Molecular basis for DNA recognition by the maternal pioneer transcription factor FoxH1

Radoslaw Pluta
Institute for Research in Biomedicine https://orcid.org/0000-0002-1676-860X
Eric Aragon
Institute for Research in Biomedicine (IRB Barcelona)
Nicholas Prescott
Memorial Sloan Kettering Cancer Center, New York https://orcid.org/0000-0002-0635-8906
Blazej Baginski
Institute for Research in Biomedicine (IRB-Barcelona), The Barcelona Institute of Science and Technology https://orcid.org/0000-0002-8246-5041
Lidia Ruiz
Instituto Investigacion Biomedica
Julia Flood
Memorial Sloan Kettering Cancer Center, New York
Pau Martin-Malpartida
Institute for Research in Biomedicine (IRB-Barcelona), The Barcelona Institute of Science and Technology https://orcid.org/0000-0001-5867-5535
Joan Massague
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-9324-8408
Yael David
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0003-1696-0025
María Macías (maria.macias@irbbarcelona.org)
Institute for Research in Biomedicine (IRB-Barcelona), The Barcelona Institute of Science and Technology / ICREA https://orcid.org/0000-0002-6915-963X

Article

Keywords: FoxH1 structure, Fast-1, Pioneer factor, FOX proteins, Forkhead domain, forkhead target, embryo development

Posted Date: April 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1522438/v1

Version of Record: A version of this preprint was published at Nature Communications on November 26th, 2022. See the published version at https://doi.org/10.1038/s41467-022-34925-y.

Abstract

Nucleosomes are barriers for the binding of most transcription factors, but pioneer factors (PFs) do bind and facilitate subsequent interactions of other proteins during transcription activation. FoxH1 is a maternal PF essential during embryonic development that interacts with specific GK-forkhead targets. How FoxH1 binds DNA targets has remained elusive for decades until now. We have determined high-resolution structures of human, frog and fish proteins bound to four DNAs. We found that the FoxH1 DNA-binding domain is almost twice the size of other FOX proteins, allowing for a highly specific binding to both minor and major grooves. Consistent with its PF activity, we also quantified that the affinity for DNA is even higher for native mononucleosomes than for linear DNA. Our structures illustrate how binding to distinct GK sites allows FoxH1 to avoid cross-regulation by other FOX proteins that also operate during the maternal-zygotic transition and select canonical TT-forkhead sites.

Introduction

The transcription factor FoxH1 (also known as Fast1) was initially identified as a SMAD protein cofactor in TGFβ signaling [1] by Whitman and coworkers in 1996 [2]. FoxH1 recognizes the activin-responsive element (ARE), which confers activin regulation of Mix.2 transcription *in vivo*, a region that does not contain the consensus DNA binding motif characteristic of other winged helix factors belonging to the FOX family. In fact, FoxH1 selects *cis* regulatory sites containing GG and GT motifs *in vivo*, as in *Gsc*, *Eomes*, *Nodal*, *Mixl* and *FoxA2*.

In addition to these roles as a SMAD partner [3], FoxH1 was later identified as the earliest maternal pioneer factor (PF) that participates in the coordination of zygotic genome activation (ZGA) [4-6]. Embryonic development begins from a single-cell zygote, which undergoes rapid cell division to form a spherical blastula. These cells reorganize into three germ layers during gastrulation, giving rise to different parts of the developing organism [7]. Embryo formation requires cells to coordinate their location and fate during all stages of development. Positional information is conveyed by gradients of secreted signaling molecules that regulate key transcription factors, which play key roles in the formation of body axes and in tissue development, by controlling cell proliferation and migration [8]. In animals, embryo development is initially directed by maternal mRNAs and proteins. During the maternal-zygotic transition (MZT), the transcription factor (TF) FoxH1 functions at the top of the regulatory hierarchy by priming enhancers for the activation of developmental genes [9-13]. As a maternal PF, FoxH1 marks super-enhancers throughout the genome at the blastula stage, characterized by relatively nucleosome-dense chromatin [4-6,9,14,15]. Basal binding of FoxH1 primes mesendoderm differentiation promoters to activate transcription [9]. At the start of gastrulation, the primitive streak and other surrounding structures produce concentration gradients of signaling molecules such as Wnt proteins, as well as BMP, Activin and Nodal, which are a subset of the transforming growth factor beta (TGFβ) family of cytokines [16,17]. FoxH1 is key to transducing Activin/Nodal signals and directing maternal regulation of Wnt/β-catenin target genes for differentiation [9,18,19]. The essential roles played by FoxH1 are underlined by the identification of several

embryo mutants that lack anteroposterior axis specification, primitive streak patterning, and heart development [20-22].

FoxH1 belongs to one of the largest TF families found in eukaryotes: the Forkhead-box or FOX proteins [23,24]. These proteins share a DNA-binding domain known as the Forkhead (FH) domain (InterPro entry IPR001766). The FH fold comprises a three-helix bundle, a double (or triple) β-sheet, and two variable regions known as Wing1 and Wing2. In general, FOX proteins interact with the canonical forkhead motif (FKH) TRTTTRY (R=A/G, Y=C/T). Intriguingly, FoxH1 has a preference for TGT**GG**ATT and TGT**GT**ATT sequences among its consensus motif TGTKKATT (K=G/T). In fact, of all the FH factors examined, only FoxH1 selects GG-containing motifs [24,25]. The GG sites are present in the TGFβ/activin response element (TARE motif) identified in the *goosecoid* (*Gsc*) promoter, and in the ARE motif that confers activin regulation of Mix.2 through a defined mechanism [2,5,9,19,26-29].

Although many structures of FOX proteins bound to TT sites have been reported in the literature, there is a dearth of structural information regarding FoxH1 recognition of the specific GK sites. Understanding how this specific DNA recognition is achieved is essential to describe its roles in embryo development and in tumors associated to elevated FoxH1 expression in adults [30-32]. To fill this knowledge gap, we determined the crystal structures of the FoxH1 FH domain bound to three TGTKKATT consensus DNA motifs based on the native *Gsc* promoter sequence, as well as a complex bound to the canonical FKH DNA. To correlate small differences in protein sequence with DNA binding, we selected FoxH1 FH domains from three vertebrates (*Homo sapiens*, *Xenopus laevis,* and *Brachydanio rerio*) for structural studies. These complexes reveal the presence of unique flanking regions surrounding the FH domain core—conserved in vertebrate FoxH1 proteins—to ensure stability and high-affinity protein-DNA binding. Furthermore, these structures illustrate the exceptional ability of FoxH1 with respect to other FOX proteins to bind to GG and GT targets has even a slight preference for binding to mononucleosomes containing the canonical FoxH1 site over linear DNA, consistent with the role of FoxH1 as a PF. Although FoxH1 can also interact with canonical FKH DNA, binding to GK target sites appears to be an essential property of this maternal PF to avoid cross-regulation by other FOX proteins that also operate during embryonic development and select TT sites.

Results

The FoxH1 FH domain binds DNA with high affinity

FoxH1 proteins contain a FH DNA-binding domain and a C-terminal region including the Engrailed Homology 1 (EH1) motif and the SMAD-interacting domain (SID) involved in protein recognition (Figure 1a, Supplementary Figure 1a) [33-35].

To study the DNA-binding capacity of the FoxH1 FH domain, we selected three DNA variants of the TGTKKATT consensus (Figure 1b, Supplementary Figure 1b) based on the native *Gsc* promoter sequence and the FoxH1 sequences from three model organisms: *Homo sapiens*, *Xenopus laevis,* and *Brachydanio*

rerio (or *Danio*). Based on the detection of helical propensities outside the canonical boundaries [36] and to optimize the FH fold, we produced several constructs with and without residues surrounding the core region (Figure 1c).

Proteins expressing only the canonical FH domain were prone to degradation, whereas the most stable constructs were those that included extended boundaries at both sides of the FH domain. The deletion of the N-terminal extension produced insoluble proteins whereas the removal of the additional C-terminal region decreased affinity for DNA (Figure 1d). Using thermal shift assays, we explored the stability of the soluble constructs and the changes upon DNA binding. For the extended FoxH1 constructs, a clear increase in the stability of the protein (between 14 ˚C and 20 ˚C) was detected in the presence of DNAs containing the TGTKKATT motifs (GG, GT, and TT, Figure 1d , Supplementary Figure 1c and Supplementary Table 1). This increase was larger for the GG and GT than for the TT motifs, including the canonical FKH binding motif (TGTTTAC) of FoxA2, a well described FH family member which we use here for comparison with FoxH1.

Surprisingly, FoxA2 selected only for TT sites, with small changes of its melting temperature—or even a decrease—in the presence of GT and GG sites, respectively (Figure 1d). These observations were corroborated by native gel electrophoresis DNA binding assays, which revealed protein-DNA interactions in the nanomolar range for the extended FoxH1 constructs and a ~100-fold reduction in this affinity when the C-terminal FoxH1 extension was absent (Figure 1e, Supplementary Figure 1d).

Overall, these results establish the ability of FoxH1 to bind with high affinity to TGTKKATT motifs with GG, GT, and TT sites, whereas FoxA2 binds selectively to the motif with TT sites. Our findings also show that the FoxH1 FH domain is larger than other FH domains, with the additional areas being highly conserved in vertebrate sequences (Figure 1e).

General features of Gsc GG-recognition by FoxH1

To decipher how the N-terminal-charged region and the hydrophobic C-terminus contribute to protein stability and DNA binding, we started by determining the crystal structures of the three extended FH domains (constructs of ~140 residues instead of the canonical ~80 residue domain, Figure 1c) bound to the 16 bp GG sequence belonging to the *Gsc* promoter. We also determined the complexes with the GT and TT variants of the TGTGGATT sequence and with the canonical FKH site for comparison with other FOX structures.

The three GG complexes were refined at 1.5 Å (*Homo*), 0.98 Å (*Brachydanio*), and 2.8 Å (*Xenopus*) resolution. When superimposed, the root-mean-square deviations (RMSD) ranged from 0.47 to 0.65 Å (Figure 2a,b and Supplementary Figure 2a). The minor differences reflect the sequence variations between the three species, these variations being located mainly at Wing2, which is slightly longer in mammals than in other vertebrates (Figure 1f). Although the human sequence includes the EH1 motif, known to fold as a short helix [35], this part is not defined in the complex, thereby confirming that this region does not belong to the compact FH domain. In all complexes, the crystallographic asymmetric unit

contains one copy of a protein-DNA complex. A summary of the data collection, refinement statistics and PDB entries is shown in Table 1.

The three FoxH1 complexes have a conserved canonical core, composed of a three-helical bundle (H1-H2-H3 helices) followed by a pair of antiparallel β strands (S1, S2). In contrast to other FOX proteins, FoxH1 core is flanked by three well-ordered regions: the N-terminal loop1 (residues 22-30, human sequence), the Wing1 and an unusually long Wing2 (more than 50 residues) (Figure 2a,b and Supplementary Figure 2a). The overall fold is further stabilized by a K cation, which tethers the C-terminus of the helix H3 to the beginning of strand S1 (Figure 2c). The cation and its coordination properties have been verified at the *CMM server* (https://cmm.minorlab.org/) [37].

In all FOX proteins (including FoxH1), helix H3 is docked into the major groove to establish direct hydrogen bonds (dHBs) using three conserved residues: Asn (position i), Arg (i+3) and His (i+4). Remarkably, the specific contacts observed in FoxH1 differ from those reported for other members of the FOX family. First, FoxH1 proteins have a conserved Asp residue at position i (Asp79, human sequence) instead of the Asn present in almost all FOX proteins (Supplementary Figure 1a). Asp79 and His83 bind to the GG pair (Gua8-Gua9) through direct HBs (dHBs) with Cyt9' and with Gua8 and Cyt8' respectively (Figure 2d). Second, in FoxH1, the conserved Arg82 participates in dHB with Gua6 and Thy7 and, in the high-resolution structures, Arg82 also interacts with Ade10' through a water-mediated HB (wHB) (Figure 2d and Supplementary Figure 2b).

A summary of the protein-DNA interactions for the human sequence and the GG site is given in Figure 2e. These contacts comprise HB interactions with three riboses, nine phosphates, and ten nucleobases, including seven bases of the FoxH1-specific DNA motif (bases 6, 7, 8, and 12 and 6', 8' and 9' of the complementary strand), and cover 15 nucleotides. Most of these contacts are also observed in the GT and TT structures described below.

Binding to the GT and TT variants. Key roles of the Asp and Arg residues in DNA recognition

We also investigated how the major groove interactions differ in complexes with GT and TT sites. Given the overall similarity of the human, frog and zebrafish FH domains and that the latter yielded the best diffracting crystals, we used this construct for the studies with the GT and TT variants. These complexes were refined at 1.2 and 2.2 Å resolution respectively, and they are highly similar to the GG structure and between them (RMSD between the structures: GT-TT: 0.59 Å, GG-GT: 0.23 Å, GG-TT: 0.74 Å) (Figure 3a).

Focusing on the specific recognition of the different bases in the major groove, the Asp residue, which under physiological pH occurs as the negatively charged aspartate form, is able to establish direct HBs with either Cyt9' in the GG and GT sites or with Ade9' in the TT site, respectively. Also, if Cyt9' gets protonated, the aspartate and the cytosine can establish a salt bridge interaction, which is much stronger than a regular HB. The His residue also binds to the three KK base pairs accommodating the orientation of its side chain. In both GT and TT complexes, the His directly contacts Thy7', Thy9 and Ade8', and also Thy8 in the TT complex only (Figure 3b). In contrast, the Asn present in other FOX proteins has a polar

side chain and either selects the Adenine bases in complexes containing TT sites or does not participate in DNA binding, as in the structure of FoxC2 bound to a GT site (PDB 6akp, [38]).

The contacts between the Arg and DNA are identical in all FoxH1 complexes with this residue contacting the common region of the DNA motifs (Figure 3c,d). The Arg side chain participates in a HB with Gua6 through a direct bidentate bond with the O6 and N7 atoms, contacts that are normally absent in most FOX proteins characterized to date including FoxA2 complexes. To corroborate these observations, we also determined the X-ray crystal structure of an extended version of FoxA2 (Figure 1c) bound to TTACT and TT DNAs (refined at 2Å). These structures confirm a non-favorable orientation of the Arg side chain as in the previously described complex bound to the TTACT site, [39] (Figure 3d, Supplementary Figure 3a-c), and the interaction with Gua6 is water-mediated. In fact, in many FOX complexes, the Arg side chain is partially sequestered by an interaction with the hydroxyl group of a neighboring Tyr, preventing the formation of bidentate bonds with the Gua base [25,38-43] (Figure 3e, Supplementary Figure 3d). However, because FoxH1 has Leu/Ile residues at the Tyr position, the Arg-Tyr interactions cannot occur, facilitating the optimal orientation of the Arg side chain towards the DNA and the bidentate interaction.

Overall, these structures illustrate how two modifications in helix H3 (Asp vs. Asn) and helix H2 (Leu/Ile vs. Tyr) allow the FH domain of FoxH1 to recognize a specific subset of FHK motifs containing GG and GT bases without losing the ability to interact with TT sites.

Both FoxH1 FH wings contribute to DNA recognition

Wing2 folds as a series of short α-helical turns and participates in a network of interactions with the domain core and with the N-terminal loop, including a salt bridge between Asp32 and Arg130 (Figure 2a,b and Figure 4a). The 5-amino acid insertion observed in mammals is located between the H6 and H7 helices and, due to its flexibility, it cannot be fully traced in the human sequence. The Wing2 region also interacts with DNA, directly with the backbone (Gln123 to Ade6', Asn124 to Ade16, and Arg126 to Gua15) and indirectly by guiding the N-terminal extension towards minor groove 1 (Figure 4 a,b and Supplementary Figure 4a,b). This network of interactions is conserved in the three species and explains our observation that constructs lacking the C-terminal extension bind DNA with 200 times lower affinity than the extended domain (Figure 1e).

The remaining interactions with DNA arise from Wing1, which specifically contacts minor groove 2. The main contacts from Wing1 to the minor groove are not affected by the substitutions of GG by GT or TT bps (Supplementary Figure 3a,b). These contacts involve a pair of specific HBs from Lys104 to Thy4 and Thy14', as well as with the ribose of Cyt15' (Figure 4c and Supplementary Figure 4b,c) in the GG complexes, and from Lys 168 (equivalent to 104) and Lys160 or Lys164 to Cyt15' in the GT and TT complexes. These direct and specific interactions with DNA from both Wings and the N-terminal loop are rare in other FOX proteins and, when observed, the interactions usually occur with the backbone DNA in a non-specific manner.

The N-terminal KYR motif binds to the minor groove

Many FOX proteins contain positively charged residues surrounding the FH domain. These residues usually act as nuclear localization signals or promote non-specific interactions with the backbone DNA, thereby enhancing overall protein-DNA affinity. In the case of FoxH1, we observed that the residues surrounding the core domain are essential to obtain stable and functional proteins. We tested several constructs, varying the cloning starting site, and found that constructs lacking the first 27 residues were prone to degradation during purification. Once we had solved the structures, we found the explanation for these observations. In the three GG, GT and TT complexes, and in the three-species investigated, the N-terminal loop, and in particular the KYR motif, is exceptionally well-ordered. This order is the result of the multiple interactions that the loop establishes with the minor groove and with Wing2. For instance, in the human GG complex, several base-specific HBs from residues 28 and 30 (Tyr28 to Ade6', and Arg30 to Thy12 and Gua13) are observed. These interactions are further supported by one wHB (Arg30 to Cyt14) and by a network of van der Waals contacts between residues in the loop and the DNA bases (Tyr28 to Thy12, Ade6', and Thy7'; Arg30 to Gua13 and Ade5'). We also detected HBs with the DNA backbone involving residues Lys26, His31, Lys33, and Tyr38, the first residue of helix H1. These interactions are highly conserved, as depicted in the snapshot showing the high-resolution 2Fo-Fc electron density map of the zebrafish-GT complex contoured at 1.0 sigma (Figure 4d). In contrast, the FoxA2 structure confirmed that its folded region involves only the canonical part of the FH domain (Figure 3d), even in the presence of extended boundaries as those present in the FoxH1 domain. The conserved and specific patterns of FoxH1 interactions with the GG, GT and TT motifs are schematically summarized in Figure 4e.

DNA shape analysis. Hoogsteen pairings in the GG complex

To quantify the differences in DNA binding and to explore the impact of the protein-DNA interactions in the DNA topology, we compared the DNA shapes of the FoxH1 complexes to other FOX structures available in the PDB and to the pair of complexes we determined for FoxA2 using Curves+ [44] (Supplementary Figure 5a,b). We detected small differences at the major groove, whereas the minor grooves were found to be slightly wider in FoxH1 complexes than in other FOX counterparts, as measured. These differences at the minor groove are probably caused by the abundant number of protein-DNA interactions present in FoxH1, which are absent in other FOX proteins. To quantify these differences, we used FoxA2 for comparison and measured the number of HBs detected in both structures, as well as the DNA area covered by each protein in the different TT complexes (Supplementary Table 2). FoxH1 duplicates the number of protein-DNA HBs detected in FoxA2 and buries an area of 1553 $Å^2$ vs. 1012 $Å^2$ covered by FoxA2.

We also noticed that, in the *Brachydanio* GG structure and in the region proximal to Wing1 binding, two bps at one end of the DNA helix form Hoogsteen pairings (HGs), instead of the more common Watson-Crick-Franklin (WCF) base pairing that we observed in all other complexes (Figure 5a,b and Supplementary Figure 5c,d). HGs require flipping of the purine base by 180º with respect to the corresponding WCF bp and the protonation of the cytosine N3. Although HGs are still infrequent in structures, their proper identification and functional implications have gained attention in the last decade [45]. Some of these functional effects are related to DNA damage and the generation of point

mutations through a mechanism that starts by introducing a mismatch of an oxidized guanine (8-oxoG) with an adenine via HG bps, which results in a G:C to T:A transversion [46]. Since these HG pairings are at the edge of the DNA in this case, they might have been further stabilized by crystal packing.

FoxH1 binding to reconstituted mononucleosomes

The DNA-binding domain of FOX proteins has a structural similarity to the folded domain present in the linker histones H1 and H5 as they all share the same winged-helix domain fold [41,47,48]. This structural convergence, originally identified in FoxA3 with H1, is thought to allow Pioneer Factors to gain access to chromatin, thus facilitating the subsequent binding of other TFs and chromatin remodeling proteins [5,9,49]. In the case of PFs, the similarity might go further as histones H1 and H5 also have extended motifs surrounding the compact domain that contribute to non-specific and specific interactions with nucleosomes [50]. To study FoxH1 binding to nucleosomes, we first used the Widom 601 model sequence using a compact 147 bp nucleosome core particle (NCP) [51] and introducing the FoxH1 motif close to the nucleosome edge. Using electrophoretic mobility shift assays (EMSA) and the 147 bp nucleosome core particle, we observed specific interactions between the protein and the compact mononucleosome. As a control for sequence-agnostic binding, we used histone H1A and the 147 bp Widom 601 nucleosome (Figure 6a,b, Supplementary Table 2).

To clarify whether FoxH1 binds to compact nucleosomes, we next sought to characterize the interactions between the FoxH1 FH domain and the native *Gsc* nucleosome sequence containing the TGTGGATT motif. To this end, we first generated mononucleosomes containing a 167 bp fragment of the native *Gsc* sequence using the boundaries mapped in the Gene Expression Omnibus (entry: GSM2842982) (Figure 6c). To quantify the affinity of protein-DNA interactions, we measured nucleosomal or linear DNA of *Gsc* as binding substrates for the FoxH1 FH domain in biolayer interferometry (BLI) assays (Figure 6d). We also measured the interactions of FoxA2 with the same DNA molecules for comparison, since at the end of the Gsc NCP there is a TGTTAAC motif that almost fits the so called 'degenerate FoxA motif on nucleosomes' [52]. We observed that both FoxH1 and FoxA2 domains interact with higher affinity to NCP mononucleosomes than to linear DNA, although the affinity is always higher for FoxH1 than for FoxA2 (Figure 6e).

We also used this binding assay to quantify the effect of point mutations in the FoxH1 FH domain. The functional role of residues located at the KYR loop was postulated after genetic analyses in zebrafish embryos (*sur/schmalspur*) [53,54] and more recently has been highlighted by mutations detected in human tumors. (https://cancer.sanger.ac.uk) [55]. The mutation of the Arg and Lys in zebrafish (R94H and K97N mutants) induces severe developmental pathologies. To confirm the role of the KYR motif in DNA binding and its functional role, we generated point mutations at the KYR positions. As observed in the K_D measurements, all mutations show reduced DNA-binding capacity with respect to the WT construct using the GG site or native *Gsc*-NCP.

The results of mononucleosome binding corroborate the role of FoxH1 and FoxA2 proteins as PFs as they show a preference for binding to mononucleosomes containing either the canonical FoxH1 site or a degenerate FoxA motif over binding to linear DNA.

Motif analysis: FoxH1 vs. FoxA2

Dynamic binding of FoxH1 allows for temporal control of the recruitment of co-factors to *cis*-regulatory modules (CRMs) during mesendodermal programming. During MZT, FoxH1, as a PF, primes CRMs for recruitment of TFs such as SMAD2/3 and FoxA2 to promote different cell fate transitions [5,9]. Before and during early stages of MZT, the abundance of maternal FoxH1 is reflected in FoxH1 binding evenly within the ±10kb region from the genome-wide transcription start site. During MZT, the level of maternal FoxH1 decreases with FoxH1 remaining persistently bound close to TSS and overlapping with FoxA-bound regions [5]. The ChIP-seq data available for FoxH1 and FoxA2 in Xenopus indicate that at stage 10.5, FoxH1 occupies only 3% of genomic sites it occupies during stage 8, and the remaining sites are largely bound by FoxA2 [5]. These observations led the authors to propose a molecular mechanism to explain this process that involves either the sequential binding of each TF to different sites within these CRMs or through direct competition between FoxH1 and FoxA2 for binding to the same sequences.

We set out to analyze whether the different structural preferences we detected for FoxH1 and FoxA2 are identifiable in ChIP-seq peaks available for both proteins at different stages and whether these differences might help explain the mechanism of action of these two TFs. We found datasets at stages 8, 9 and 10.5 for FoxH1 and at 10.5 for FoxA2 and considered that FoxH1 protein decreases at stage 10.5, which is the stage that shows an increase in FoxA2 (chip-atlas and in the Gene Expression Omnibus GEO database, [5,56] (Methods section). When comparing ChIP-seq peaks from stages 8/9 and 10.5, we found that the FoxH1-specific GG and GT motifs are present only in regions where FoxH1 binds, and in the common regions of both proteins, FoxH1 uses TT motifs for binding to DNA (Supplementary Figure 6a,b). FoxH1 motifs bearing GK motifs are much more enriched at stage 10.5 than stage 8/9 (x7.31 enrichment vs. x4.42), while TT motifs do not show this difference. As expected, FoxA2 TT motif enrichment does not vary significantly in these datasets (Supplementary Figure 6b).

Together with the structural information we gathered for the DNA binding specificities of FoxA2 and FoxH1 FH domains, these bioinformatic results are compatible with a new hypothesis where FoxA2 substitutes for FoxH1 preferentially at TT sites, with FoxH1 remaining bound to GK sites even when the concentration of this protein is low. When both TFs interact with a given accessible CRM, because FoxA2 cannot interact with GK sites efficiently, they do so at those CRMs that contain either a FKH TT site or several adjacent TT and GK binding sites (Figure 6e).

Discussion

FoxH1 has a unique fold and DNA binding ability

Our structural studies reveal three essential features of the FoxH1 sequence that enable this winged helix factor to recognize DNA through a network of interactions with both major and minor grooves. First, FoxH1 has an extended fold that almost duplicates the FH domain length when compared to other FOX proteins. Second, it also has two specific amino acid modifications, Asp versus Asn and Leu/Ile versus Tyr, which are essential to interact with the major groove and with both GG and TT sites. Under physiological pH, the Asp residue appears as the negatively charged aspartate form and is able to establish direct HBs with either the cytosine or adenine bases, respectively, or even through a strong salt bridge interaction if the cytosine is positively charged [57]. In contrast, the Asn present in other FOX proteins has a polar side chain and selects only the adenine bases in complexes containing TT sites. Acidic residues, such as aspartate and glutamate, are not very common at protein-DNA interfaces of transcription factors, although their presence has also been reported in other DNA complexes, such as that of the telomere-interacting protein RAP1 and several restriction endonucleases and Zn fingers [58,59]. The infrequent occurrence of these two acidic amino acids is probably explained by electrostatic repulsions with the phosphate groups of the DNA when the carboxylic group does not participate in specific interactions with cytosine bases [58,60]. And third, the substitution of Tyr for Leu/Ile has an allosteric effect on DNA binding. The presence of Leu/Ile (instead of the aromatic Tyr) provides rotational freedom to the Arg, thereby facilitating the formation of bidentate HBs with Gua6, which increases protein-DNA affinity and specificity. As FoxH1 has the largest structure of a FH domain characterized to date, the specific interactions with DNA are extended from the initially proposed canonical motif to include two additional bps at the 5' site and one more at the 3' site, ATTGTGGATTG. At the 5’ site, these additional interactions occur through contacts from Lys104 (human sequence) located in Wing1 with Thy4 and the complementary base of Ade3 (Thy14’). The additional contacts at the 3’ site arise from interactions of the KYR loop and the large Wing2, which recognize Gua13.

Intrigued by the importance of the KYR motif for FoxH1-DNA contacts, we analyzed sequences of other FOX proteins and found that FoxA2/3 and FoxQ1 could have such N-terminal motifs. As an example, we investigated whether the N-terminal region of FoxA2 could also contribute to the specific minor groove recognition by determining the structures of an extended domain with similar boundaries as those of FoxH1. However, these complexes showed that the extended N- and C- termini are predominantly disordered and neither adopt additional tertiary structures nor engage in specific DNA binding, probably because FoxA2 does not have the unusually large Wing2 of FoxH1, which guides the KYR motif towards the DNA minor groove.

In accordance with the documented role of FoxH1 as a PF, we confirmed that the extended FH domain of FoxH1 binds with nanomolar affinity *in vitro* to native *Gsc* mononucleosomes with the specific FoxH1 site next to the entry-exit ends. Interestingly, FoxH1 displays a preference for nucleosomal DNA over linear DNA in binding assays (Figure 6d,e). Thus, the presence of the extra base pairs in the motif seems to enhance the overall affinity and shape complementarity between the protein and DNA, even when the DNA motif is wrapped around a histone octamer. These observations, both of *in vitro* binding phenomena and of *in vivo* motif positioning, align with similar data for other PFs, such as Oct4 and Sox2 [61,62].

The structural comparison of the DNA binding specificities displayed by FoxH1 and FoxA2 has shed light on the interaction between these factors when they bind to CRM during early development, an open question for the past several years [5]. Given the preference of FoxA2 for TT vs. GK sites, if both TFs can bind to common promoters at the same binding site then this feature will occur only in promoters containing TT sites. The presence of GK sites in many essential FoxH1 targets may have been optimized, together with these key differences in the FoxH1 structure, thus avoiding the mis-regulation of essential genes that contain GK sites by other FOX proteins optimized to select TT motifs.

In addition, the specific structural properties of FoxH1 provide important insights into the mechanism of DNA recognition by this PF. These results have the potential to facilitate the design of selective small-molecule compounds to expand our understanding of the roles of FoxH1 in development and in reprogramming. This knowledge might also assist the development of new therapeutic strategies to target increased expression of FoxH1 in acute myeloid leukemia [32] due to its distinct fold and DNA sequence preferences

Declarations

Acknowledgements

We thank the staff of the Mass Spectrometry Core Facility (Universitat de Barcelona) for support, the Protein Expression Core Facility (IRB Barcelona) for providing reagents and the Automated Crystallography Platform staff (IRB Barcelona-CSIC) and ALBA Synchrotron staff for support with the experiments performed at the BL13-XALOC beamline. We also thank J. Cordero for preliminary experiments. We also thank C.D. Lima (MSKCC) and the Tri-Institutional Therapeutics Discovery Institute for providing access to biolayer interferometry instrumentation.

Funding

R.P. and B.B are co-funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie COFUND actions of IRB Barcelona and the PROBIST and PREBIST Postdoc and Predoc Programmes, respectively (agreements IRBPostPro2.0_600404, PROBIST_754510, and PREBIST_754558). N.A.P has received support from the National Science Foundation Graduate Research Fellowship 2017239554, NIH grant T32 GM115327-Tan and NIH grant F99CA264420.

This work was supported by the Spanish MINECO program (BFU2014-53787-P and BFU2017-82675-P, M.J.M), and access to ALBA through the BAG proposals 2018092972, and 2020094472. The work was also financed through the Spanish Ministry of Science and Education and the National Investigation Agency MCIN/AEI/ 10.13039/501100011033 and the European Regional Development Fund (ERDF) (BFU2014-53787-P and BFU2017-82675-P, M.J.M), by AGAUR (SGR-50, M.J.M) and by NIH grants R35-CA252978 (J.M.) and P30-CA008748 (MSKCC). Y.D. is supported by the Josie Robertson Foundation, the Pershing Square Sohn Cancer Research Alliance, the NIH (CCSG core grant P30 CA008748, MSK SPORE P50 CA192937, and R35 GM138386), the Parker Institute for Cancer Immunotherapy (PICI), and the Anna

Fuller Trust. In addition, the David lab is supported by Mr. William H. Goodwin and Mrs. Alice Goodwin and the Commonwealth Foundation for Cancer Research and the Center for Experimental Therapeutics at MSKCC. We gratefully acknowledge institutional funding from the CERCA Programme of the Government of Catalonia, IRB Barcelona, the BBVA Foundation, and the Spanish Ministry of Science and Education through the Centres of Excellence Severo Ochoa Award. M.J.M is an ICREA Programme Investigator.

Author contributions

M.J.M and J.M. designed the project. R.P. collected X-ray data, determined the structures and analyzed them with P.M.M. and M.J.M. E.A. cloned, expressed and purified all proteins and characterized their folding properties in solution. R.P. and B.B. screened crystallization conditions. E.A., L.R. and B.B. optimized the *in vitro* reconstruction of Widom601 NCPs for EMSAs, N.P. and J.R.F. optimized the *in vitro* reconstruction of native *Gsc* and Widom601 NCPs and analyzed the quantitative binding assays with Y.D. L.R. performed the DSC assays. P.M.M. analyzed the ChIP-Seq and MNase data. All authors contributed ideas to the project. M.J.M. R.P. and P.M.M. wrote the manuscript with contributions from all other authors.

Competing interest statement

J.M. owns company stock of Scholar Rock. The remaining authors declare no competing interests.

Data availability

PDB accession codes are: hFoxH1-GG: 7YZB; bFoxH1-GG: 7YZ7; bFoxH1-GT: 7YZA; bFoxH1-TT: 7YZC; FoxA2-TTACT: 7YZE; FoxA2-TT: 7YZF; xFoxH1-GG: 7YZG. Source data are provided with this paper.

Materials And Methods

Protein expression and purification

Wild-type FoxH1 (*Brachydanio rerio:* Uniprot Q9I9E1, aa 86-210, *Homo sapiens*: Uniprot: P75593, aa 1-185 and *Xenopus laevis*: Uniprot P70056, aa 97-236) were cloned in pOPINS using a synthesized DNA template with optimized codons for bacterial expression (Thermo Fisher Scientific). Point mutations were produced by site-directed mutagenesis PCR reactions and confirmed by DNA sequencing (GATC Biotech). FoxA2 protein (*Homo sapiens*, Uniprot: Q9Y261, aa 153-273). All proteins were expressed fused to an N-terminal His-tag SUMO-tag followed by the Ulp1 peptidase cleavage site in *E. coli* B834(DE3) strain essentially as described [9,28,63]. Cells were grown at 37°C in Terrific Broth and induced with IPTG (0.5 mM) at an OD600 of 0.8. After overnight expression at 20°C, bacterial cultures were centrifuged and cells were lysed at 4°C (EmulsiFlex-C5, Avestin) in 50 mM Tris, 400 mM NaCl, 40 mM imidazole, 1 mM TCEP and Tween 20 0.2% V/V pH 8 at 25°C in the presence of lysozyme and DNase I. Supernatants containing the soluble proteins were diluted until a conductivity of 10 mS/cm was reached, then loaded into HiTrap™ SP HP 5 mL column and eluted by a NaCl gradient to remove non-specifically bound bacterial DNA, using a

NGC™ Quest 10 Plus Chromatography System (Bio-Rad). Fractions containing the protein of interest were pooled, dialyzed to reduce the NaCl concentration and cleaved with recombinant Ulp1 (SUMO protease) overnight at 4°C. Cleaved proteins were loaded into a HiTrap™ SP HP 5 mL column to separate the SUMO tag from the FoxH1/FoxA2 FH domains. Finally, all FH domains were purified by size exclusion chromatography on a HiLoad™ 16/600 Superdex™ 75 pg (GE Healthcare Life Science) in 50 mM Tris, 150 mM NaCl and 1 TCEP at pH 7.2 at 25°C (buffer A) and kept at -80°C.

Nucleosome preparation and binding assays

DNA Preparation. A 167 bp fragment from the native Gsc promoter, containing the FoxH1 binding site, was amplified by polymerase chain reaction. A 40X reaction was prepared using Phusion polymerase (Thermo Scientific) following the manufacturer's instructions, in the presence of 5'-biotinylated forward (BIO-5'-ACAAGGCCTGAAAAGAGATTGTGGATTGCGA-3') reverse primers (5'-TCGGGCGGAGGGAGTTGTTAACTGCGGCGGCAC-3'). Amplicon was purified using the QIAquick PCR Purification kit from Qiagen and eluted in water. Eluent was then pooled, lyophilized, and resuspended to a final concentration of approximately 1.5 mg/mL.

Human Core Histone Preparation. Each of the four canonical human core histones (H2A, H2B, H3.2, and H4) in a pET3 vector was independently transformed into *E. coli* BL21 (DE3) cells and grown at 37 ºC until reaching an OD600 of 0.6-0.8, at which time isopropyl-β-d-thiogalactopyranoside (IPTG) was added to a final concentration of 0.5 mM. Cultures were grown for another 3-4 h before harvesting by centrifugation at 5,000 x *g* for 10 min and freezing at -20ºC.

For all histones, the pellet from a 1 L expression culture was thawed and resuspended in a lysis buffer containing 1 X PBS and 1 mM DTT. Sample was lysed by rod sonication with a Branson digital sonifier (40% amplitude, 5 s on, 10 s off, 90 s on-time). Lysate was then clarified by centrifugation at 20,000 x *g* for 25 min. Supernatant was discarded, and the insoluble pellet was then resuspended in a buffer containing 6 M guanidine hydrochloride, 1 X PBS, and 1 mM DTT. Inclusion bodies were extracted from the resuspended pellet by rotating overnight at 4ºC. Samples were then clarified as before by centrifugation, and the supernatant was then passed through a 0.22 µm filter. Next, core histones were purified by RP-HPLC on a preparative scale C-18 column. HPLC buffer A was 0.1% (v/v) TFA in water, and HPLC buffer B was 90% (v/v) acetonitrile and 0.1% (v/v) TFA in water. Core histone purification used a gradient from 40% to 70% buffer B. Finally, purified histones were analyzed for mass and purity by LC/ESI-MS. Purified histones were aliquoted, lyophilized, and stored as powder at -70ºC.

Histone Octamer Preparation. Core histones were individually dissolved in an unfolding buffer (6 M guanidine HCl, 20 mM Tris pH 7.6, 10 mM NaCl, 1 mM EDTA, 1 mM DTT) and quantified by A280. Histones were combined at a 1:1:0.95:0.95 molar ratio of H2A:H2B:H3.2:H4. Pooled histones were diluted to a total concentration of approximately 1.0 mg/mL and dialyzed against refolding buffer (2 M NaCl, 10 mM Tris pH 7.6, 1 mM EDTA, 1 mM DTT) with three exchanges, each of which lasted for 6-12 h. The mixture was then recovered, concentrated to a volume of less than 1 mL on a 30 kDa centrifugal filtration

concentrator, and cleared by centrifugation for 5 min at 17,000 x *g* at 4ºC. Supernatant was then resolved using a Superdex200 Increase 10/300GL on an AKTA FPLC column. Octamer-containing fractions were pooled, concentrated, and diluted 50% with glycerol before long-term storage at -20ºC.

Nucleosome Assembly. Assembly reactions were performed at 2-5 µM (Gsc167 DNA) scale via serial salt dialysis. The histone octamer: DNA ratio used was optimized empirically to favor full conversion of free DNA into nucleosomes, and all steps were performed at 4ºC. DNA, octamers, and buffer were combined to a final volume of 20 µL in buffer composed of 2 M NaCl, 10 mM Tris pH 7.6, 1 mM EDTA, 1 mM DTT and placed into 7,000 MWCO Slide-A-Lyzer Mini dialysis buttons pre-moistened in 200 mL of initial buffer (1.4 M NaCl, 10 mM Tris pH 7.6, 0.1 mM EDTA, 1 mM DTT). After 1 h of dialysis in initial buffer, a peristaltic pump was used to add a total 350 mL of dilution buffer (10 mM NaCl, 10 mM Tris pH 7.6, 0.1 mM EDTA, 1 mM DTT) at a rate of 1 mL per min. Samples were moved to another 350 mL of dilution buffer 1-2 h after the peristaltic pump transfer was completed, and was allowed to dialyze in this new buffer for another 6-12 h. Samples were transferred to 300 mL of fresh dilution buffer and allowed to dialyze for a final 1-2 h before harvesting from the dialysis cassettes. After samples were removed from dialysis, they were then subject to centrifugation at 17,000 x *g* for 5 min to remove any precipitate that formed over the course of dialysis. Finally, the quality of nucleosome assembly was analyzed by native PAGE using 5% acrylamide, 0.5 X TBE gels. Nucleosomes of suitable quality were pooled and quantified by A260.

Biolayer Interferometry was used to characterize binding kinetics between FoxH1 FH domain and either free or nucleosomal Gsc167 DNA on an Octet Red96e system (Sartorius). All reagents (DNA, nucleosomes, FoxH1) were exchanged into the following assay buffer for BLI experiments: 20 mM Tris pH 7.6, 100 mM NaCl, 2 mM DTT, 0.02% (v/v) Tween-20, 0.01% (w/v) BSA. Gsc167 DNA and nucleosomes were both diluted to concentrations of approximately 5 ng/µL DNA. A two-fold dilution series of FoxH1 was prepared, starting from a concentration of 125 nM and going down to 3.9 nM. Prior to experiments, streptavidin biosensors were pre-moistened and blocked in the assay buffer for at least 20 min. Binding kinetics experiments were performed in standard kinetics mode at 25ºC with the plate being rotated at 800 rpm throughout the assay. Loading of nucleosome or DNA ligand onto the sensors was limited to a threshold of 0.25 nm, analyte association was measured for 180 s, and dissociation for 240 s. Data analysis was performed using the Octet Data Analysis software, and data were fit to a 1:1 binding model for estimation of kinetic parameters. Data reported are derived from global fitting of replicates at five different protein concentrations, referenced against sensors with no FoxH1 added, with error values calculated by the analysis software. Goodness of fit was analyzed by visually examining residual plots for the fitted curves, as well as the R^2 and X^2 values of each fit.

Electrophoretic mobility shift assay (EMSA)

Duplex Cy®5-DNA was annealed using complementary single-strand HPLC purified DNAs. DNAs were mixed at equimolar concentrations (3 mM) in 20 mM Tris pH 7.0 at 25°C and 10 mM NaCl, heated at 90°C

for 3 min and cooled down to room temperature for 2 h. Protein-DNA binding reactions were carried out for 30 min at 4°C in 10 µL of binding buffer (100 mM Tris, 10% glycerol). A fixed concentration of 5'-end Cy5-labeled (Biomers, Germany) duplex DNA (7.5 nM) was incubated with increasing amounts of the different FoxH1 FH domain constructs. Electrophoresis was performed in native 8% polyacrylamide gels (1.5 mm thick), prepared with 29:1 30% acrylamide solution (PanReac AppliChem). The gels were run for 30 min in TG buffer at 150 V at 4°C and exposed to a Typhoon imager (GE Healthcare).

Differential scanning fluorimetry (DSF) assay

DSF assays were performed using a LightCycler 480 real-time PCR device (Roche, Switzerland) as described previously. Protein (at a final concentration of 2.0 mg/ml) and SYPRO Orange (Sigma USA, at a final concentration of 5.0 mg/ml) were mixed and placed into the instrument at a heating rate of 1°C/min. The fluorescence intensity vs. temperature (melting curve) was measured, and a melting temperature (Tm) was calculated from the maximum value of the first derivative of the curve using the in-house programme HTSDSF explorer (https://github.com/maciaslab/htsdsf_explorer) [64].

Crystallization

The protein-DNA complexes were prepared by mixing protein with DNA at a 1:1.2 molar ratio. FoxH1 complexes were in buffer A (20 mM Tris-HCl pH 7.2, 100 mM NaCl, 10 mM potassium acetate, 2 mM TCEP). FoxA2 complexes were in buffer B (40 mM Tris-HCl pH 7.2, 120 mM NaCl, 40 mM ammonium acetate, 20 mM magnesium chloride, 2 mM TCEP). The complexes were screened for crystal growth at the IBMB-IRB Barcelona Automated Crystallography Platform (PAC) using sitting-drop vapor diffusion in the SWISSCI 96-well format 3-lens plates with 25 µL of the reservoir solutions. All crystals grew within a few days. The conditions supporting the crystal growth were as followed:

- **hFoxH1-GG complex** at 4.0 mg/mL protein concentration was mixed with 20% PEG 3350, 0.2 M ammonium chloride pH 6.3 reservoir solution at 200:100 nL reservoir to sample ratio; crystals grew at 4°C; crystals were flash cooled and stored in liquid nitrogen using cryo-solution composed of 21% PEG 3350, 0.13 M ammonium chloride pH 6.3, and 18% ethylene glycol
- **bFoxH1-GG complex** at 4.0 mg/mL protein concentration was mixed with 35% PEG Smear Low* reservoir solution at 100:200 nL reservoir to sample ratio; crystals grew at 4°C; crystals were directly flash cooled and stored in liquid nitrogen
- **bFoxH1-GT complex** at 4.0 mg/mL protein concentration was mixed with 41% PEG Smear Low* reservoir solution at 150:300 nL reservoir to sample ratio; crystals grew at 4°C; crystals were directly flash cooled and stored in liquid nitrogen
- **bFoxH1-TT complex** at 4.5 mg/mL protein concentration was mixed with 20% PEG Smear High*, 0.1 M sodium acetate pH 4.5 reservoir solution at 100:200 nL reservoir to sample ratio; crystals grew at

4°C; crystals were flash cooled and stored in liquid nitrogen using cryo-solution composed of 14% PEG Smear High*, 0.07 M sodium acetate pH 4.5, 18% glycerol, and 12% PEG 400

- **xFoxH1-GG complex** at 4.0 mg/mL protein concentration was mixed with 30% PEG 8000, 0.2 M ammonium sulfate, 0.1 M sodium cacodylate reservoir solution at 100:200 nL reservoir to sample ratio; crystals grew at 20°C; crystals were flash cooled and stored in liquid nitrogen using cryo-solution composed of 21% PEG 8000, 0.14 M ammonium sulfate, 18% glycerol, and 12% PEG 400
- **FoxA2-TTACT complex** at 4.0 mg/mL protein concentration was mixed with 25% PEG 3350, 0.1 M Bis-Tris pH 6.5 reservoir solution at 150:150 nL reservoir to sample ratio; crystals grew at 4°C; crystals were flash cooled and stored in liquid nitrogen using cryo-solution composed of 25% PEG 3350, 0.06 M Bis-Tris pH 5.5, and 12% glycerol and,
- **FoxA2-TT complex** at 4.0 mg/mL protein concentration was mixed with 25% PEG 3350, 0.2 M ammonium sulfate, 0.1 M Bis-Tris pH 6.5 reservoir solution at 100:200 nL reservoir to sample ratio; crystals grew at 4°C; crystals were flash cooled and stored in liquid nitrogen using cryo-solution composed of 25% PEG 3350, 0.12 M ammonium sulfate, 0.06 M Bis-Tris pH 5.5, and 12% glycerol.

*** The PEG Smears** are made by mixing PEG stocks (50% concentration) at an equal volume:

- PEG Smear Low is a mix of PEGs: 400, 500 MME, 600, and 1000.

- PEG Smear High is a mix of PEGs: 6000, 8000, and 10000.

Data collection and structure determination

Diffraction data used for the structure determination were recorded at the ALBA beamline BL13-XALOC (Barcelona, Spain) and at the ESRF beamline ID30a3 (Grenoble, France). The data were processed, scaled, and merged with autoPROC 81 applying the anisotropy correction by STARANISO [65]. The $CC_{1/2}$ criterion was used for selecting the diffraction resolution cut-off [66]. Initial phases were obtained by molecular replacement using PHASER [67,68] as part of the CCP4 and PHENIX suites (search model FoxK2, PDB code: 2c6y). REFMAC [69] phenix.refine [70] and BUSTER [71] were employed for the refinement, and COOT 24 for the manual improvement of the models. The PDB-REDO server was used for the selection of data resolution cut-off (paired-refinement), structure model optimization, and refinement [72]. UCSF Chimera [73] was used to prepare figures and calculate RMSD values for structural comparisons and Curves+ [44] for DNA analysis.

Motif analysis

We downloaded ChIP-Seq data (bed format) from the Gene Expression Omnibus (GEO) Database, with accession numbers GSM2263597 (FoxA2, stage 10.5), GSM2263590 (FoxH1, stage 8), GSM2263591 (FoxH1, stage 9) and the reanalysis of the data from GSE53652 [56] available at GEO series GSE85273 [5].

We then analyzed the presence of FoxH1 (both TGTGKATT and TGTTKATT) and FoxA2 (TRTTTAC, as described in Jaspar 2022 [74] motifs in the form of enrichment with respect to reshuffled primary sequences using SEA 5.4.1 [75] in each of the datasets. Fasta files for the analysis were generated from the bed files and *X. tropicalis* 7.1 genome (http://www.xenbase.org/, RRID:SCR_003280) using BEDTools 2.24 [76]. We used STREME 5.4.1[77], a motif discovery program, to ensure that the FoxA2 motif, as defined in JASPAR, was the most enriched in FoxA2 peaks. In fact, we obtained this motif ranked with a p-value of 1.6e-42. Default options were used for both SEA 5.4.1 and STREME 5.4.1 programs. We used BEDTools to split the FoxA2 dataset into two. Set 1 contained those peaks previously occupied by FoxH1 and set 2 those peaks that were not occupied by FoxH1. These regions were also analyzed for the presence of FoxH1 and FoxA2 binding motifs.

References

1 Massagué, J. TGF-beta signal transduction. *Annu Rev Biochem* **67**, 753-791, doi:10.1146/annurev.biochem.67.1.753 (1998).

2 Chen, X., Rubock, M. J. & Whitman, M. A transcriptional partner for MAD proteins in TGF-beta signalling. *Nature* **383**, 691-696, doi:10.1038/383691a0 (1996).

3 Macias, M. J., Martin-Malpartida, P. & Massagué, J. Structural determinants of Smad function in TGF-beta signaling. *Trends Biochem Sci* **40**, 296-308, doi:10.1016/j.tibs.2015.03.012 (2015).

4 Bogdanovic, O., van Heeringen, S. J. & Veenstra, G. J. The epigenome in early vertebrate development. *Genesis* **50**, 192-206, doi:10.1002/dvg.20831 (2012).

5 Charney, R. M. *et al.* Foxh1 Occupies cis-Regulatory Modules Prior to Dynamic Transcription Factor Interactions Controlling the Mesendoderm Gene Program. *Dev Cell* **40**, 595-607 e594, doi:10.1016/j.devcel.2017.02.017 (2017).

6 Landsberger, N. & Wolffe, A. P. Remodeling of regulatory nucleoprotein complexes on the Xenopus hsp70 promoter during meiotic maturation of the Xenopus oocyte. *EMBO J* **16**, 4361-4373, doi:10.1093/emboj/16.14.4361 (1997).

7 Vastenhouw, N. L., Cao, W. X. & Lipshitz, H. D. The maternal-to-zygotic transition revisited. *Development* **146**, doi:10.1242/dev.161471 (2019).

8 Muller, P., Rogers, K. W., Yu, S. R., Brand, M. & Schier, A. F. Morphogen transport. *Development* **140**, 1621-1638, doi:10.1242/dev.083519 (2013).

9 Aragon, E. *et al.* Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-beta signaling. *Genes Dev* **33**, 1506-1524, doi:10.1101/gad.330837.119 (2019).

10 Blitz, I. L. & Cho, K. W. Y. Control of zygotic genome activation in Xenopus. *Curr Top Dev Biol* **145**, 167-204, doi:10.1016/bs.ctdb.2021.03.003 (2021).

11 Gao, P. *et al.* Transcriptional regulatory network controlling the ontogeny of hematopoietic stem cells. *Genes Dev* **34**, 950-964, doi:10.1101/gad.338202.120 (2020).

12 Gentsch, G. E., Owens, N. D. L. & Smith, J. C. The Spatiotemporal Control of Zygotic Genome Activation. *iScience* **16**, 485-498, doi:10.1016/j.isci.2019.06.013 (2019).

13 Joseph, S. R. *et al.* Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos. *Elife* **6**, doi:10.7554/eLife.23326 (2017).

14 Larson, E. D., Marsh, A. J. & Harrison, M. M. Pioneering the developmental frontier. *Mol Cell* **81**, 1640-1650, doi:10.1016/j.molcel.2021.02.020 (2021).

15 Michael, A. K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460-1465, doi:10.1126/science.abb0074 (2020).

16 Massagué, J. How cells read TGF-beta signals. *Nat Rev Mol Cell Biol* **1**, 169-178, doi:10.1038/35043051 (2000).

17 Massagué, J. TGFbeta signalling in context. *Nat Rev Mol Cell Biol* **13**, 616-630, doi:10.1038/nrm3434 (2012).

18 Afouda, B. A. *et al.* Foxh1/Nodal Defines Context-Specific Direct Maternal Wnt/beta-Catenin Target Gene Regulation in Early Development. *iScience* **23**, 101314, doi:10.1016/j.isci.2020.101314 (2020).

19 Attisano, L. & Lee-Hoeflich, S. T. The Smads. *Genome Biol* **2**, 3010.3011-3018 (2001).

20 Hoodless, P. A. *et al.* MADR1, a MAD-related protein that functions in BMP2 signaling pathways. *Cell* **85**, 489-500 (1996).

21 von Both, I. *et al.* Foxh1 is essential for development of the anterior heart field. *Dev Cell* **7**, 331-345, doi:10.1016/j.devcel.2004.07.023 (2004).

22 Yamamoto, M. *et al.* The transcription factor FoxH1 (FAST) mediates Nodal signaling during anterior-posterior patterning and node formation in the mouse. *Genes Dev* **15**, 1242-1256, doi:10.1101/gad. 883901 (2001).

23 Lam, E. W., Brosens, J. J., Gomes, A. R. & Koo, C. Y. Forkhead box proteins: tuning forks for transcriptional harmony. *Nat Rev Cancer* **13**, 482-495, doi:10.1038/nrc3539 (2013).

24 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665, doi:10.1016/j.cell.2018.01.029 (2018).

25 Dai, S., Qu, L., Li, J. & Chen, Y. Toward a mechanistic understanding of DNA binding by forkhead transcription factors and its perturbation by pathogenic mutations. *Nucleic Acids Res* **49**, 10235-10249, doi:10.1093/nar/gkab807 (2021).

26 Chen, X. *et al.* Smad4 and FAST-1 in the assembly of activin-responsive factor. *Nature* **389**, 85-89, doi:10.1038/38008 (1997).

27 Labbe, E., Silvestri, C., Hoodless, P. A., Wrana, J. L. & Attisano, L. Smad2 and Smad3 positively and negatively regulate TGF beta-dependent transcription through the forkhead DNA-binding protein FAST2. *Mol Cell* **2**, 109-120 (1998).

28 Martin-Malpartida, P. *et al.* Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors. *Nat Commun* **8**, 2070, doi:10.1038/s41467-017-02054-6 (2017).

29 Zhou, S., Zawel, L., Lengauer, C., Kinzler, K. W. & Vogelstein, B. Characterization of human FAST-1, a TGF beta and activin signal transducer. *Mol Cell* **2**, 121-127, doi:10.1016/s1097-2765(00)80120-3 (1998).

30 Zhang, Y. *et al.* High throughput determination of TGFbeta1/SMAD3 targets in A549 lung epithelial cells. *PLoS One* **6**, e20319, doi:10.1371/journal.pone.0020319 (2011).

31 Zhang, J. *et al.* FOXH1 promotes lung cancer progression by activating the Wnt/beta-catenin signaling pathway. *Cancer Cell Int* **21**, 293, doi:10.1186/s12935-021-01995-9 (2021).

32 Loizou, E. *et al.* A Gain-of-Function p53-Mutant Oncogene Promotes Cell Fate Plasticity and Myeloid Leukemia through the Pluripotency Factor FOXH1. *Cancer Discov* **9**, 962-979, doi:10.1158/2159-8290.CD-18-1391 (2019).

33 Jimenez, G., Verrijzer, C. P. & Ish-Horowicz, D. A conserved motif in goosecoid mediates groucho-dependent repression in Drosophila embryos. *Mol Cell Biol* **19**, 2080-2087, doi:10.1128/MCB.19.3.2080 (1999).

34 Miyazono, K. I. *et al.* Hydrophobic patches on SMAD2 and SMAD3 determine selective binding to cofactors. *Sci Signal* **11**, doi:10.1126/scisignal.aao7227 (2018).

35 Jennings, B. H. *et al.* Molecular recognition of transcriptional repressor motifs by the WD domain of the Groucho/TLE corepressor. *Mol Cell* **22**, 645-655, doi:10.1016/j.molcel.2006.04.024 (2006).

36 Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* **43**, W389-394, doi:10.1093/nar/gkv332 (2015).

37 Zheng, H. *et al.* Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nat Protoc* **9**, 156-170, doi:10.1038/nprot.2013.172 (2014).

38 Chen, X. *et al.* Structural basis for DNA recognition by FOXC2. *Nucleic Acids Res* **47**, 3752-3764, doi:10.1093/nar/gkz077 (2019).

39 Li, J. *et al.* Structure of the Forkhead Domain of FOXA2 Bound to a Complete DNA Consensus Site. *Biochemistry* **56**, 3745-3753, doi:10.1021/acs.biochem.7b00211 (2017).

40 Bandukwala, H. S. *et al.* Structure of a domain-swapped FOXP3 dimer on DNA and its function in regulatory T cells. *Immunity* **34**, 479-491, doi:10.1016/j.immuni.2011.02.017 (2011).

41 Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-420, doi:10.1038/364412a0 (1993).

42 Li, J. *et al.* Mechanism of forkhead transcription factors binding to a novel palindromic DNA site. *Nucleic Acids Res* **49**, 3573-3583, doi:10.1093/nar/gkab086 (2021).

43 Rogers, J. M. *et al.* Bispecific Forkhead Transcription Factor FoxN3 Recognizes Two Distinct Motifs with Different DNA Shapes. *Mol Cell* **74**, 245-253 e246, doi:10.1016/j.molcel.2019.01.019 (2019).

44 Blanchet, C., Pasi, M., Zakrzewska, K. & Lavery, R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* **39**, W68-73, doi:10.1093/nar/gkr316 (2011).

45 Nikolova, E. N. *et al.* Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470**, 498-502, doi:10.1038/nature09775 (2011).

46 Xu, Y. *et al.* Hoogsteen base pairs increase the susceptibility of double-stranded DNA to cytotoxic damage. *J Biol Chem* **295**, 15933-15947, doi:10.1074/jbc.RA120.014530 (2020).

47 Widom, J. Chromatin structure: linking structure to function with histone H1. *Curr Biol* **8**, R788-791, doi:10.1016/s0960-9822(07)00500-3 (1998).

48 Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**, 219-223, doi:10.1038/362219a0 (1993).

49 Zaret, K. S. Pioneer Transcription Factors Initiating Gene Network Changes. *Annu Rev Genet* **54**, 367-385, doi:10.1146/annurev-genet-030220-015007 (2020).

50 Wang, S. *et al.* Linker histone defines structure and self-association behaviour of the 177 bp human chromatosome. *Sci Rep* **11**, 380, doi:10.1038/s41598-020-79654-8 (2021).

51 Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol* **276**, 19-42, doi:10.1006/jmbi.1997.1494 (1998).

52 Meers, M. P., Janssens, D. H. & Henikoff, S. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Mol Cell* **75**, 562-575 e565, doi:10.1016/j.molcel.2019.05.025 (2019).

53 Pogoda, H. M., Solnica-Krezel, L., Driever, W. & Meyer, D. The zebrafish forkhead transcription factor FoxH1/Fast1 is a modulator of nodal signaling required for organizer formation. *Curr Biol* **10**, 1041-1049, doi:10.1016/s0960-9822(00)00669-2 (2000).

54 Sirotkin, H. I., Gates, M. A., Kelly, P. D., Schier, A. F. & Talbot, W. S. Fast1 is required for the development of dorsal axial structures in zebrafish. *Curr Biol* **10**, 1051-1054, doi:10.1016/s0960-9822(00)00679-5 (2000).

55 Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947, doi:10.1093/nar/gky1015 (2019).

56 Chiu, W. T. *et al.* Genome-wide view of TGFbeta/Foxh1 regulation of the early mesendoderm program. *Development* **141**, 4537-4547, doi:10.1242/dev.107227 (2014).

57 Benabou, S., Mazzini, S., Avino, A., Eritja, R. & Gargallo, R. A pH-dependent bolt involving cytosine bases located in the lateral loops of antiparallel G-quadruplex structures within the SMARCA4 gene promotor. *Sci Rep* **9**, 15807, doi:10.1038/s41598-019-52311-5 (2019).

58 Corona, R. I. & Guo, J. T. Statistical analysis of structural determinants for protein-DNA-binding specificity. *Proteins* **84**, 1147-1161, doi:10.1002/prot.25061 (2016).

59 Konig, P., Giraldo, R., Chapman, L. & Rhodes, D. The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell* **85**, 125-136, doi:10.1016/s0092-8674(00)81088-0 (1996).

60 Jantz, D. & Berg, J. M. Probing the DNA-binding affinity and specificity of designed zinc finger proteins. *Biophys J* **98**, 852-860, doi:10.1016/j.bpj.2009.11.021 (2010).

61 Gadea, F. C. & Nikolova, E. Nucleosome topology and DNA sequence modulate the engagement of pioneer factors SOX2 and OCT4. *BioRxiv* (2022).

62 Echigoya, K. *et al.* Nucleosome binding by the pioneer transcription factor OCT4. *Sci Rep* **10**, 11832, doi:10.1038/s41598-020-68850-1 (2020).

63 Aragon, E. *et al.* A Smad action turnover switch operated by WW domain readers of a phosphoserine code. *Genes Dev* **25**, 1275-1288, doi:10.1101/gad.2060811 (2011).

64 Martin-Malpartida, P. *et al.* HTSDSF explorer, a novel tool to analyze high-throughput DSF screenings. *Journal of Molecular Biology*, 167372, doi:https://doi.org/10.1016/j.jmb.2021.167372 (2021).

65 Tickle, I. J. *et al.* STARANISO. Global Phasing Ltd., Cambridge, UK. (2018).

66 Diederichs, K. & Karplus, P. A. Better models by discarding data? *Acta Cryst. D,* **69** 1215-1222 (2013).

67 McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr D Biol Crystallogr* **63**, 32-41, doi:10.1107/S0907444906045975 (2007).

68 Medina, E. *et al.* Three-Dimensional Domain Swapping Changes the Folding Mechanism of the Forkhead Domain of FoxP1. *Biophysical journal* **110**, 2349–2360, doi:https://doi.org/10.1016/j.bpj.2016.04.043 (2016).

69 Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**, 240-255, doi:10.1107/S0907444996012255 (1997).

70 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877, doi:10.1107/S2059798319011471 (2019).

71 Smart, O. S. *et al.* Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr D Biol Crystallogr* **68**, 368-380, doi:10.1107/S0907444911056058 (2012).

72 Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213-220, doi:10.1107/S2052252514009324 (2014).

73 Pettersen *et al.* UCSF Chimera - a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

74 Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**, D165-D173, doi:10.1093/nar/gkab1113 (2022).

75 Bailey, T. L. & Grant, C. H. SEA: Simple Enrichment Analysis of motifs. *biorxiv.org* (2021).

76 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

77 Bailey, T. L. STREME: Accurate and versatile sequence motif discovery. *Bioinformatics*, doi:10.1093/bioinformatics/btab203 (2021).

78 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

79 Laskowski, R. A. & Swindells, M. B. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model* **51**, 2778-2786, doi:10.1021/ci200227u (2011).

80 Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr D Biol Crystallogr* **67**, 293-302, doi:10.1107/S0907444911007773 (2011).

Table

Table 1. X-ray data collection and refinement statistics

Structure	hFoxH1-GG	bFoxH1-GG	bFoxH1-GT	bFoxH1-TT	FoxA2-TTACT	FoxA2-TT	xFoxH1-GG
PDB entry ID	7YZB	7YZ7	7YZA	7YZC	7YZE	7YZF	7YZG
Data collection[#^]							
Beamline	ALBA BL13	ALBA BL13	ALBA BL13	ALBA BL13	ALBA BL13	ALBA BL13	ALBA BL13
Space group	$P2_1$	$C2$	$C2$	$C222_1$	$C222_1$	$C222_1$	$P2_12_12_1$
a, *b*, *c* (Å)	36.19, 78.03, 51.88	100.12, 30.09, 76.66	99.39, 29.92, 75.72	36.18, 96.56, 153.81	46.04, 92.37, 118.80	45.46, 92.70, 116.84	46.30, 78.72, 103.25
α, β, γ (°)	90.00, 100.46, 90.00	90.00, 108.70, 90.00	90.00, 108.25, 90.00	90.00, 90.0, 90.00	90.00, 90.0, 90.00	90.00, 90.0, 90.00	90.00, 90.0, 90.00
Resolution (Å)[*]	51.02-1.47 (1.62-1.47)	47.24-0.98 (1.00-0.98)	46.74-1.19 (1.25-1.19)	76.91-2.17 (2.34-2.17)	46.18-1.98 (2.17-1.98)	58.42-2.18 (2.38-2.18)	62.60-2.82 (3.10-2.82)
R_{meas} (%)	11.2 (174.8)	4.6 (136.5)	3.8 (73.4)	13.8 (125.9)	4.1 (213.4)	5.6 (177.0)	32.1 (272.1)
R_{pim} (%)	2.8 (45.4)	1.8 (68.0)	2.0 (46.9)	5.5 (48.8)	1.2 (58.4)	2.4 (74.2)	8.3 (65.8)
$I/\sigma(I)$	16.2 (1.5)	17.7 (0.9)	15.4 (1.4)	7.9 (2.1)	26.3 (1.2)	19.9 (1.4)	8.2 (1.2)
$CC_{1/2}$	0.996 (0.912)	0.999 (0.365)	0.999 (0.554)	0.995 (0.552)	0.999 (0.573)	0.998 (0.593)	0.997 (0.484)
Completeness:							
Spherical (%)	76.2 (15.1)	89.9 (20.7)	85.1 (32.1)	82.5 (21.3) 92.3	84.1 (23.2)	69.2 (15.3)	58.0 (10.4)
Ellipsoidal (%)	93.1 (50.6)	93.4 (31.5)	89.1 (41.7)	(44.7)	93.9 (53.5)	86.5 (52.2)	90.8 (68.0)
Multiplicity	16.8 (14.7)	6.1 (3.7)	3.1 (2.1)	6.3 (6.4)	13.1 (13.2)	9.9 (10.3)	14.7 (17.0)
Refinement							
Resolution (Å)[*]	51.02-1.47	47.28-0.98	46.78-1.18	76.91-2.17	46.18-1.99	58.42-2.18	62.6-2.82
No. of unique reflections	34784	105944	54975	11564	15058	8314	5541
R_{work} / R_{free} (%)	16.5/19.2	13.8/14.8	14.2/16.4	20.8/23.6	20.8/21.4	22.2/25.7	20.0/25.5
Protein (aa)	127	120	121	116	87	86	121
DNA (bp)	16	16	16	16	16	16	16
No. non-H atoms							
All	1867	1936	1873	1660	1409	1388	1639
Ligand/ion	1	1	1	8	1	1	0
Water	138	257	208	46	87	5	0
Mean B factors							
Overall	41.7	18.4	16.7	30.4	68.4	45.6	67.9
Protein	40.9	17.4	15.9	29.0	58.7	42.1	63.7
DNA	42.9	17.0	15.5	30.9	79.7	49.3	74.2
Ligand/ion	50.2	16.1	12.4	59.4	71.9	92.5	NA
Water	42.3	26.1	24.7	47.2	56.2	60.8	NA
R.M.S.D.							
Bond length (Å)	0.007	0.011	0.011	0.015	0.009	0.009	0.005
Bond angle (°)	1.29	1.54	1.60	1.74	1.20	1.17	0.76
Ramachandran:							
Favored (%)	100.0	99.2	99.2	96.4	98.8	98.8	93.3
Allowed(%)	0.0	0.8	0.8	1.8	1.2	1.2	6.7
Outliers (%)	0.0	0.0	0.0	1.8	0.0	0.0	0.0

#Data for the hFoxH1-GG structure come from five merged data sets, and from three merged data sets for the xFoxH1-GG.

^Anisotropy correction by STARANISO/autoPROC with the default setting used for the determination of the resolution cutoff [65,80].

*Values in parentheses are for highest-resolution shell. Resolution cut-off based on paired-refinement protocol implemented in the PDB-REDO server [72].

Figures

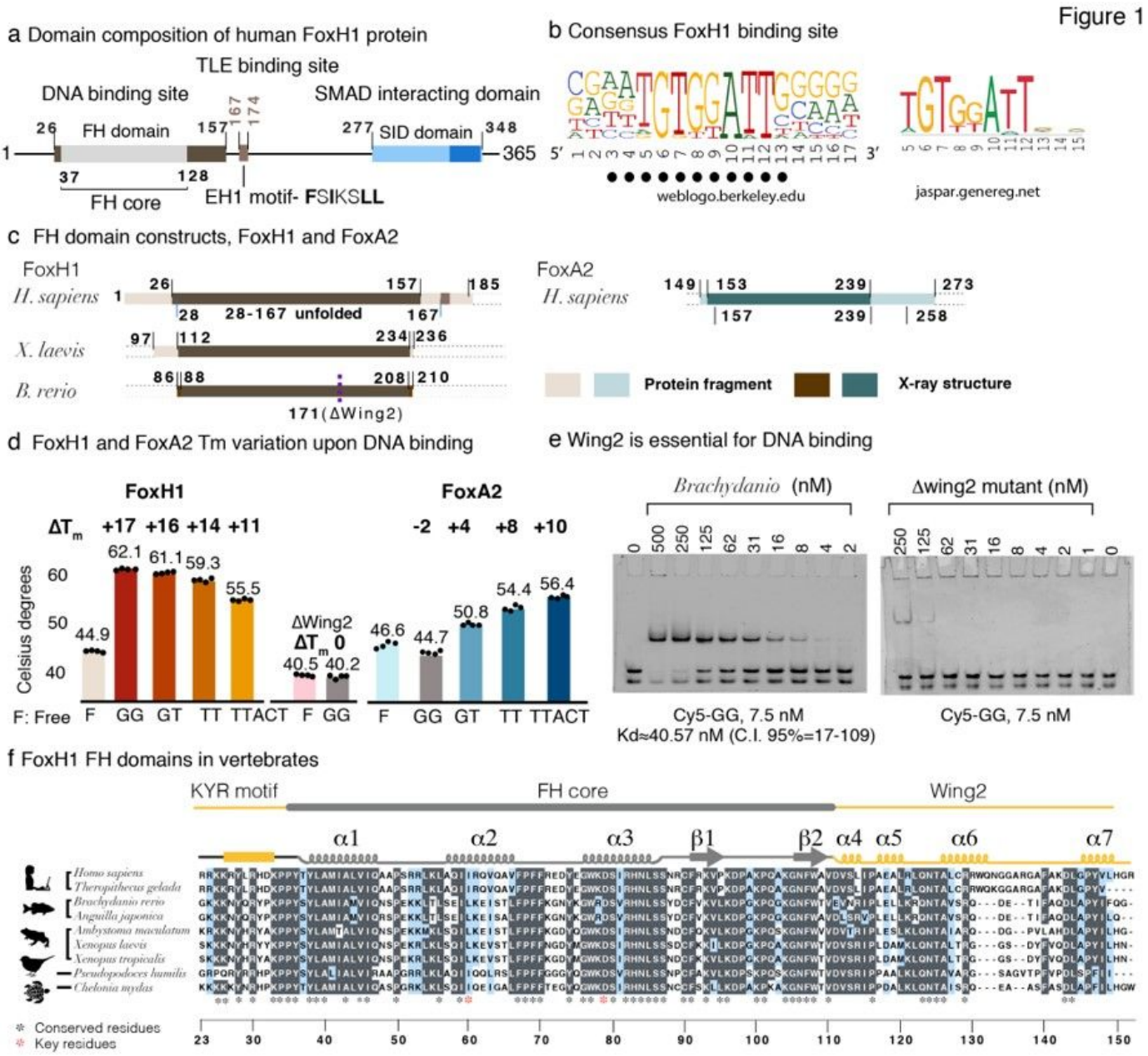

Figure 1

a. Domain composition of FoxH1 proteins. The FH and SID domains are conserved in vertebrates.

b. Consensus DNA binding. The sequences used to derive this consensus are shown in Supplementary Figure 1d. Base pairs found to participate in specific protein contacts are indicated. This consensus is almost identical to the JASPAR profile https://jaspar.genereg.net/.

c. Constructs used for the structural studies indicating the expressed proteins and the region observed in the X-ray structures. The human sequence contains the TLE binding site, but this region does not contribute to the extended FH domain structure.

d. Melting temperature modification in the presence of the different DNAs for the human FoxH1 and FoxA2 domains. The stability of the construct without the C-terminal extension (ΔWing2) is not affected by the presence of DNA. FoxA2 incubation with the GG site induced a decrease in its melting temperature. Melting temperatures correspond to two repetitions and three replicates. Values are summarized in Supplementary Table 1.

e. Titration of a 16 bp cy5-labeled DNA derived from the native Gsc sequence containing the FoxH1 GG motif, followed by native electrophoretic mobility shift assay (EMSA) at 4oC using the extended FH domain and the ΔC construct. The ΔC construct loses its DNA binding capacity compared to the WT.

Sequence alignment of FoxH1 FH domain in vertebrates. Secondary structure elements observed in the human complex are depicted on the top of the alignment. The FH core domain and the extended N- and C-termini are also indicated. Alignments were generated with MAFFT [78] and BoxShade server (https://bio.tools/boxshade).

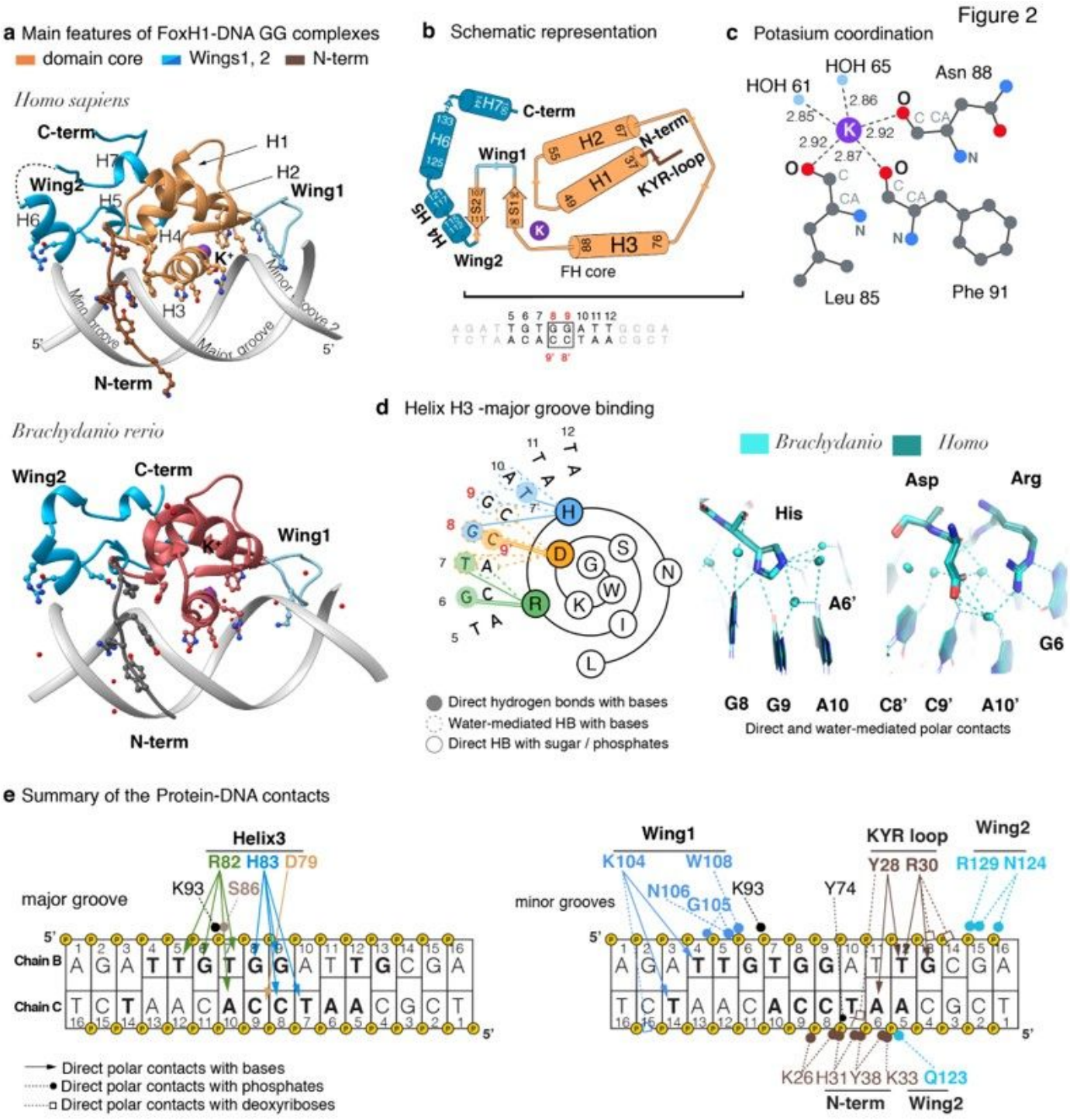

Figure 2

a. Structures of the human and zebrafish FoxH1-DNA complex bound to the GG motif displaying the crystallographic asymmetric unit. The core domain is shown in dark orange (human) and dark pink (zebrafish), whereas the N- and the extended Wing2 are colored in brown and cyan, respectively.

b. Schematic drawing of the secondary structure elements and the DNA sequence.

c. The potassium cation (K^+) is coordinated by four carbonyl oxygen atoms from the polypeptide backbone corresponding to Leu85, Asn 88, Phe91 and two well-ordered water molecules. The ion is indicated as a violet sphere (Ligplot+ V2.2) [79].

d. 2D Wenxiang diagram (H3) showing the side chains and specific base contacts with the major groove. Direct and water-mediated HBs with bases and contacts with sugar and phosphates are indicated. Snapshots highlighting direct and water-mediated polar contacts with nucleobases (distance up to 3.6 Å). In both complexes the side chain orientation and the contacts are almost identical.

e. Cartoon representation of the protein DNA interactions observed for the human FoxH1-GG complex. FoxH1 recognizes a DNA segment of 15 bp, through a rich network of direct and water-mediated contacts

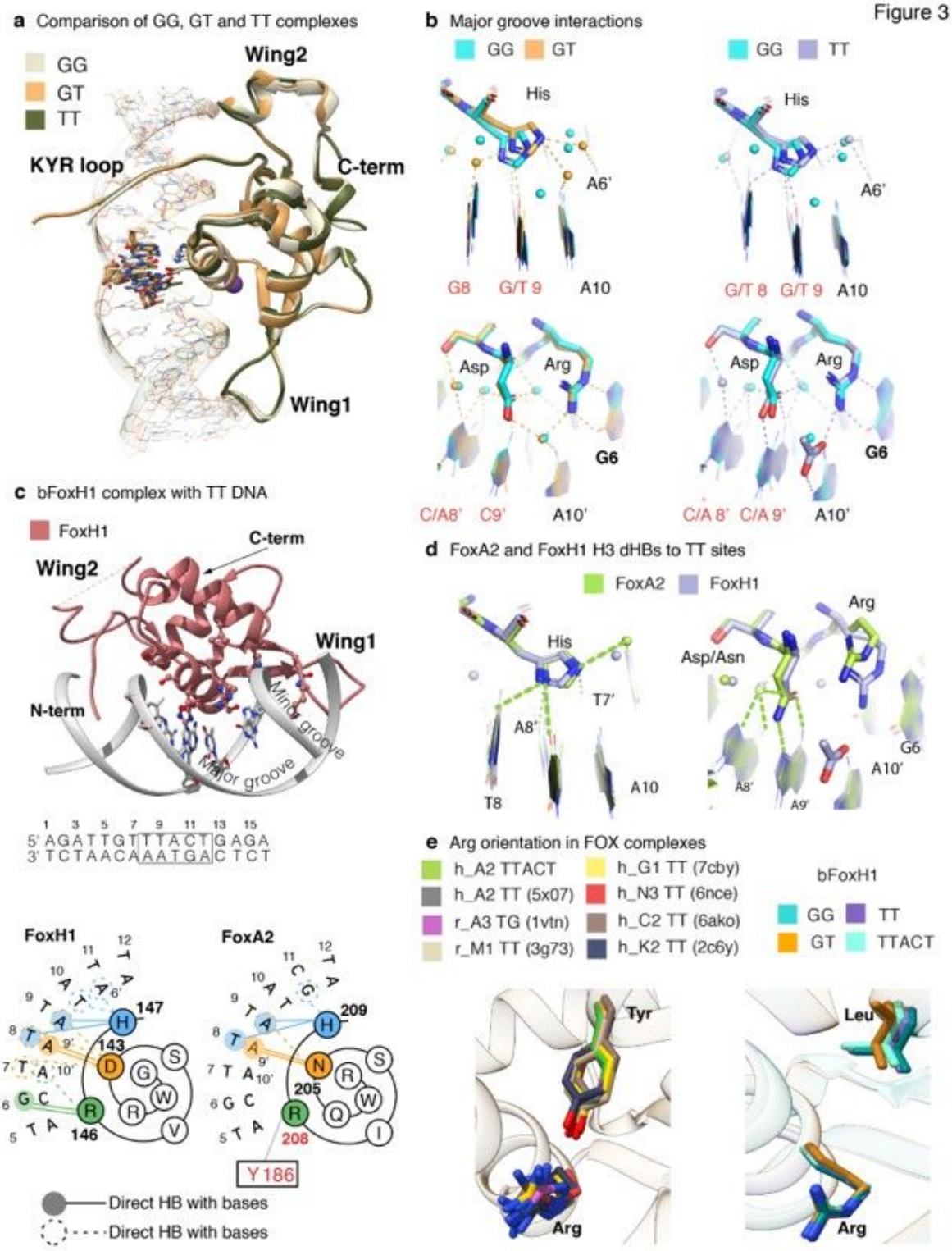

Figure 3

a. Superposition of the three FoxH1-DNA complexes with the GG, GT and TT sites.

b. Snapshots highlighting direct and water-mediated polar contacts with nucleobases (distance up to 3.6 Å). The contacts of Asp and Arg with the DNA are similar to the GG structure because they interact with the common part of the three DNA molecules although in the TT complex the Asp side chain loses the

solvent-based contact with A10'. The main differences concentrated at the His residue, which binds to the modified sites. In both GT and TT complexes, the histidine directly contacts T7', T9 and A8', and also to G8, A6' and A10 in the GT complex whereas in the TT complex, the contacts with A6' and T9' are water-mediated. Additionally, in the TT complex, the His forms a HB with T8 and there is a new water-mediated contact between Ser-A8'. Close up view of the H3-DNA dHB for the GT complex.

c. FoxH1 (coral) complex bound to the TT site. The interactions of H3 with the DNA are summarized using a 2D Wenxiang diagram and compared to those of FoxA2. Cartoon representations of the FoxA2-DNA complexes and the superposition to the FoxH1 structure are shown in Supplementary Figure 3.

d. Comparison of the specific HBs between the H3 and the TT site for both FoxH1 (side chains shown in light blue) and FoxA2 complexes (light green) determined in this work. In FoxA2, the bidendate Arg-Gua6 HB and also Ser-H2O-A8' and water-mediated contacts between the Asn side chain with T7 and T8 are lost.

e. Different rotamers of the Arg residue in eight FOX complexes (PDB entries indicated) and in FoxH1. In FoxH1, the Arg residue is close to the DNA and participates in direct contacts, whereas in the remaining complexes the side chain is rotated away and does not contact DNA efficiently (Supplementary Figure 3d).

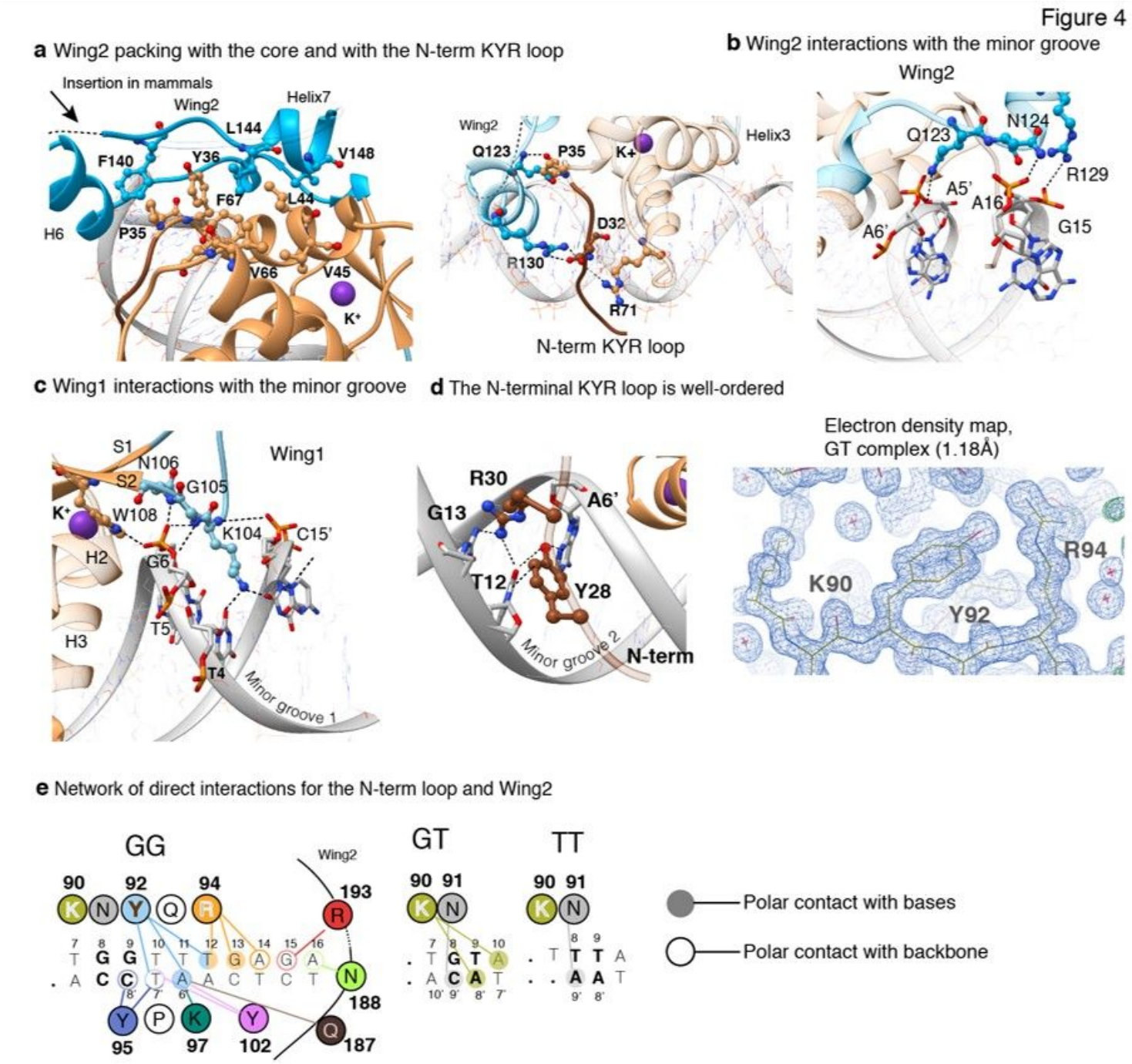

Figure 4

a. Two views of the Wing2 and the KYR loop displaying side chain–side chain contacts contributing to the distinctive FoxH1 FH fold.

b. Wing2 binding to minor groove 1. HBs are indicated with dashed lines. Protein residues, nucleotides and phosphates involved in the interaction are shown and labeled.

c. DNA contacts observed from Wing1 to minor groove 2. The 2Fo-Fc map contoured at 1.0 sigma is shown for the GT complex refined at 1.8Å.

d. The KYR loop is well-ordered in the three GG, GT and TT complexes.

e. Schematic representation with the summary of all direct protein-DNA contacts of the KYR loop. Specific differences are indicated as a cartoon.

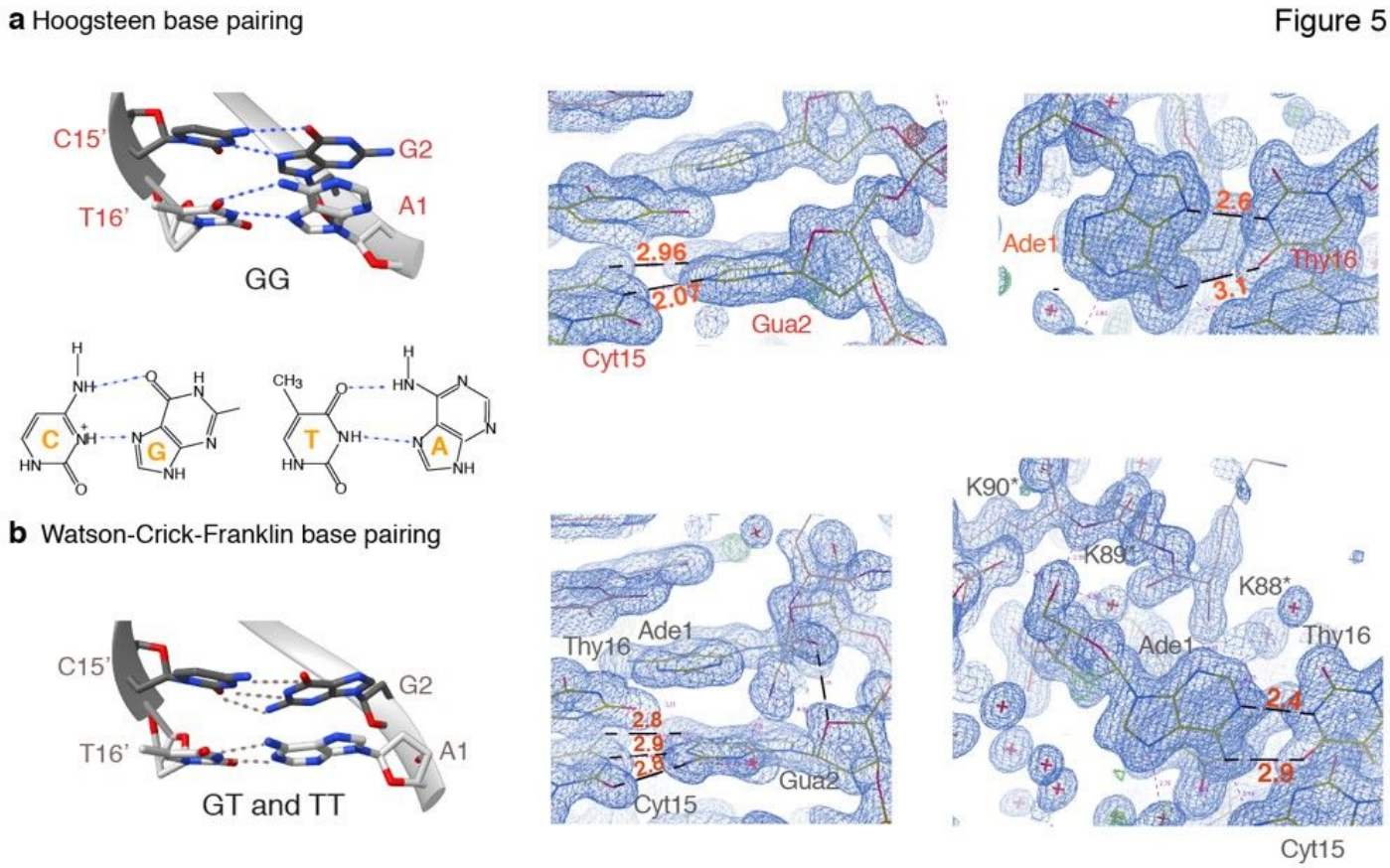

Figure 5

a. Comparison of Watson-Crick-Franklin (WCF) bps observed in the GT and TT complexes to the Hoogsteen bps observed in the GG complex. Two snapshots showing the crystal environment surrounding the two bps of the bFoxH1-GG structure and polar contacts up to 3.3 Å distances, which form HG bps (Ade1:Thy16 and Gua2: Cyt15).

b. The Watson-Crick-Franklin (WCF) bps. Larger regions of the maps are shown as Supplementary Figure 5c,d. Both DNA structures originate from crystals grown in similar conditions and identical space groups, and they have nearly identical cell units. In all cases, the 2Fo-Fc maps are contoured at 1.0 sigma.

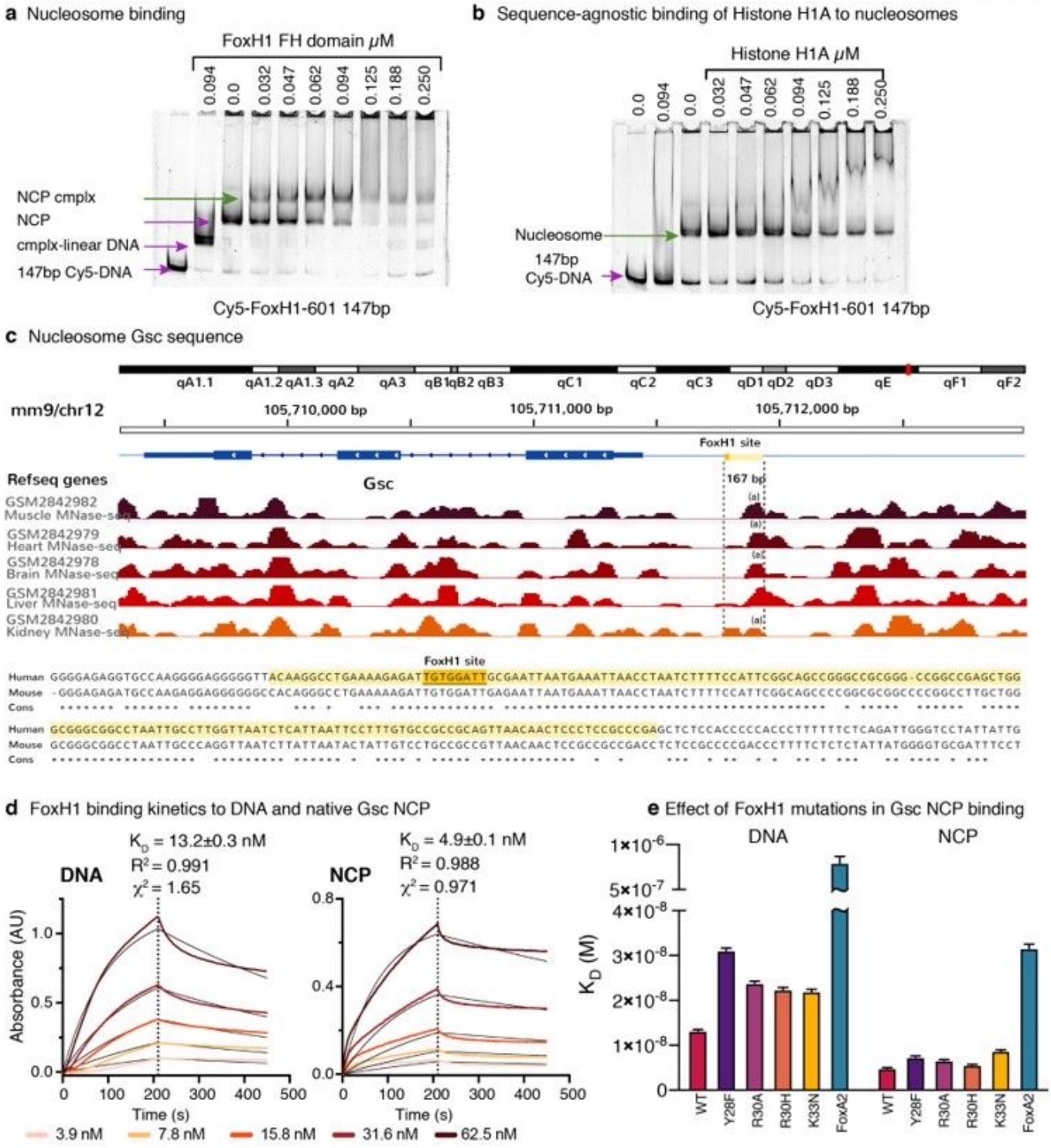

Figure 6

a. Titration of the Cy5-labeled Widom601-FoxH1 NCP-147 bp with FoxH1 FH domain followed by EMSA. Solid arrows indicate the different species.

b. Sequence-agnostic binding of the Histone H1A FH domain and the same 146 bp NCP used in panel a.

c. Native nucleosome selection. *Gsc*167 boundaries (dashed lines) were selected based on the MNase seq information described in GSM2842982 (mouse). The boundaries are more similar in Muscle, Heart, Brain and Liver than in Kidney datasets, with differences perhaps indicating NCP dynamics. For the *Gsc* sequence, we have used the (a) site. The *Gsc* human NCP sequence (in light yellow) contains the FoxH1 site (highlighted in orange). Both human and mouse sequences are highly conserved.

d. Binding kinetics between FoxH1 FH domain and either free *Gsc*167 or nucleosomes using Biolayer Interferometry.

e. Effect of point mutations of the KYR loop in DNA binding affinity. Native *Gsc* NCP and the GG site were used in these experiments. All WT and FoxH1 mutated proteins bind well to DNA and NCP, although the mutations show a 2-3 fold affinity reduction.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesandFigures.docx