant to this work. The NuRD binding region (aa1-12) and Q-rich sequences (aa207-242) are followed by seven zinc finger domains arranged in three clusters (ZFC1, aa382-432; ZFC2, aa566-648; and ZFC4, aa870-920). The variant depletion value for SALL4 (Vp) is given, along with domain-level variant depletion values (VdVp ratios) under the domain labels. A construct of ZFC4 containing zinc fingers 6 (green) and 7 (blue) was used for structural analysis. Figure was generated using Plot Protein. (b) The palindromic DNA sequence co-crystallised with SALL4. Strand A and B are coloured yellow and white respectively. (c) An overview of the mouse SALL4 ZFC4 DNA complex, showing a single protein chain interacting with the DNA. Colour schemes match those of schematics in (a) and (b). (d) An all-atom model of SALL4 ZFC4 complex with DNA fitted into a SAXS envelope. The complex is rotated 90° around the vertical axis relative to (c).

Figure 2 SALL4 ZFC4 binds DNA with polar contacts. (a) Alignment of Znf6 and Znf7 showing standard position annotation for zinc finger helix residues along with secondary structure elements (arrows are beta strands, cylinder is an alpha helix) below. (b) Schematic overview of direct interactions between SALL4 ZFC4 and AT-rich DNA. (c-e) Zoomed in views of side chain interactions with AT-rich DNA.

Figure 3 ZFC4 patient missense mutations are likely to interfere with DNA binding. (a) Model of point mutations H898R and R900W (mouse numbering, pink carbon atoms) superposed on wild type structure to show alterations in proteins structure. H898R would disrupt zinc ion binding. R900W likely disrupts the interface between Znf6 and Znf7. (b) Zoomed-in view showing the position of the G921D mutation in the major groove.

Figure 4 ZFC4 patient missense mutations reduce SALL4 binding to DNA and alter localisation in nuclei. (a) Diagram showing SALL4 ZFC4 wild type and mutant constructs used in this study. (b) EMSA showing binding interactions of SALL4 ZFC4 wild type and mutant proteins with an AT-rich DNA motif. Protein concentrations used in titration points are shown below the gel. (b) Immunofluorescence of SALL4 wild type and mutant proteins in 3T3 cells transfected with expression constructs, with DAPI staining for comparison. Scale bars are 5µm.

Figure 5 ZFC4 mutations do not alter sequence preference. (a) Diagram summarising the HT-SELEX procedure to determine ZFC4 binding specificity. (b) Relative enrichment of 6-mer DNA motifs categorized by total number of A/Ts at cycle 1, 3 and 6 of HT-SELEX with SALL4 ZFC4 wild-type (blue), R900W (purple), G921D (green) and negative control (grey) samples. Error bars indicate the variability (SD) in three independent replicate experiments. (c) Venn diagram showing the overlap of the top 100 enriched 6-mer DNA motifs at cycle 6 of HT-SELEX with SALL4 ZFC4 wild-type and mutant proteins.

Supplementary files:

1. Supplementary figures and tables:
    a. Figures S1-S5
    b. Table S1-S3
2. Supplementary Data File 1 – SELEX dataset

References

Al-Baradie, R., K. Yamada, C. St Hilaire, W. M. Chan, C. Andrews, N. McIntosh, M. Nakano, E. J. Martonyi, W. R. Raymond, S. Okumura, M. M. Okihiro and E. C. Engle (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**(5): 1195-1199.

Borozdin, W., D. Boehm, M. Leipoldt, C. Wilhelm, W. Reardon, J. Clayton-Smith, K. Becker, H. Muhlendyck, R. Winter, O. Giray, F. Silan and J. Kohlhase (2004). "SALL4 deletions are a common cause of Okihiro and acro-renal-ocular syndromes and confirm haploinsufficiency as the pathogenic mechanism." J Med Genet **41**(9): e113.

Borozdin, W., M. J. Wright, R. C. Hennekam, M. C. Hannibal, Y. J. Crow, T. E. Neumann and J. Kohlhase (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**(8): e102.

Chambers, I., D. Colby, M. Robertson, J. Nichols, S. Lee, S. Tweedie and A. Smith (2003). "Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells." Cell **113**(5): 643-655.

Chen, V. B., W. B. Arendall, 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson (2010). "MolProbity: all-atom structure validation for macromolecular crystallography." Acta Crystallogr D Biol Crystallogr **66**(Pt 1): 12-21.

Deak, G. and A. G. Cook (2022). "Missense Variants Reveal Functional Insights Into the Human ARID Family of Gene Regulators." J Mol Biol **434**(9): 167529.

Diehl, A., W. Mu, D. Batista and M. Gunay-Aygun (2015). "An atypical 0.73 MB microduplication of 22q11.21 and a novel SALL4 missense mutation associated with thumb agenesis and radioulnar synostosis." Am J Med Genet A **167**(7): 1644-1649.

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

Elrod-Erickson, M., M. A. Rould, L. Nekludova and C. O. Pabo (1996). "Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions." Structure **4**(10): 1171-1180.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2126-2132.

Emsley, P., B. Lohkamp, W. G. Scott and K. Cowtan (2010). "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 486-501.

Franke, D. and D. I. Svergun (2009). "DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering." Journal of Applied Crystallography **42**(2): 342-346.

Garton, M., H. S. Najafabadi, F. W. Schmitges, E. Radovani, T. R. Hughes and P. M. Kim (2015). "A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity." Nucleic Acids Res **43**(19): 9147-9157.

Hooper, M., K. Hardy, A. Handyside, S. Hunter and M. Monk (1987). "HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells." Nature **326**(6110): 292-295.

Jolma, A., T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpaa, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen and J. Taipale (2010). "Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities." Genome Res **20**(6): 861-873.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alfoldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, C. Genome Aggregation Database, B. M. Neale, M. J. Daly and D. G. MacArthur (2020). "The mutational constraint spectrum quantified from variation in 141,456 humans." Nature **581**(7809): 434-443.

Kohlhase, J., D. Chitayat, D. Kotzot, S. Ceylaner, U. G. Froster, S. Fuchs, T. Montgomery and B. Rosler (2005). "SALL4 mutations in Okihiro syndrome (Duane-radial ray syndrome), acro-renal-ocular syndrome, and related disorders." Hum Mutat **26**(3): 176-183.

Kohlhase, J., M. Heinrich, L. Schubert, M. Liebers, A. Kispert, F. Laccone, P. Turnpenny, R. M. Winter and W. Reardon (2002). "Okihiro syndrome is caused by SALL4 mutations." Hum Mol Genet **11**(23): 2979-2987.

Kohlhase, J., L. Schubert, M. Liebers, A. Rauch, K. Becker, S. N. Mohammed, R. Newbury-Ecob and W. Reardon (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**(7): 473-478.

Konarev, P. V., V. V. Volkov, A. V. Sokolova, M. H. J. Koch and D. I. Svergun (2003). "PRIMUS: a Windows PC-based system for small-angle scattering data analysis." Journal of Applied Crystallography **36**(5): 1277-1282.

Kong, N. R., M. A. Bassal, H. K. Tan, J. V. Kurland, K. J. Yong, J. J. Young, Y. Yang, F. Li, J. D. Lee, Y. Liu, C. S. Wu, A. Stein, H. R. Luo, L. E. Silberstein, M. L. Bulyk, D. G. Tenen and L. Chai (2021). "Zinc Finger Protein SALL4 Functions through an AT-Rich Motif to Regulate Gene Expression." Cell Rep **34**(1): 108574.

Koshiba-Takeuchi, K., J. K. Takeuchi, E. P. Arruda, I. S. Kathiriya, R. Mo, C. C. Hui, D. Srivastava and B. G. Bruneau (2006). "Cooperative and antagonistic interactions between Sall4 and Tbx5 pattern the mouse limb and heart." Nat Genet **38**(2): 175-183.

Landrum, M. J., J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman and D. R. Maglott (2018). "ClinVar: improving access to variant interpretations and supporting evidence." Nucleic Acids Res **46**(D1): D1062-D1067.

Lauberth, S. M. and M. Rauchman (2006). "A conserved 12-amino acid motif in Sall1 recruits the nucleosome remodeling and deacetylase corepressor complex." J Biol Chem **281**(33): 23922-23931.

Manalastas-Cantos, K., P. V. Konarev, N. R. Hajizadeh, A. G. Kikhney, M. V. Petoukhov, D. S. Molodenskiy, A. Panjkovich, H. D. T. Mertens, A. Gruzinov, C. Borges, C. M. Jeffries, D. I. Svergun and D. Franke (2021). "ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis." J Appl Crystallogr **54**(Pt 1): 343-355.

Matyskiela, M. E., T. Clayton, X. Zheng, C. Mayne, E. Tran, A. Carpenter, B. Pagarigan, J. McDonald, M. Rolfe, L. G. Hamann, G. Lu and P. P. Chamberlain (2020). "Crystal structure of the SALL4-pomalidomide-cereblon-DDB1 complex." Nat Struct Mol Biol **27**(4): 319-322.

Matyskiela, M. E., S. Couto, X. Zheng, G. Lu, J. Hui, K. Stamp, C. Drew, Y. Ren, M. Wang, A. Carpenter, C. W. Lee, T. Clayton, W. Fang, C. C. Lu, M. Riley, P. Abdubek, K. Blease, J. Hartke, G. Kumar, R. Vessey, M. Rolfe, L. G. Hamann and P. P. Chamberlain (2018). "SALL4 mediates teratogenicity as a thalidomide-dependent cereblon substrate." Nat Chem Biol **14**(10): 981-987.

McCoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read (2007). "Phaser crystallographic software." J Appl Crystallogr **40**(Pt 4): 658-674.

Miertus, J., W. Borozdin, V. Frecer, G. Tonini, S. Bertok, A. Amoroso, S. Miertus and J. Kohlhase (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**(1-2): 154-161.

Miller, A., M. Ralser, S. L. Kloet, R. Loos, R. Nishinakamura, P. Bertone, M. Vermeulen and B. Hendrich (2016). "Sall4 controls differentiation of pluripotent cells independently of the Nucleosome Remodelling and Deacetylation (NuRD) complex." Development **143**(17): 3074-3084.

Panjkovich, A. and D. I. Svergun (2018). "CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data." Bioinformatics **34**(11): 1944-1946.

Pantier, R., K. Chhatbar, T. Quante, K. Skourti-Stathaki, J. Cholewa-Waclaw, G. Alston, B. Alexander-Howden, H. Y. Lee, A. G. Cook, C. G. Spruijt, M. Vermeulen, J. Selfridge and A. Bird (2021). "SALL4 controls cell fate in response to DNA base composition." Mol Cell **81**(4): 845-858 e848.

Patel, A., X. Zhang, R. M. Blumenthal and X. Cheng (2017). "Structural basis of human PR/SET domain 9 (PRDM9) allele C-specific recognition of its cognate DNA sequence." J Biol Chem **292**(39): 15994-16002.

Pettersen, E. F., T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris and T. E. Ferrin (2021). "UCSF ChimeraX: Structure visualization for researchers, educators, and developers." Protein Sci **30**(1): 70-82.

Rao, S., S. Zhen, S. Roumiantsev, L. T. McDonald, G. C. Yuan and S. H. Orkin (2010). "Differential roles of Sall4 isoforms in embryonic stem cell pluripotency." Mol Cell Biol **30**(22): 5364-5380.

Sakaki-Yumoto, M., C. Kobayashi, A. Sato, S. Fujimura, Y. Matsumoto, M. Takasato, T. Kodama, H. Aburatani, M. Asashima, N. Yoshida and R. Nishinakamura (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**(15): 3005-3013.

Schrodinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.

Stein, N. (2008). "CHAINSAW: a program for mutating pdb files used as templates in molecular replacement." Journal of Applied Crystallography **41**(3): 641-643.

Svergun, D., C. Barberato and M. H. J. Koch (1995). "CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates." Journal of Applied Crystallography **28**(6): 768-773.

Sweetman, D. and A. Munsterberg (2006). "The vertebrate spalt genes in development and disease." Dev Biol **293**(2): 285-293.

Sweetman, D., T. Smith, E. R. Farrell, A. Chantry and A. Munsterberg (2003). "The conserved glutamine-rich region of chick csal1 and csal3 mediates protein interactions with other spalt family members. Implications for Townes-Brocks syndrome." J Biol Chem **278**(8): 6560-6566.

Tickle, I. J., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., and C. Vonrhein, Bricogne, G. (2018). STARANISO. Cambridge, United Kingdom, Global Phasing Ltd.

Turner, T. (2013). "Plot protein: visualization of mutations." J Clin Bioinforma **3**(1): 14.

Vonrhein, C., C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack and G. Bricogne (2011). "Data processing and analysis with the autoPROC toolbox." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 293-302.

Wolfe, S. A., L. Nekludova and C. O. Pabo (2000). "DNA recognition by Cys2His2 zinc finger proteins." Annu Rev Biophys Biomol Struct **29**: 183-212.

Yamashita, K., A. Sato, M. Asashima, P. C. Wang and R. Nishinakamura (2007). "Mouse homolog of SALL1, a causative gene for Townes-Brocks syndrome, binds to A/T-rich sequences in pericentric heterochromatin via its C-terminal zinc finger domains." Genes Cells **12**(2): 171-182.

Figure 1

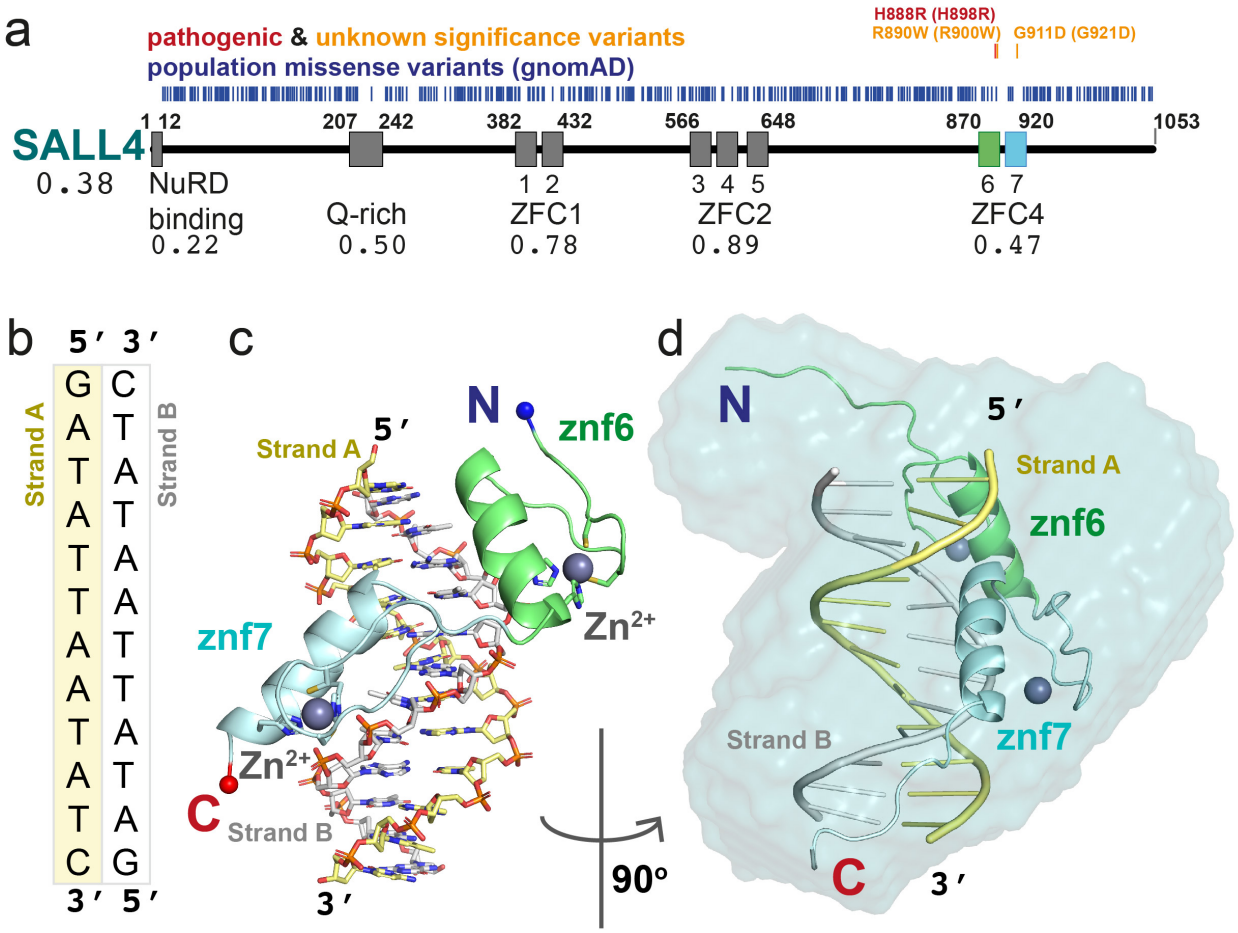

Figure 2

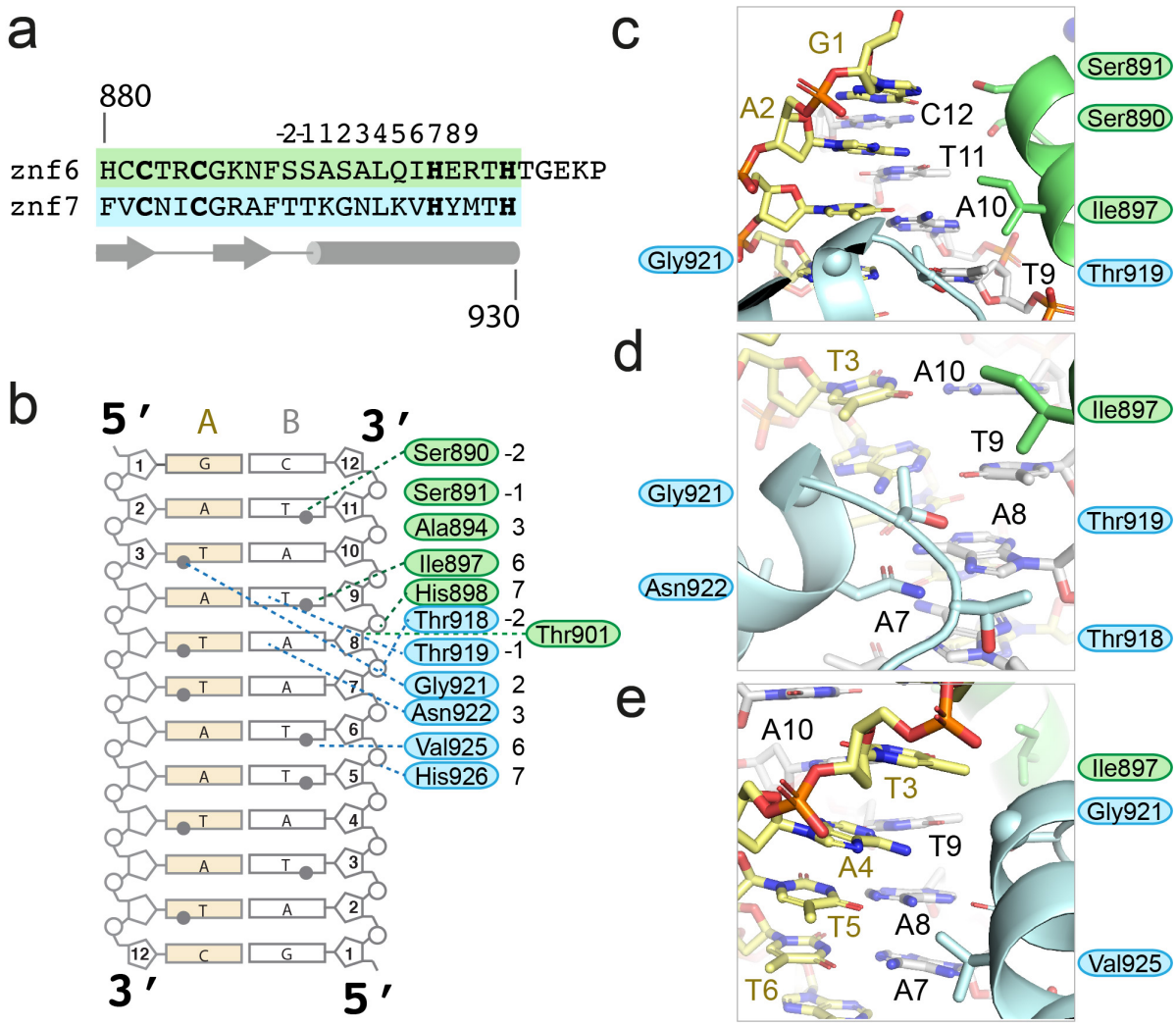

Figure 3

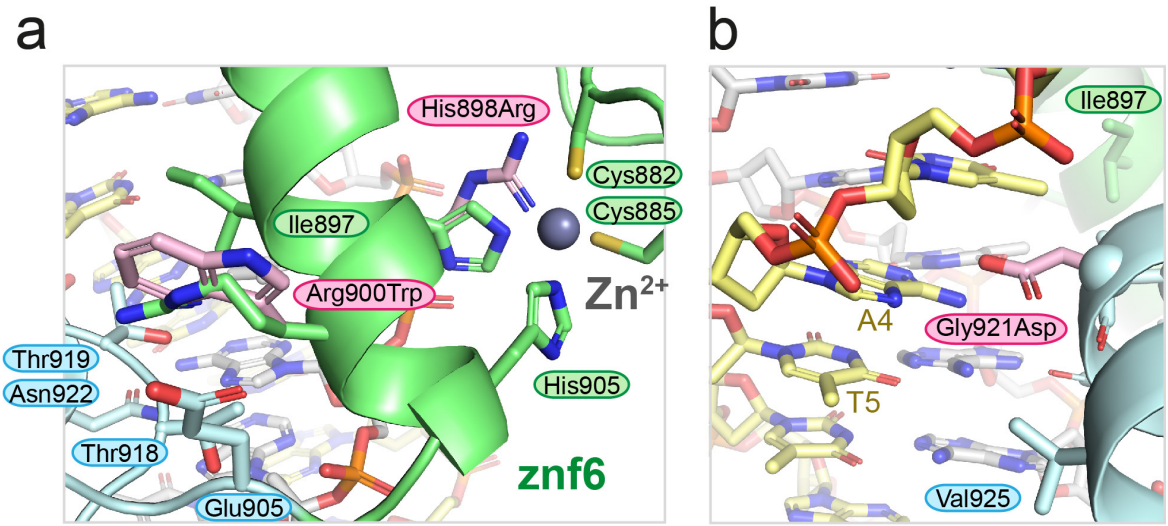

Figure 4

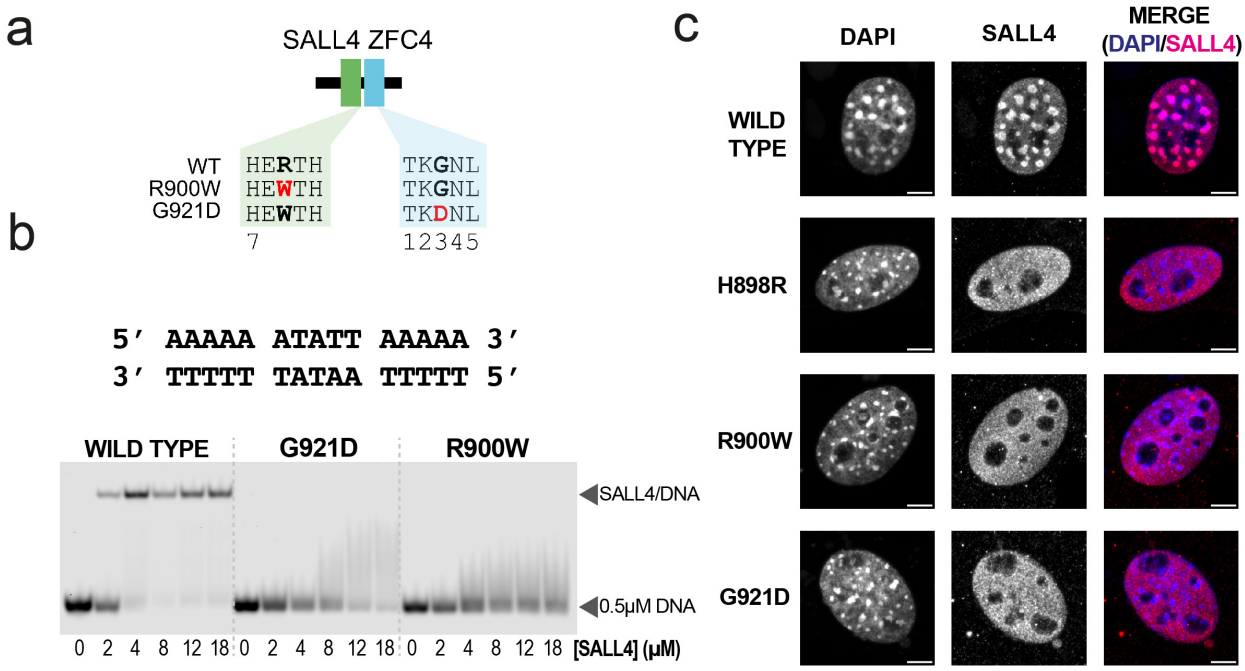

Figure 5