

TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF- β signaling

Ewelina Guca¹, David Suñol¹, Lidia Ruiz¹, Agnieszka Konkol¹, Jorge Cordero¹, Carles Torner¹, Eric Aragon¹, Pau Martin-Malpartida¹, Antoni Riera^{1,2} and Maria J. Macias^{1,3,*}

¹Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldri Reixac, 10, Barcelona 08028, Spain, ²Departament de Química Inorgànica i Orgànica, Secció de Química Orgànica, Universitat de Barcelona, Martí i Franquès 1-11, 08028, Barcelona, Spain and ³ICREA, Passeig Lluís Companys 23, 08010-Barcelona, Spain

Received May 16, 2018; Revised June 26, 2018; Editorial Decision July 14, 2018; Accepted July 17, 2018

ABSTRACT

TGIF1 is a multifunctional protein that represses TGF- β -activated transcription by interacting with Smad2-Smad4 complexes. We found that the complex structure of TGIF1–HD bound to the TGACA motif revealed a combined binding mode that involves the HD core and the major groove, on the one hand, and the amino-terminal (N-term) arm and the minor groove of the DNA, on the other. We also show that TGIF1–HD interacts with the MH1 domain of Smad proteins, thereby indicating that TGIF1–HD is also a protein-binding domain. Moreover, the formation of the HD-MH1 complex partially hinders the DNA-binding site of the complex, preventing the efficient interaction of TGIF1–HD with DNA. We propose that the binding of the TGIF1 C-term to the Smad2-MH2 domain brings both the HD and MH1 domain into close proximity. This local proximity facilitates the interaction of these DNA-binding domains, thus strengthening the formation of the protein complex *versus* DNA binding. Once the protein complex has been formed, the TGIF1-Smad system would be released from promoters/enhancers, thereby illustrating one of the mechanisms used by TGIF1 to exert its function as an active repressor of Smad-induced TGF- β signaling.

INTRODUCTION

Smad transcription factors comprise a family of highly conserved proteins that are the main signal transducers of the Transforming Growth Factor beta (TGF- β) signaling pathways. These pathways, including those regulated by Bone

morphogenetic proteins (BMPs), activin, nodal and others, play essential roles in embryo development, differentiated tissue homeostasis and immune responses (1). In humans, mutations in several core components of TGF- β signaling inactivate the tumor suppressor function of the network and facilitate cancer cell survival (2–5). The eight members of the Smad family are divided into three distinct sub-types: receptor-activated (R-Smads), inhibitory (I-Smads), and common (Co-Smad or Smad4). R-Smads and Co-Smads carry a DNA-binding domain (MH1) and a protein–protein interaction region composed of the linker and the MH2 domain (reviewed in (5,6)). Upon activation, R-Smads form trimers with Smad4 (Co-Smad) and translocate into the nucleus, where they interact with *cis* regulatory elements and with transcription factors, co-activators and co-repressors. They also interact with ubiquitin ligases, which label them for degradation (7–11). A large body of biochemical results have illustrated several steps of the signaling process, which starts with the transmission of signals from the TGF- β receptor and ends with the induction or inhibition of the expression of specific genes. However, the specific requirements of Smad protein interactions with cofactors and repressors have not been characterized to date. Such information would open up possibilities for drug design and chemical biology-based studies addressing Smad-mediated signaling pathways.

To characterize the interaction of Smad proteins with cofactors, we focused on the 5'TG3' interacting factor 1 (TGIF1), whose interaction with R-Smad-Smad4 complexes represses TGF- β signaling (12,13). Given that Smad and TGIF1 proteins bind directly to DNA, we also looked into how these protein-protein complexes affect DNA binding. In fact, TGIF1-DNA binding occurs through the conserved homeobox DNA-interacting domain (homeodomain, HD), located at the N-terminus of TGIF1 (14). HDs are three-helical bundle structures that are widely dis-

*To whom correspondence should be addressed. Tel: +34 934037189; Email: maria.macias@irbbarcelona.org

tributed in animals, fungi and plants (Pfam entry:PF00046) (15). TGIF1–HD belongs to a large superclass defined by the presence of a three-amino acid loop extension (TALE) between the first two helices (Figure 1A) (16). The middle and last helices constitute the helix-turn-helix (HTH) motif, with the last helix, $\alpha 3$, being dedicated to interactions with DNA. TGIF1 recognizes the 5'-TGACA-3' motif of DNA. This property is unusual for HD proteins, which normally show a preference for TA-rich motifs (15). In addition to the HD, TGIF1 also contains a C-terminal region, which harbors two protein interaction sites. The first site, named C-term domain, interacts with the MH2 domain of Smad proteins and with other cofactors (shown in beige, Figure 1B) (13). The second site, which corresponds to the most C-term part of TGIF1, is a binding site for the E3 ubiquitin ligase Fbxw7, a component of the SCF complex (SKP1, CUL1 and F-box protein), which targets TGIF1 for degradation (red, Figure 1B) (17).

To characterize the interactions between Smad and TGIF1 proteins, we first analyzed the structural properties of the domains present in TGIF1. We started with the complex structure of TGIF1–HD bound to DNA (at 2.4 Å resolution). The DNA-binding interface is defined by specific hydrogen bonds (HBs) from conserved residues of the $\alpha 3$ helix, and the DNA major groove, composed of the 5'-TGACA-3' motif. In addition, three Arg residues (165–166–167), located at the N-term of the domain, participate in crucial interactions with the minor groove. These additional interactions contribute to the stability and specificity of the complex, increasing the number of base pairs (bp) recognized by TGIF1–HD outside the TGACA site to eight.

We also observed that the recombinant C-term region of TGIF1 is mostly unstructured in solution and that this region participates in weak interactions with the Smad2 MH2 domain *in vitro* when using isolated domains. We also found that the HD and C-term domains interact weakly with one another *in vitro*, suggesting that full-length TGIF1 protein presents an equilibrium of active-open and inactive-closed conformations. Since TGIF1 is an active repressor of Smad signaling and both proteins form stable and functional complexes in cells, we hypothesized that other domains in the proteins (in addition to the C-term and MH2 regions) provide secondary interactions between the proteins, perhaps acting in a synergic manner. In fact, we found that TGIF1–HD interacts with the MH1 domains of Smad2 and Smad4 proteins and that these interactions affect the access of the protein complexes to DNA.

In the context of full-length complexes and native conditions, we believe that the interaction of the C-term region of TGIF1 with the MH2 domain of Smad2 stabilizes the open conformation of TGIF1 and favors its interaction with DNA. However, binding to the MH2 domain also brings the MH1 domain of Smad proteins closer to the HD domain of TGIF1, thereby promoting the interaction between the DNA-binding domains. The formation of a HD-MH1 complex would reduce access of the Smad–TGIF1 system to DNA, thus providing an elegant mechanism for the repression of Smad-driven gene regulation.

MATERIALS AND METHODS

Cloning, expression and purification

The cDNA encoding the human TGIF1 protein (Isoform 2, Uniprot entry Q15583) was used as a template for cloning all domain constructs (pCMV5 Flag-TGIF1, Addgene Plasmid #14047, provided by the laboratory of Dr J. Massagué). All constructs were amplified by PCR.

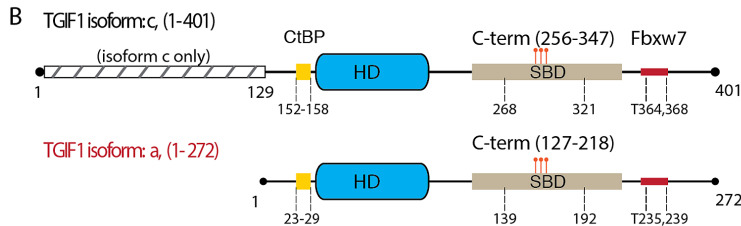
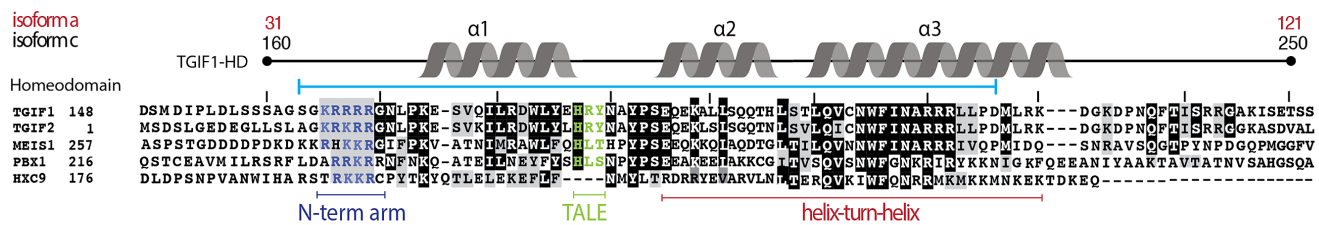
HD fragments were cloned into pCoofy18 vector (Addgene Plasmid# 43975) using RecA_f recombinase (New England Biolabs, Massachusetts, USA), including kanamycin resistance, T7-lactose protein promoter, N-terminal 10xHis tag and 3C protease cleaving site between the tag and the protein. R167A/R168A mutants were prepared by site-directed mutagenesis using QuikChange site-directed mutagenesis kit (Stratagene) and oligonucleotide primers containing either one or two Arginine to Alanine mutations. C-term fragments were cloned into the pETM11 vector (EMBL, Heidelberg), which has a TEV cleavage site between the N-His tag and the protein.

The Smad2-EEE clone (188–467 fragment) (Isoform 1, Uniprot entry Q15796) was a gift from Dr Zinn-Justin (CEA, Gif-sur-Yvette, France). The insert was cloned into a pTEM10 vector (EMBL, Heidelberg), which contains an N-term 6xHis tag and no cleaving site between the tag and the insert.

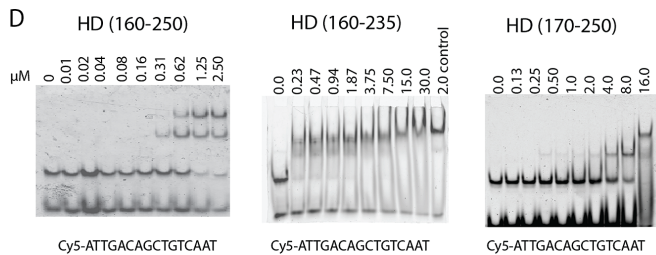
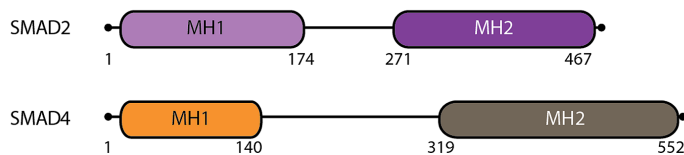
Smad4MH1(10–140) Uniprot (Q13485) and Smad2MH1(10–174) Uniprot (Q15796) domains were cloned into pETM11 vectors. Both pTEM10 and pETM11 vectors were a gift from Gunter Stier.

All clones were confirmed by DNA sequencing (Macrogen, Amsterdam, The Netherlands) and expressed in Rosetta (Novagen, Darmstadt, Germany) *Escherichia coli* cells were grown using either Luria-Bertani (LB) or Terrific Broth (TB) media for unlabeled samples. Minimal media (M9) containing D₂O (99.95%, Silantes), ¹⁵NH₄Cl and/or D-[¹³C] glucose (Cambridge isotopes) were used as sole hydrogen, nitrogen, and carbon sources to prepare the labeled samples. Bacterial cultures were grown at 37°C, induced at OD₆₀₀ 0.8 with 0.2–0.5 mM IPTG, and left overnight at 20°C. The cultures were then centrifuged at 4000 g for 10 min, and the pellet was re-suspended in lysis buffer (20 mM Tris/HCl pH 7.4, 500 mM NaCl), supplemented with lysozyme, DNase I and PMSF. The bacteria were lysed mechanically using an EmulsiFlex-C5 (Avestin). The soluble supernatants were purified by nickel-affinity chromatography on Ni²⁺ affinity resin (ABT Beads, Madrid, Spain), previously equilibrated with the lysis buffer. During the washing step, a solution of 1 M NaCl was added in order to remove the non-specifically bound DNA. Proteins were eluted with 500 mM imidazole, digested with 3C protease or TEV (at 4°C overnight), and purified further by size-exclusion chromatography on a HiLoad™ Superdex 75 16/60 prepgrade column (GE Healthcare) equilibrated in 20 mM Sodium Phosphate (pH 6.4), 150 mM NaCl and 1 mM TCEP (Melford UK). For ITC or crystallization experiments of HD constructs, the last step of purification was performed using 40 mM Tris/HCl buffer pH 7.4, 150 mM NaCl and 1 mM TCEP. The purified proteins were ver-

A Sequence comparison of human Homeodomains



C Smad proteins



E ITC titrations of HD constructs with the ATTGACAGCTGTCAAT site

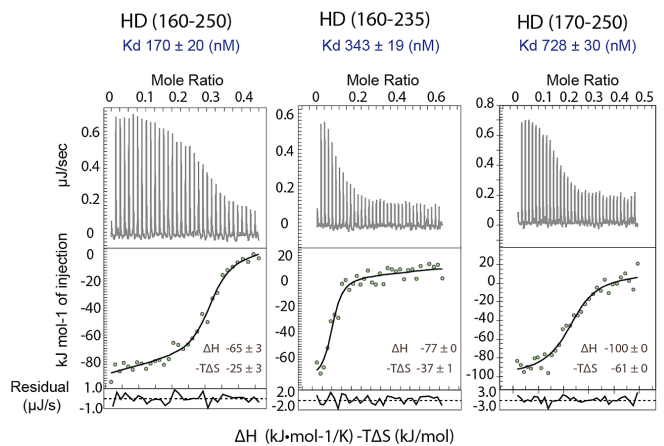


Figure 1. Definition of the TGIF1-HD, sequence comparison and DNA binding. (A) Sequence comparison of human homeodomains generated using T-Coffee, conserved residues are highlighted as black/gray boxes. The boundaries used for the structural determination and the elements of secondary structure observed in TGIF1-HD are shown and labeled. The conservation of the N-term arm (dark blue), the TALE motif (green), the helix-turn-helix (red), and the canonical length of the HD (cyan) are shown. (B) Schematic representation of TGIF1-HD and C-term constructs used in this work. The phosphorylation sites described for the C-term region are shown as red bars. Numbers correspond to a and c isoforms. (C) Schematic representation of full-length Smad2 and Smad4 proteins, with the domain boundaries specified. (D) EMSA titrations of the three selected HD constructs shown in A. The three experiments were carried out using the same DNA for comparison. Concentrations are shown in μM . (E) K_D , ΔH and $-T\Delta S$ values determined for the three selected HD constructs using the canonical DNA 5'-ATTGACAGCTGTCAAT-3'. Experiments were acquired at 20°C in Tris/HCl buffer, pH 7.4. Data were fitted using the independent model assuming a single binding site. Differences in affinity values with respect to the estimated values observed in the EMSAs are probably due to inaccuracy measurements of the Cy5-DNA used in the gel shift assays.

ified by MALDI-TOF mass spectrometry. Concentrations were measured at $A_{280\text{ nm}}$ and $A_{205\text{ nm}}$ using a NanoDrop-One system, (Thermo Fisher Scientific, Waltham, USA).

DNA annealing

Duplex DNAs were annealed using complementary single-strand HPLC-purified DNAs (Metabion, Germany). ssDNAs were desalted and dissolved in annealing buffer (10 mM Tris/HCl pH 7.4, 20 mM NaCl), mixed at 1 mM final concentration, heated at 95°C for 3 min, and cooled for 2 h at room temperature. DNAs were aliquoted and stored at -20°C. Concentrations were measured at $A_{260\text{ nm}}$ using a NanoDrop-One system, (Thermo Fisher Scientific, Waltham, USA).

Electrophoretic mobility shift assay (EMSA)

EMSA assays were carried out with all TGIF1-HD constructs and their canonical DNA was labeled with Cy5 fluorophore (Cy5-ATTGACAGCTGTCAAT) (at 3.75 and 7.5 nM concentrations). Reactions were carried out for 30 min at room temperature in 10 μl of binding buffer (40 mM Tris pH 7.2, 150 mM NaCl, 1 mM TCEP). Electrophoreses were performed in non-denaturing 12.0% (19:1) polyacrylamide gels. The gels run for 1 h in 1 \times TG buffer (25 mM Tris, pH 8.4, 192 mM Glycine) at 90 V at 20°C. Competition assays with Smad MH1 domains were performed using increasing amounts of either Smad proteins or TGIF1-HD, using two different cy5-labeled oligos. The gels were exposed to a Typhoon imager (GE Healthcare, Uppsala Sweden) using a wavelength of

678/694 nm (excitation/emission maximum) for the Cy5 fluorophore. Bands were quantified using ImageJ. The Cy5 oligos were 5'-ATTATTGCGCCGAAACGCA-3' and 5'-AAGCCGTCGGGGCCGCGCCGGGGCCGGAACGA-3'.

EMSA-western blot

To detect the presence of Smad4 MH1 in the competitive assays, we transferred the proteins from the native acrylamide gel to a polyvinylidene difluoride (PVDF) membrane. The efficiency of the transfer step was validated using REVERT (Total Protein Stain Kit, LI-COR Biosciences, 926-11010). The stain was removed using the REVERT Reversal Solution (LI-COR Biosciences, 926-11013). After the addition of blocking buffer, the membrane was incubated with the anti-Smad4 antibody (abcam, ab208804, Smad4-antibody, N-terminal, rabbit), and the secondary antibody was labeled with horseradish peroxidase (HRP). Target proteins were detected using the Odyssey Fc Digital Imaging System. Bands were quantified using ImageJ.

Isothermal titration calorimetry (ITC)

All ITC measurements were carried out with protein and DNA samples dissolved in the same buffer (40 mM Tris pH 7.4, 150 mM NaCl and 1 mM TCEP), and degassed. Protein and DNA concentrations were determined using a NanoDrop ONE system and their predicted extinction coefficients (ProtParam EXPASy). The DNA concentration was determined assuming that all DNA is present as dsDNA. The ITC measurements were performed at 20°C using a Nano ITC calorimeter (TA Instruments). The NanoAnalyze software (TA Instruments) was used to study the binding isotherms. Baseline controls were acquired with buffer and pure DNA solutions. Titrations were performed as a set of 34 injections of DNA (0.1–0.15 mM) into protein (0.06–0.07 mM). Each ITC titration was performed at least in duplicates. Fittings were performed using the independent binding sites model.

Nuclear magnetic resonance (NMR)

NMR data of the HD and C-term domain of TGIF1 were recorded on a Bruker Avance III 600-MHz spectrometer (IRB Barcelona) or on a 750 MHz (Leibniz Institute of Molecular Pharmacology, FMP Berlin) equipped with quadruple (^1H , ^{13}C , ^{15}N , ^{31}P) resonance cryogenic probe heads and z-pulse field gradient unit.

Backbone ^1H , ^{13}C and ^{15}N resonance assignments for the domains were obtained at 277 K (C-term) by analyzing the 3D CBCANH, HNCANH and HN(CO)CACB experiments. Experiments were acquired as Band-Selective Excitation Short-Transient-type experiments (BEST) with TROSY and Non-Uniform Sampling (NUS) (18,19). The protein concentration was 200 μM , in 20 mM HEPES, 80 mM NaCl, 2 mM DTT, 100 μM PMSF at pH 6.4 and 5% of D_2O . Spectra were processed with TopSpin v3.5 Bruker Software and were evaluated with CCPNMR Analysis and CARA. Chemical shifts of the TGIF1 C-term domain have been deposited in the Biological Magnetic Resonance Data

Bank, BMRB entry 2746. Assignment of the HD was publicly available at the BRMB, entry 17971. We adjusted the variations caused by differences in sequence boundaries and experimental conditions to our data.

For the titration experiments, a Non-Uniform Sampling (NUS) acquisition strategy was also used to reduce experimental time and increase resolution. 2D HSQC experiments were performed using 100 μM HD protein concentration in 20 mM HEPES, 150 mM NaCl, 1 mM TCEP at pH 5.25, 10% D_2O at 298 K.

p38 α phosphorylation was performed by adding ATP (at 2 mM final concentration), 5 mM MgCl_2 and 10 μl p38 α at 0.1 $\mu\text{g}/\mu\text{l}$ (SignalChem, Richmond, Canada) for every 100 μl of protein solution in 20 mM HEPES, 100 mM NaCl, 2 mM DTT at pH 6.8 and 5% D_2O . The reaction was incubated overnight at 298K without stirring.

p38 α phosphorylation of TGIF1 (256–339, 80 μM) was monitored by real-time NMR experiments (2D HSQC at 277 K), and 21 non-stop SOFAST-HMQC experiments (34 min 42 s each) were recorded at 298 K. The intensities of each ^1H - ^{15}N unambiguously assigned peak were fitted using the mono-exponential Equation (1) with the GraphPad Prism software.

$$(y = y_0^*(1 - \exp(-K*x))) \quad (1)$$

The final product was characterized by analyzing a pair of 3D CBCANH-BEST, HNCANH-BEST experiments using 200 μM protein (20 mM HEPES, 100 mM NaCl, 2 mM DTT, 1.5 mM EDTA, 100 μM PMSF at pH 6.4). 5% D_2O was added to a final volume of 320 μl .

Further phosphorylation with Casein Kinase I (CK1) was done by adding 2 mM ATP, 1x protein kinase buffer (provided by the manufacturer) and 8 μl of CK1 at 1 000 000 U/ml (New England Biolabs, Massachusetts, USA) for every 100 μl of protein solution. The reaction was incubated overnight at 298 K.

The phosphorylation reactions were also confirmed by Mass Spectrometry, using a MALDI-ToF/ToF MS instrument (4700 Proteomics Analyzer), since the addition of each phosphate group represents an increment of 78 Da to the total protein MW. p38 α phosphorylates the ^{15}N HD sample at two positions (as we observed an increment of 156 Da in the protein mass, (from 10001.0 up to 10157.4 Da). Phosphorylation by CK1 increases the MW further, up to 10236.4 Da (only one site is phosphorylated). The phosphorylated sites were specifically identified by triple resonance NMR spectroscopy.

X-ray data collection and refinement

TGIF HD 150–248 and HD 161–250 R167A/R168A were concentrated to 10 $\text{mg}\cdot\text{ml}^{-1}$ prior to the addition of the DNAs dissolved in the annealing buffer. The final protein:DNA ratios were 1:2 and 1:3, respectively. Screenings and optimizations were prepared by mixing 100 nl of the complex solution and 100 nl of reservoir solution in 96-well plates. Crystals of the complexes were grown by sitting-drop vapor diffusion at 4°C. The best diffracting crystals were obtained in: 0.1 M sodium acetate tri-hydrate pH 4.5, 30% (v/v) PEG 1500 and in 0.2 M L-Proline, 0.1M HEPES pH 7.5, 24% (v/v) PEG 1000 in crystallization buffer for

DNA complexes of TGIF HD 150–248 and HD 161–250 R167A/R168A mutant, respectively.

Crystals were cryoprotected in mother liquid solution supplemented with 25% PEG 500 MME in both cases and frozen in liquid nitrogen. The X-ray data were collected at 100 K at the beamlines ID23-1 and ID29 at the ESRF Grenoble, France. The data were scaled using XDS, Scala and Truncate from the CCP4 software and phased using molecular replacement with structure of MEIS HD complex with DNA (PDB ID: 4XRM) as a model search, the statistics are listed in Table 1. The structures were refined using Coot and Phenix (20,21) and graphical representations were prepared using PyMOL and Chimera (22,23).

Peptide synthesis

The phosphorylated peptide was prepared using Fmoc solid-phase peptide synthesis by manual coupling (0.10–0.15 mmol scale). Crude peptides were purified by RP-HPLC using an AKTApurifier10 (GE Healthcare) or a HPLC delta 600 system (Waters) and the C4 Sephasil and/or SunFire C18 columns. Fractions containing the desired peptides were identified by MALDI-TOF mass spectrometry, lyophilized, and stored at -20°C .

HD models with HPE mutations

All models were generated using the WT crystal structure as the template and applying molecular dynamic simulations (MDS) using openMM and Amber14 as force field. The settings for the MDS were as follows: Langevin integrator, at 300 K, 1 ps \wedge -1 Friction coefficient, 100 000 steps of 0.002 ps, for a total simulation time of 200 ps. All models were generated in the absence of DNA, since functional information for several of these mutated proteins indicated a decreased DNA binding capacity when compared to the WT. As a control, we also ran MDS for the WT in the absence of DNA. In this case, the structure of the HD is maintained unaltered during the simulation.

RESULTS

TGIF1–HD binds the TGACA motif with high specificity

The presence of Arg and Lys residues at the N-term of HDs is common, especially in HDs that belong to the TALE sub-family. In the case of TGIF1, several Arg and Lys residues surround the HD (Figure 1A). To characterize the potential roles of these residues in the specific recognition of the DNA motif, we designed two constructs, HD (160–250) and HD (151–248), with positively charged residues at both termini, and two additional constructs, HD (171–250) and HD (160–235), with and without the charged N- or C-term residues, respectively (Figure 1B). Binding was performed using electrophoretic mobility shift assays (EMSAs), by titrating a constant concentration of Cy-5-labeled dsDNA containing the 5'-TGACA-3' motif (14), and various recombinant protein constructs. These experiments (Figure 1D) revealed a loss of apparent affinity for DNA when the N-term truncated HD (171–250) construct was titrated. The remaining samples, HD (160–250) and HD (160–235),

bound DNA with similar affinities. Using Isothermal Titration Calorimetry (ITC), we observed that HD (160–250), containing the extended HD at both termini, had the lowest dissociation constant of all the samples (K_D of 169.9 ± 20 nM). In contrast, the constructs lacking either the positively charged motif preceding the HD core (HD (171–250)), or the residues after (HD (160–235)), bound to the same dsDNA four or two times weaker than the WT (728.1 ± 3 and 343.2 ± 20 nM), respectively (Figure 1E).

Overall, these binding experiments reveal that the efficient interaction of with DNA requires the participation of the HD core and that the flanking (positively charged) residues surrounding the domain enhance the global affinity of the interaction.

TGIF1–HD interacts with both major and minor grooves

To decipher all key interactions with DNA, we tested a range of crystallization conditions, varying TGIF1–HD constructs and dsDNA lengths. All DNAs share the presence of the TGACA site. High-resolution diffracting crystals were obtained with HD construct (151–248) bound to a 16mer palindromic DNA, which is the first crystal structure of a TGIF1–HD bound to DNA determined to date. The structure was refined at 2.42 Å resolution and determined by molecular replacement using Meis1-HD bound to DNA (PDB ID: 4XRM) as a model. The structure of TGIF1–HD bound to DNA, space group $P2_12_12_1$, contains two TGIF-HD monomers and one dsDNA per asymmetric unit (ASU). Residues 168–234 (chain A), 164–229 (chain B), and the 16mer dsDNA are well-defined in the electron density map (Supplementary Figure S1). In the biological unit, the second protein monomer is bound on the opposite site of the palindromic DNA.

TGIF1–HD folds into a tight globular helical structure, with all characteristic secondary structural elements of the Homeodomain family, namely, the N-terminal arm and the 3-helix core domain (boundaries represented on top of Figure 2A). In the complex, each TGIF1–HD monomer binds through specific interactions of the $^{219}\text{ArgArgArg}^{221}$ binding site, located in the $\alpha 3$ helix, with the 5'-TGACA-3' site (Figure 2A and Table 1). This $\alpha 3$ helix is nicely accommodated into the major groove defined by the 5'-TGACA-3' motif (Figure 2B) and stabilized by a network of HBs (distances ≤ 3.7 Å), such as those between the guanidine group of the Arg220 side-chain and both G11 and T12 bases, and also from the guanidine group of Arg221, to T12, A5' and G4' (located in the complementary DNA chain). The global interaction with the DNA is further stabilized by HBs from the guanidine group of Arg219 to the phosphate group of C9. In addition, other residues located at various secondary structure elements also contribute to additional HBs to the DNA backbone (Figure 2B,C). The protein–DNA interactions are further stabilized at the complex interface by a set of well-ordered water molecules (W) that provide 15 additional HBs between DNA phosphate groups and the protein. For instance, W13 stabilizes the contacts of Tyr191, and Pro192 with G8, and W18, and those of C9 with the carboxyl group of Ile216. At the N-term arm, W54 contributes to the HB between G4' (base) and the guanidine group of Arg167 (Figure 2E and Supplementary Figure S2).

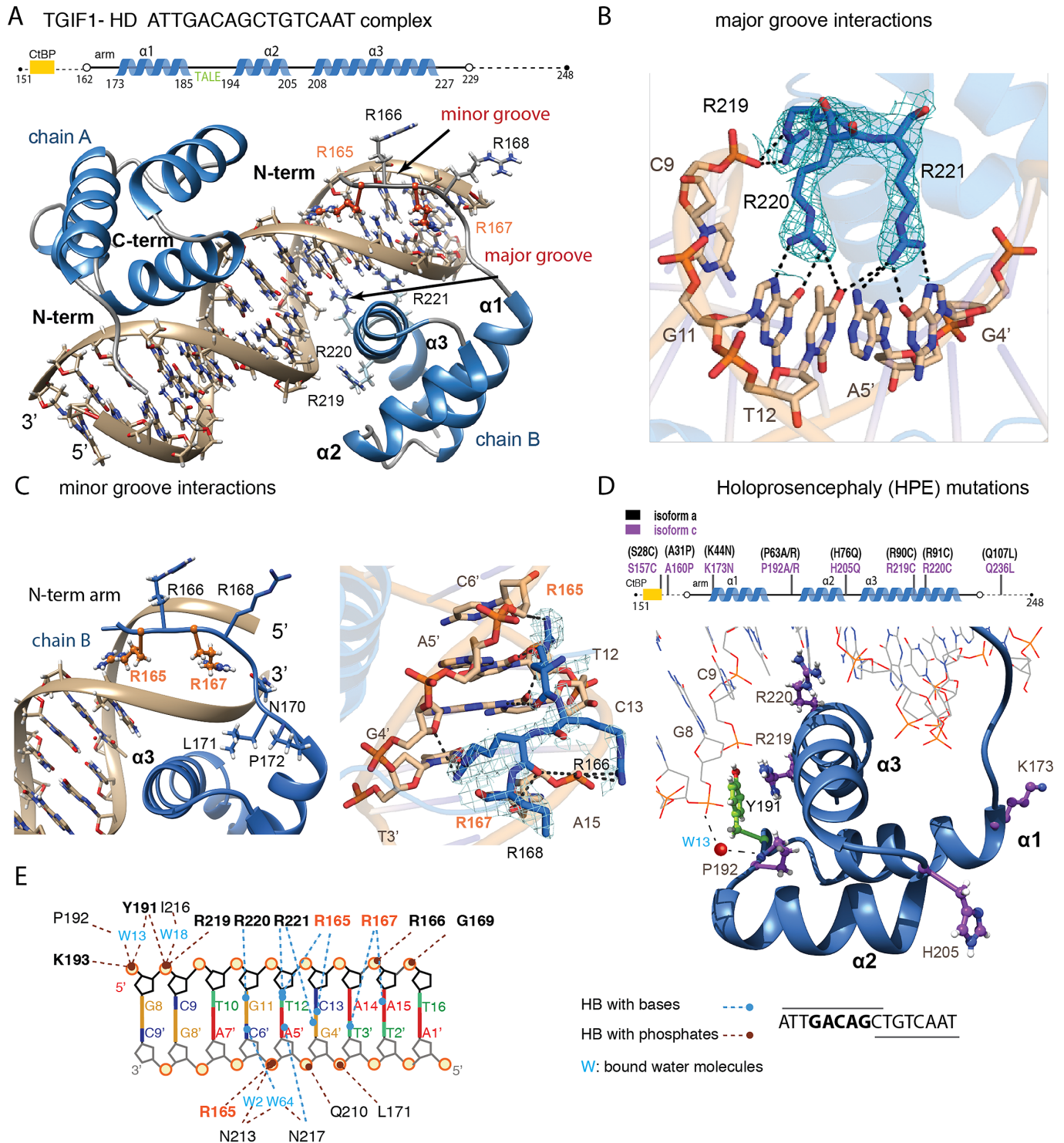


Figure 2. Complex of TGIF1-HD bound to the TGACAGCTGTCA site. **(A)** Overall structure of TGIF1-HD (blue) bound to the canonical DNA sequence 5'-ATTGACAGCTGTCAAT-3' shown as a ribbon representation (gold). The biological unit (chains A and B), the major and minor DNA grooves, and the elements of secondary structure are labeled. **(B)** Close-up view of the intermolecular interactions of the major groove. Three arginine side chains (Arg219, Arg220 and Arg221) protrude from the recognition helix3 and penetrate the major groove (hydrogen bonds indicated with dashed lines). Omit electron density map ($2F_o - F_c$) contoured at $\sigma = 1$ (in teal) is shown only for these side chains for clarity. Electron density maps for other regions are shown in Supplementary Figure S1A. **(C)** Representation of the interactions of the minor groove with Arg165 and 167 (shown as orange sticks), to illustrate how their side chains penetrate the groove. Minor groove recognition by N-terminal Arg165-Arg168 with ($2F_o - F_c$) omit maps contoured at $\sigma = 1$. Intermolecular electrostatic interactions and hydrogen bonds are indicated by dashed lines, Arg165 and Arg167 are shown as orange sticks, whereas Arg166 and 168, which do not participate in interactions, are labeled in black. **(D)** Schematic representation of the human holoprosencephaly mutations located at the TGIF1-HD. The side chains corresponding to five mutations are depicted as violet sticks in the structure of the WT HD and are labeled. Models calculated for each mutation are shown in Supplementary Figure S3. **(E)** Schematic representation of TGIF1-DNA interactions. Indicated hydrogen bonds correspond to contacts with bases (blue dashed lines); with phosphates (red); or mediated by water molecules (cyan). Interactions are shown for one monomer (PDB ID: 6FQP). A more detailed representation of these contacts is shown as Supplementary Figure S2).

Table 1. Data collection and refinement statistics for the two X-ray structures

PDB code	TGIF1- HD 6FQP	TGIF1- HD R167A/R168A 6FQQ
Data collection		
Wavelength (Å)	0.972422	1.07227
Space group	<i>P</i> 2 ₁ 2 ₁ 2 ₁	<i>P</i> 2 ₁ 2 ₁ 2 ₁
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	50.85, 67.25, 102.31	60.06, 93.02, 100.71
α , β , γ (°)	90, 90, 90	90, 90, 90
Resolution (Å)	45.54–2.42 (2.51–2.42)	45.11–3.25 (3.36–3.25)
<i>R</i> _{merge} (%)	2.75 (23.2)	8.1 (28.8)
<i>I</i> / σ <i>I</i>	17.3 (3.5)	4.4 (1.7)
Completeness (%)	98.9 (98.7)	99.01 (97.12)
Multiplicity	2.0	6.7
Refinement		
No. reflections overall	13749	9236
<i>R</i> _{work} / <i>R</i> _{free} (%)	20.5/24.1	21.5/27.3
Number of atoms		
Protein	1083	1979
DNA	656	1312
water	68	32
<i>B</i> -factors (Å ²)		
Protein	47.52	58.66
DNA	47.31	61.19
Water	35.23	39.47
R.M.S. deviations		
Bond length (Å)	0.013	0.012
Bond angles (°)	1.38	1.30
Ramachandran plot		
Favored (%)	92.25	93.67
Allowed (%)	7.75	5.48
Outliers (%)	0.0	0.84

*Values in parentheses are for highest resolution shell.

Holoprosencephaly (HPE) and tumor mutations detected in TGIF1–HD

Mutations or deletions of the human TGIF1 gene have been associated with HPE, the most common structural malformation of the forebrain and face in humans (24). Six of these mutations, Lys173Asn, Pro192Ala/Arg, Arg219Cys, Arg220Cys and His205Gln, are located in the HD (residues 44, 63, 90, 91 and 76 in the most common isoform of TGIF1). The first three residues are highly conserved in the TALE HD family and, with the exception of His 205, the remaining mutations are directly located at the DNA-binding interface (Figure 2B and C). Pro129 to Arg is also mutated in rectal and colorectal adenocarcinomas, as well as in cutaneous melanomas (COSMIC database, (25)).

Previous experiments performed in cell lines (26), as well as in *in vitro* assays (27), revealed that only the mutated residues located at positions 192 and 219 failed to repress TGF β -activated transcription and impair DNA binding. The structure of the WT HD-DNA complex can provide structural grounds for describing possible mechanisms for the loss of function or for other potential roles of these mutated proteins related to the fold of TGIF1–HD. To il-

lustrate these potential effects, we generated models carrying each of mutation applying MDS using openMM and Amber14 force field (28). Our models indicate that the Lys173Asn change might not have a high impact on the structure: however, this modification can affect the interactions of the HD with Smad proteins (Figure 6A and B). With respect to His205, which is located in the structure far from the DNA-binding sites, the mutation to Gln affects the orientation of the N-term arm. This new orientation is favored by interactions of the arm with the Gln side chain, thus reducing the capacity of the N-term arm to access the minor groove (Supplementary Figure S3).

The Arg219, Arg220 and Pro192 residues belong to the DNA binding region. In the WT structure, Arg219 and Arg220 contact the DNA directly, whereas Pro192 does so through a water-mediated HB (Figure 2D). The mutation of either Arg219 or Arg220 to Cys will not permit the formation of HBs to the DNA, thus drastically reducing the DNA binding capacity of the mutant, in agreement with the results described in the literature for these point mutations, both *in vitro* and in cell line experiments (26,27).

In the case of Pro192Ala/Arg mutations, the effect observed *in vitro* and in cell lines of losing DNA binding capacities could be explained as a change of conformations of

the loop connecting the $\alpha 1$ and $\alpha 2$ helices induced by the substitution of the Pro residue. The loop configuration observed in the WT structure is possible thanks to the rigid characteristics of the Pro amino acid, which adopts a *trans* configuration with respect to Tyr191, whose side chain also contacts the DNA (shown in chartreuse in Figure 2D). In the models, the substitution of this key Pro to either Ala or to Arg cannot restrict the structure of the turn, thus adding additional flexibility to the system and decreasing the optimal interaction with DNA, Supplementary Figure S3.

Role of the N-term arm

There is abundant information in the literature regarding the contribution of N-term arms to DNA specificity in HD complexes (29,30). Although the main features in these structures are conserved, the Arg residues in the N-term arm involved in the interactions vary from one complex to another.

In the TGIF1–HD structure, we observed direct HB interactions between residues located at the N-term arm preceding the HD core (Arg165 and Arg167, chain B), and in loop 1 (Gly169 and Leu171) with the DNA minor groove. These 17 HBs expand the recognition area from the core 5'-TGACA-3' motif, to cover a total of eight base pairs 5'-TTGACACG-3' (Figure 2C and D). Of the four consecutive Arg residues present in the N-term arm (Arg165–168), Arg167 participates in HBs with T3', G4' and A15 nucleotides (nitrogenous bases and sugars), whereas Arg166 and Arg168 are partially defined, thereby indicating rotamer/backbone flexibility (Figure 2C). The flexibility is marked in chain A, where the electron density allowed us to trace only one Arg side-chain in the minor groove (Supplementary Figure S1).

To further characterize the role of the Arg residues present in the N-term arm, we prepared a mutant protein (R167A/R168A) and compared its binding properties with respect to the WT. We observed a 2-fold reduction in the dissociation constant of the double mutant with respect to the WT (K_D 388.6 \pm 65 nM). We also determined the X-ray crystal structure of the R167A/R168A mutant bound to DNA, at 3.25 Å resolution (PDB ID: 6FQQ, Figure 3A), using the WT structure as the search model for molecular replacement. The ASU of the complex (space group: $P2_12_12_1$) contains two monomers bound to the dsDNA. As expected, the main differences between the WT and mutated complexes are observed at the minor groove, where the mutations are located (Figure 3B). The best-defined model (chain D) the electron density starts at residue 167, revealing that the introduction of the Arg167/168 mutations to Alanine enhances the flexibility of the arm, thus preventing us from tracing all residues preceding 167, including Arg165 (Figure 3C, electron density map Supplementary Figure S4). None of the Ala side chains introduced formed direct HBs with DNA, but the carboxyl group of Ala168 interacted with G4' *via* a water molecule (W32), which resembles the position of Arg167 in the WT structure.

To analyze the similarities and differences of the TGIF1 N-term arm with respect to other HD complexes, we compared the overall DNA topology of TGIF1 WT and mutated complexes to that of other HD structures with ordered

N-term arms. We observed that the HD of TGIF1 is very similar to that of Engrailed (PDB: 1HDD), but different from Meis1 (PDB ID: 4XRM) (Figure 3D,E). In fact, in the three complexes, the length of the N-term arm differs, and the turn preceding the $\alpha 1$ helix adopts slightly different orientations (Figure 3E). The TGIF1 arm is the longest of the three complexes, which is stabilized by the interactions between two Arg side chains (165 and 167) and the DNA. In the case of Engrailed, only the Arg5 side chain penetrates the interior of the groove, to occupy the equivalent position of Arg167 in TGIF1, whereas in Meis1, the crystallized construct is too short at the N-term, resulting in the arm being poorly accommodated in the groove (Figure 3E). Concerning the turn, HD sequences contain a conserved aromatic/aliphatic residue (Phe in Engrailed and Meis1, and Leu171 in TGIF1), which contacts the helical part of the HD, pre-orienting the arm towards the DNA minor groove (Figure 3E). The presence of Pro172 in the TGIF1–HD sequence further stabilizes the turn and facilitates the orientation of Arg167 and Arg165 side chains to the interior of the groove.

We consider that the different DNA shapes observed in the minor grooves of the three complexes (analyzed with Curves, (31)) are the result of the small number of protein-DNA interactions observed in Engrailed, TGIF1 mutant and Meis1 complexes compared to the TGIF1 WT counterpart (Figure 3D). Consequently, the minor groove of the Meis1 complex is deeper and narrower than that of TGIF1 or Engrailed, thereby indicating that both proteins and DNAs adapt their conformations to maximize shape complementarity at the binding interface.

The TGIF1 C-term domain is intrinsically disordered in solution

The C-term domain of TGIF1 holds the Smad-binding domain, which includes several phosphorylatable residues (Figure 4A). Three of these residues, Ser286, 291 and 294, are phosphorylated *in vivo*, although the specific roles of these phosphorylations are not fully understood (32–35). These three phosphorylation sites are conserved in mammals, Ser219 being present in all vertebrates (Figure 4B).

Secondary structure predictions performed on the C-term fragment did not reveal the presence of secondary structure elements (Figure 4C, left) (36). To evaluate the structural properties in solution, we prepared a ^{15}N – ^{13}C -labeled C-term sample and acquired triple resonance NMR experiments to facilitate the assignment of the backbone resonances. The absence of elements of secondary structure was confirmed by $\text{C}\alpha$ and $\text{C}\beta$ chemical shift differences, whose negative values (–0.5 ppm) indicated that most residues adopted a highly extended conformational ensemble in solution (Figure 4C, right).

We also examined whether the phosphorylation of the three Ser residues affects the secondary structure of the domain. For this purpose, and due to the size of the domain (~90 residues), we phosphorylated the recombinant samples using appropriate kinases *in vitro*, as an alternative to peptide synthesis. To ensure the preparation of homogeneously phosphorylated samples, we followed the reaction progress by NMR and by mass spectrometry (MS).

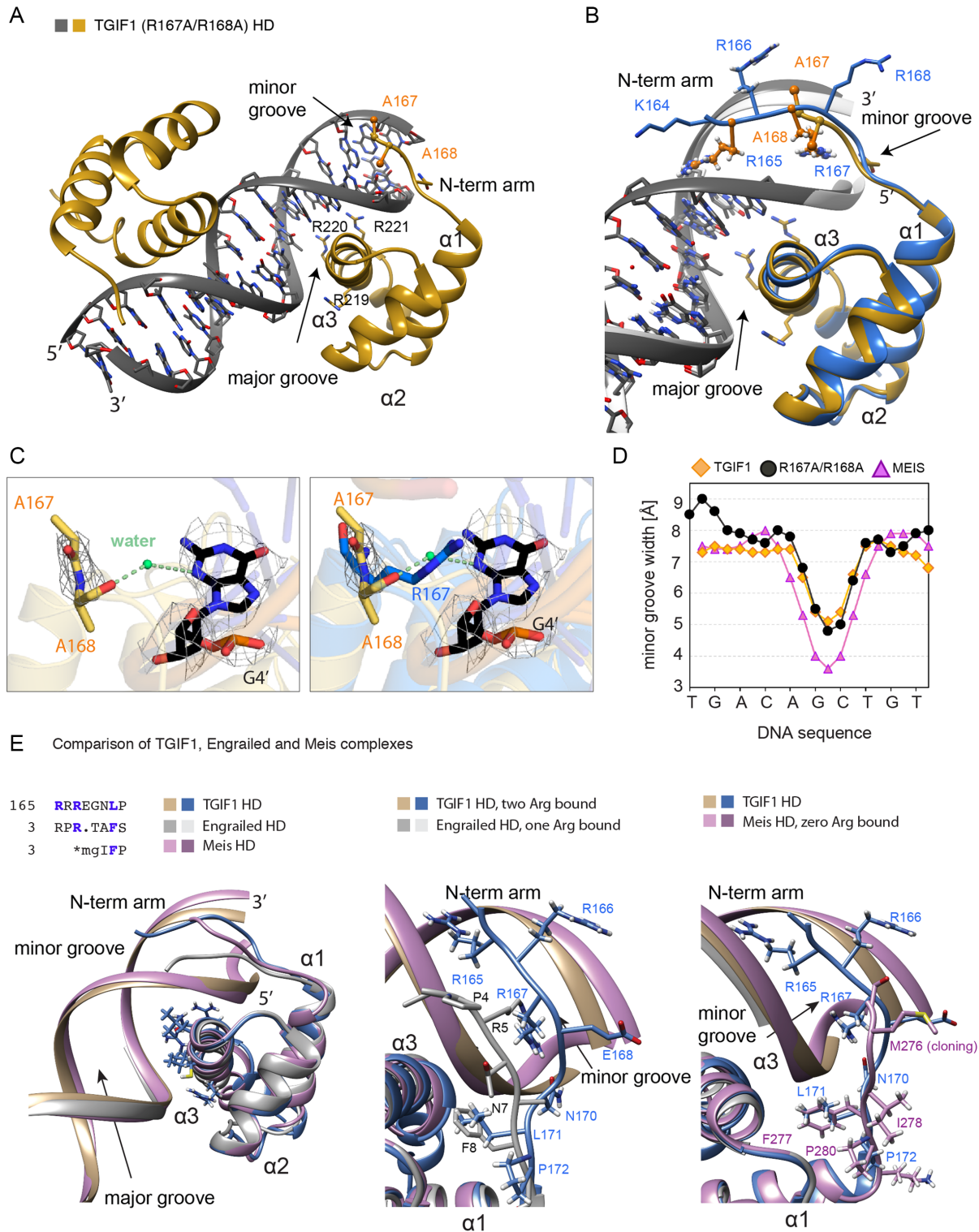


Figure 3. Structure of the R167A/R168A TGIF1–HD complex. (A) Structure of TGIF1–HD mutant (R167A/R168A) (gold) in complex with 5'-ATTGACAGCTGTCAAT-3 (light gray) shown as a ribbon representation. Elements of secondary structure, the Arg involved in the major groove recognition and the mutated Ala residues are shown as orange sticks and are labeled. (B) Superimposition of the WT complex (protein blue, DNA in light gray) and double-mutant structures colored as above. The two structures show great similarity (backbone heavy atoms comparison of 0.25 Å), differences concentrate at the N-term part of the protein (less well defined in the mutant) and at the DNA minor groove. (C) Close-up view of the interactions corresponding to the Ala167 and Ala168 mutations. Ala168–G4' interaction is stabilized by a water molecule, which resembles the position of Arg167 in the WT complex (shown below). (D) Minor groove width analyses of the DNA bound to TGIF1–HD WT (orange), R167A/R168A (black) and Meis1 (magenta). The widths were calculated using Curves+. (E) Superposition of the TGIF1–HD structure on that of Engrailed and Meis complexes. N-term arms and the different contacts observed with the minor groove are highlighted for each complex.

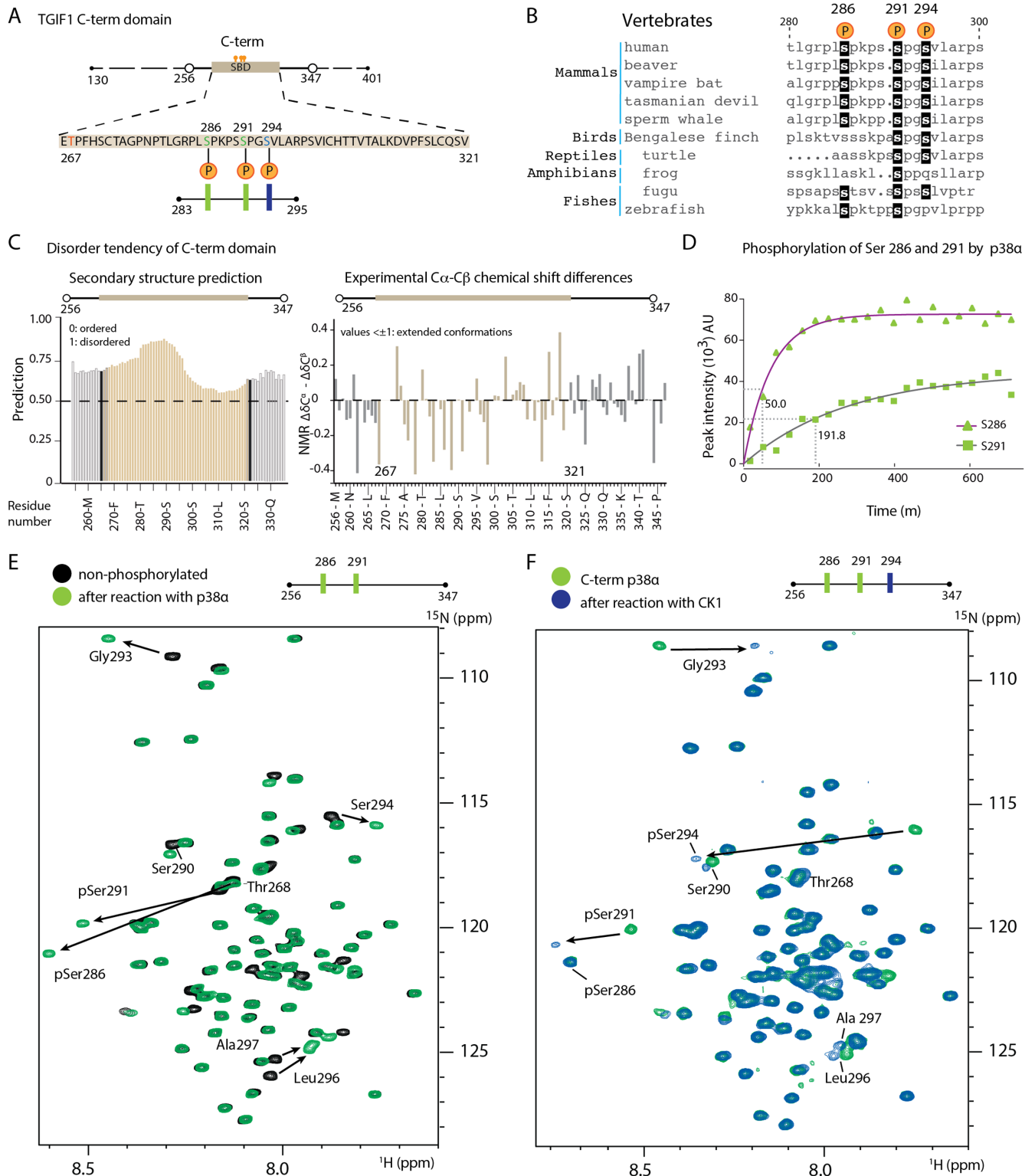


Figure 4. TGIF1 C-term domain is unstructured in solution. (A) Schematic representation of the C-term constructs used in the study. The Smad binding sequence is shown in detail. Phosphorylatable residues are highlighted in green (p38α sites) and blue (CK1). Below, a schematic peptide sequence with the three phosphorylated Ser residues, as prepared by solid-phase peptide synthesis. (B) Sequence alignment of the C-term central region of TGIF1 for various vertebrate species, displaying the degree of conservation of the three phosphorylatable Ser residues. (C) Disorder propensity of the C-term sequence characterized by secondary structure prediction (left) and by NMR chemical shift differences (right). (D) Time-resolved phosphorylation curve of TGIF1 C-term indicating the sequential phosphorylation of Ser286 and Ser291, respectively, by p38α. (E) HSQC experiments displaying the dispersion of the amide chemical shifts of ¹⁵N-labeled TGIF1 C-term domain (shown in black) or after phosphorylation by p38α (in green). (F) HSQC experiments displaying the phosphorylation of Ser294 by CK1 (shown in blue).

The reaction was performed sequentially, with the addition of p38 α first, which yielded a bi-phosphorylated ^{15}N - ^{13}C -labeled domain, with an increase of 156 Da in the molecular weight, corresponding to the addition of two phosphate groups. We followed the chemical shift changes of the corresponding phosphorylated Ser residues and determined that Ser286 phosphorylated four times faster than Ser291 (Figure 4D). Afterwards, Ser294 was phosphorylated using CK1 (Figure 4E and F) and confirmed by MS as an additional increment of 78 Da. The HSQC experiments and the $^{13}\text{C}\alpha$ - $^{13}\text{C}\beta$ analysis indicated that phosphorylation affected the chemical shifts mostly of the residues involved in the interaction, without introducing large structural rearrangements. The lack of a defined structure smooths the way for the access of the specific kinases for phosphorylation and it would represent an advantage to perform the domain function. Furthermore, the intrinsic flexibility (with and without the presence of phosphorylated residues) could facilitate the interactions with other proteins such as Smad2, (13), without involving large conformational changes.

Closed conformation of TGIF1

The presence of disordered regions in proteins is often correlated with transient interactions that facilitate conformational exchange. We first examined whether the C-term domain participates in interactions with the HD by facilitating the presence of closed conformations (inter/intramolecular), which might prevent non-specific access of the HD to DNA.

To this end, we performed NMR titration experiments, adding increasing amounts of the HD fragment to the ^{15}N -labeled C-term. Comparison of the HSQCs revealed chemical shift differences and a decrease in the intensity of several peaks in the C-term (Figure 5A). Several of the affected residues are located at the center of the C-term, in the proximity of the phosphorylatable Ser residues. We also performed a complementary titration using a ^{15}N -labeled HD and a pTGIF1 C-term peptide with the three phosphorylated sites (Figure 5B). The residues displaying chemical shift changes are indicated on the structure of the HD (Figure 5C). The phosphorylated peptide binds to the HD with a K_D of $4 \pm 1 \mu\text{M}$, as determined by ITC (Figure 5D), whereas the chemical shift differences (CSD) introduced by the non-phosphorylated recombinant C-term, in comparison to those introduced by the peptide, indicate an affinity several orders or magnitude weaker, thus preventing the quantification of this interaction using ITC (Figure 5D and F).

These results suggested that, in the context of the full-length protein, TGIF1 can populate an equilibrium of closed/open conformations, regulated by interactions of the HD and the C-term. The open/closed ratio can be switched by C-term phosphorylation and/or by the presence or absence of other protein/DNA partners that bind to the domains (13).

TGIF1 binding to Smad domains

TGIF1 repression of Smad-dependent TGF- β signaling has been proposed to involve several regions of TGIF1 and

Smad2 proteins (13). Two of these regions of TGIF1 are located in the C-term domain, whereas the third region partially overlaps with the HD. The presence of several TGIF1 sites led us to hypothesize that the global interaction between Smad and TGIF1 proteins occurs through a synergic mechanism involving several Smad/TGIF1 binding sites that participate in the interaction of the full-length proteins in an orchestrated manner.

To test this hypothesis, we selected the ^{15}N -labeled C-term domain and the linker-MH2 (EEE mutant) of Smad2 as the titrant and followed the interaction by NMR. The mutant mimics the phosphorylation state of activated Smad2 and forms homotrimers in solution (6). Upon the addition of increasing amounts of Smad2 linker-MH2 domain in a step-wise manner (up to ~ 6 eq. excess), only small chemical shifts and very weak differences in the intensities of some residues were observed at the C-term domain (data not shown), suggesting that the interactions between the two proteins were weak, with a K_D value in the millimolar range. When the titration was performed using the phosphorylated C-term variant, slightly larger chemical shift changes were detected in the region surrounding the both pSer286 and pSer291 residues (Figure 5E and F). These interactions, however, did not introduce many CSD as we would have expected for two systems that interact with K_D values in the micromolar range (Figure 5F). Other attempts to measure these interactions, including microscale thermophoresis, did not provide clear results due to the tendency due to the tendency of the MH2 domain to aggregate with greater concentration.

Overall, these experiments indicate that the interaction of the isolated C-term with the Smad2 linker-MH2 domain is weak and that phosphorylation slightly improves its apparent affinity. Since the TGIF1 protein exerts its repressor function through several regions of the protein, including the C-term addressed here, and also in a HDAC-dependent/independent manner (13), additional interactions of Smad2 and HDAC proteins and cofactors, or in the presence of the heterotrimeric complex with Smad4, may enhance the overall affinity of these complexes with TGIF1.

Functional implications of homeodomain-MH1 domain interactions

A possibility to enhance the interaction between Smad proteins and TGIF1 might involve other domains present in Smad proteins and in TGIF1, in addition to the interactions between the MH2 domains and the TGIF1 C-term. In our search for other potential interactions with Smad proteins, we examined whether TGIF1-HD binds to Smad2/Smad4 MH1 domains in a similar manner to that proposed for the HDs of HOXC9 and *distal-less*-like DLX1 proteins (37,38). Using NMR-based titration experiments, we followed the amide CSD of TGIF1-HD upon the addition of increasing amounts of either Smad4 or Smad2 MH1 domains. The presence of the MH1 domains induced CSDs of TGIF1-HD residues located mainly at the $\alpha 1$ and $\alpha 3$ helices, and clustered on one side of the structure (Figure 6A and B). The affected residues are highly conserved in HD domains, thereby suggesting that other HD-containing proteins, also

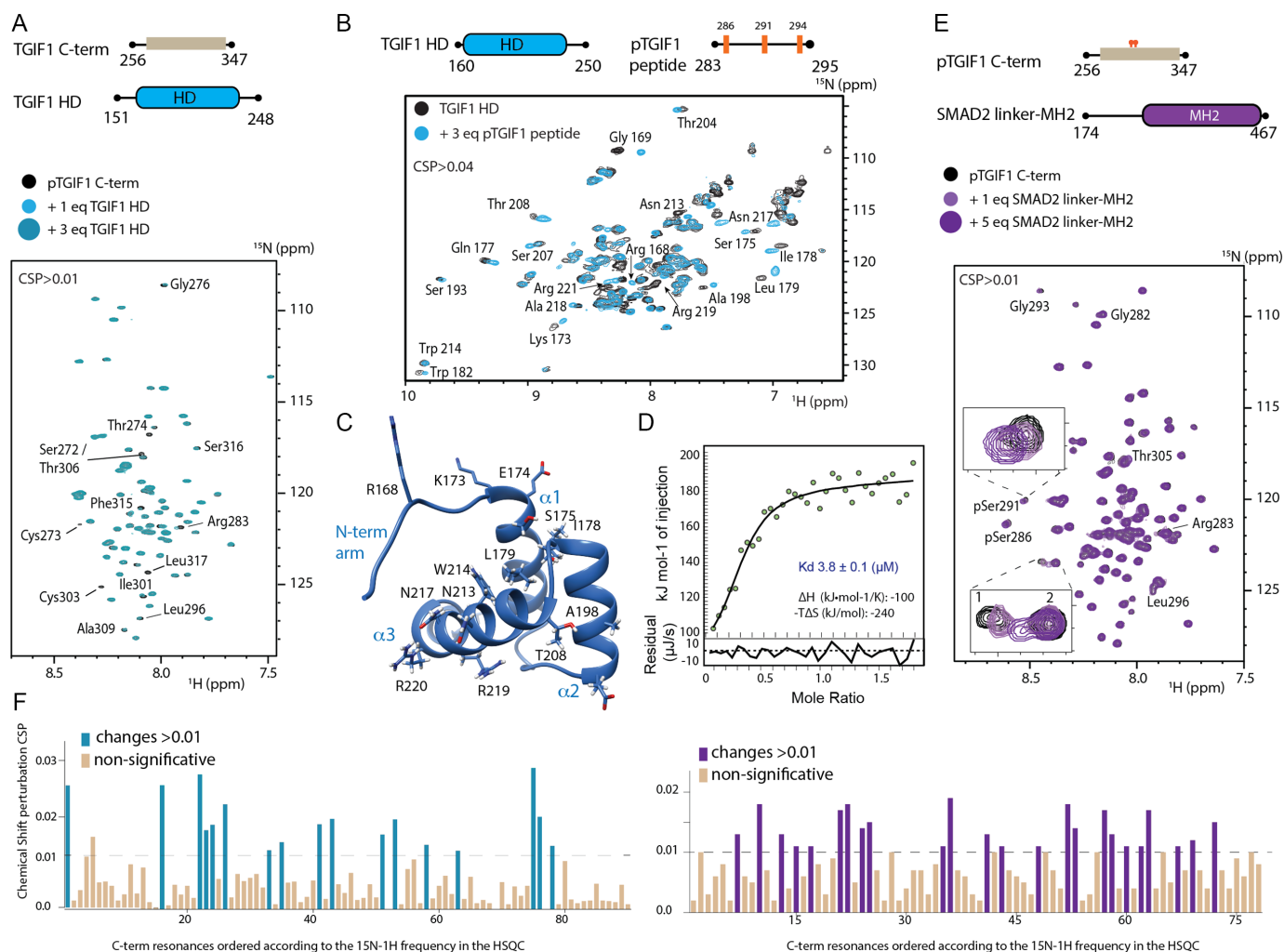


Figure 5. TGIF1 domains are also protein-protein binding sites. (A) HSQC titration of the ^{15}N -labeled C-term fragment (non-phosphorylated), with increasing amounts of HD protein. Affected residues are labeled. Several of these residues are in proximity to the phosphorylatable Ser residues. (B) HSQC titration of the ^{15}N -labeled HD using the phosphorylated peptide as titrant. (C) HD binding site for phosphorylated C-term peptide. Residues affected upon addition of increasing amounts of the peptide are plotted on the structure of the domain. (D) Isothermal calorimetry titration of the interaction between TGIF1-HD and the phosphorylated peptide. Experiments were performed at 20°C in Tris/HCl buffer, pH 7.4. Data were fitted using the independent model assuming a single binding site. (E) HSQC titration of the ^{15}N -labeled C-term fragment (bis-phosphorylated), with increasing amounts of Smad2 linker-MH2 construct. Affected residues are indicated. Peaks labeled as 1 and 2 correspond to two unassigned resonances. (F) Chemical shift perturbations detected for the C-term domain in the presence of the HD (left) or in the presence of the MH2 domain of Smad2.

known to interact with Smads, might bind through a common protein-protein HD-binding site (Figure 6C and D).

As both the DNA- and MH1-binding site involves the $\alpha 3$ helix, we set about studying how Smad-TGIF1 interactions affect the DNA-binding capacity of the proteins. For this purpose, we incubated the HD-DNA complex with increasing concentrations of either Smad4 or with Smad2-MH1 domains, using two ratios of TGIF1 protein/DNA complexes (54 and 212 nM protein and 3.7 nM DNA). In the EMSA, the shifted bands corresponding to the TGIF1-DNA (54 nM) complex almost disappeared (the presence of free DNA is observed) after the addition of 1 equiv. of Smad4-MH1 domain (Figure 6E). Similar results were obtained with the Smad2-MH1 domain. In this case, the increase of the amount of unbound DNA is observed at 0.5 equiv. (30 nM) of Smad2 MH1 domain. As Smad4/2-MH1 domains do not interact with the DNA motif of TGIF1,

these results indicate that the presence of these domains interfere with the formation of the TGIF1-DNA complex.

We also studied whether the interaction of TGIF1-HD with Smad proteins prevented them from binding to DNA. The competition mechanism through direct binding to DNA has been reported for the repression of TGIF1 by Meis2a (myeloid ecotropic viral integration site 2), which shares with TGIF1 an overlapping and complementary common binding site in the activator ACT sequence of the D_{1A} promoter (39). However, Smad DNA motifs (SBE and 5GC motifs) (40) differ in sequence from those recognized by TGIF1-HD and, in fact, the interaction of HD with Smad DNA occurred with dissociation constants 100 times weaker than those of its canonical binding site (Figure 6F). Using a complementary titration in which the Smad4 MH1 domain bound to a 5GC motif was titrated with increasing amounts of TGIF1-HD, we observed a decrease in the in-

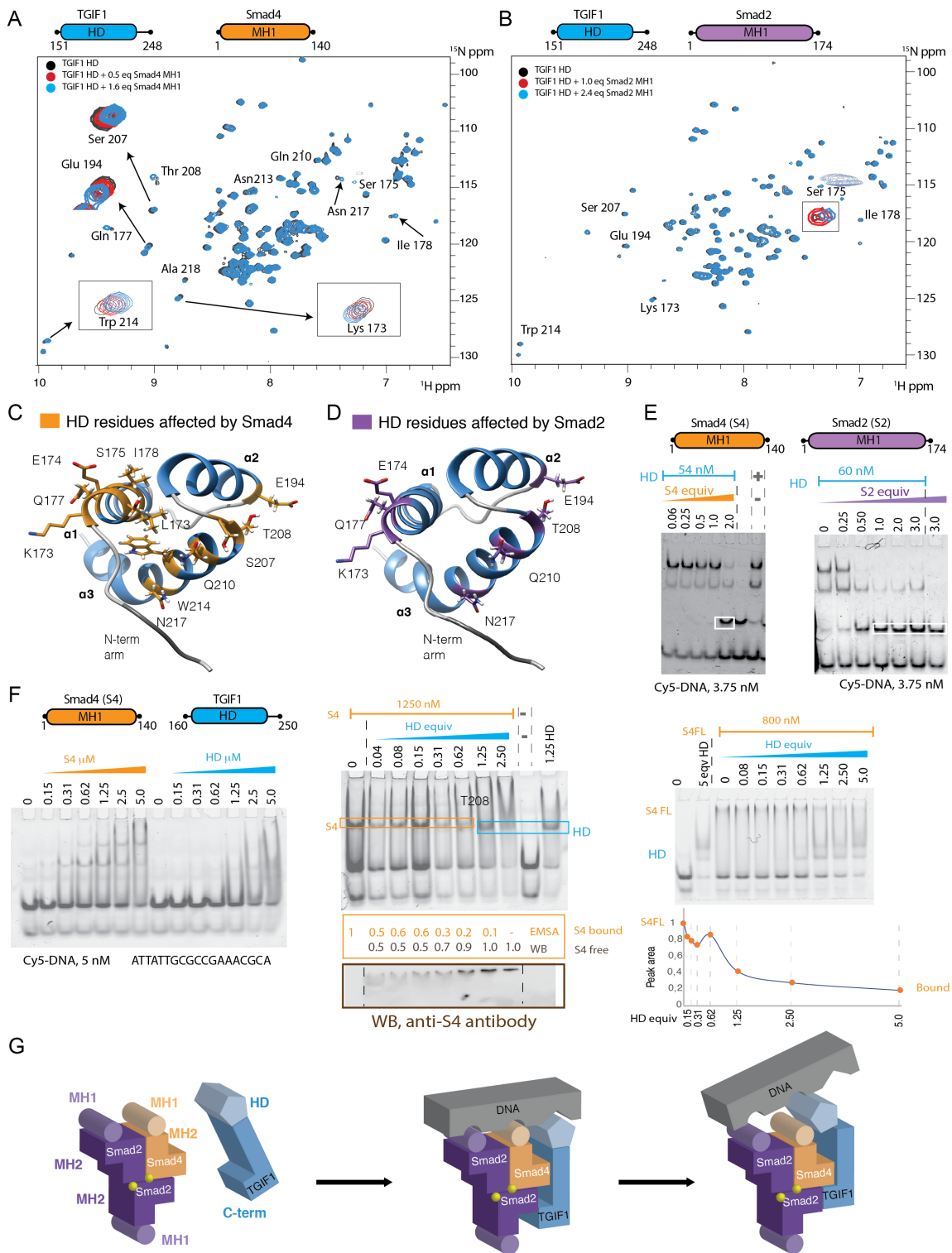


Figure 6. TGIF1–HD interacts with the MH1 domains of Smad2/4 proteins. (A) HSQC-based NMR titration of TGIF1–HD with increasing concentrations of Smad4 MH1 domain. Residues affected by chemical shift perturbations are indicated. (B) As in (A) but using Smad2 MH1 domain as the titrant. HD residues affected by chemical shift perturbations are indicated. (C) Residues identified in (A) displayed in the HD structure (monomer B, PDB entry 6FQP). HD affected residues by Smad4 are highlighted in orange sticks. (D) HD residues affected by Smad2 (B) are shown in violet. (E) Control of TGIF1–HD in the absence of Smad proteins. Competition assays using two concentrations of TGIF1–HD bound to its canonical DNA (cy5 5'-ATTGACAGCTGCAAT-3') and increasing amounts of Smad4 MH1 domain (left) and Smad2 MH1 (right) domains. (F) Control of the Smad4 MH1 and TGIF1–HD binding to the GGCGG site. TGIF1–HD binds to this site at high protein concentrations, probably due to nonspecific binding. Competition assay of Smad4 MH1 domain interaction with the GGCGG site in the presence of increasing amounts of TGIF1–HD. The bound Smad4 protein disappears from the DNA complex, as revealed by the western blot using anti-Smad4 antibody. A similar competition is performed using Smad4 FL protein. (G) Cartoon representation of TGIF1 interaction with Smad2/4 heterotrimeric complexes and the DNA. The formation of the TGIF1–Smad complex brings the Smad MH1 domain and the TGIF1 HD into close proximity, which perturbs the interaction of the protein complexes with DNA.

tensity corresponding to the Smad4 MH1-DNA complex, the recovery of unbound Smad4 MH1 domain (detected by western blot using an anti-Smad4 antibody and qualitative quantified), and the presence of the excess of TGIF1-HD nonspecifically bound to the Smad binding site (Figure 6F). We also repeated this interaction using Smad4 full-length (FL) protein and a DNA fragment containing several interaction sites for MH1 domains. In this case, since the Smad4 FL protein and TGIF1 HD complexes run very differently, we quantified the amount of free and bound Smad4 protein during the competition assay.

Taken together, these results indicate that TGIF1 repression of Smad-dependent TGF- β signaling involves several domains, namely the MH1 and MH2 of Smads and the HD and C-term of TGIF1 respectively, and influences the DNA binding capacity of the protein complexes. A possible mechanism for the repression is shown as a cartoon representation in Figure 6G. In this cartoon, we propose that several contacts between TGIF1 and Smad proteins in the presence of DNA facilitate the removal of the Smad-TGIF1 protein complex from the DNA, thereby revealing a possible mechanism by which TGIF1 represses Smad signaling.

DISCUSSION

Transcription factors (TFs) recognize DNA through interdependent effects that include the recognition of the DNA geometry and the nucleotide sequence (direct/indirect base readout). The final predominant mechanism is TF-dependent, since the presence of additional cofactors (specific to each TF) modulates the interactions with the promoters and enhancers, thus determining the transcription or repression of genes (41–43). The HD family of TF is well conserved and present in all eukaryotes. These proteins exert a wide range of functions in transcriptional activation and repression (44).

TGIF1 is a transcription repressor of the Smad-driven TGF- β signaling network. It is a ubiquitously expressed HD-containing protein, initially identified as a binder of the retinoid X receptor (RXR) response element (14,45). Our results reveal how the interactions of TGIF1-HD occur with DNA and with Smad proteins. The binding to DNA involves specific contacts with residues located at the $\alpha 3$ helix and the major groove and also with residues in the N-term arm of the protein and in the minor groove. Our results (Figure 1D and E) are in agreement with observations reported in the literature, the contribution of N-term arms to increasing the affinity of HD interactions with DNA has been highlighted (46). However, not all HDs bind to the same canonical DNA, thereby suggesting that the differences in K_D values for any given complex are probably related to the recognition of both major and minor grooves, the latter often considered to be sequence-unspecific. K_D values have been determined for several HD sequences and their different canonical DNAs using ITC. These values, all within the nM range, correspond to sequences with N-term arms containing several consecutive Arg and Lys residues. The values vary from one complex to another, as so those measured for TGIF1 (200 nM), PBX1HD (360 nM) (47), Engrailed (800 nM), and consensus-HD (8 nM), a *de novo*

designed HD sequence that interacts with the Engrailed DNA motif (46).

The TGIF1-HD complex (repressor) differs from other HD complexes (mostly activators), both in the selected DNA targets and in the efficient interaction with the minor groove, whose topology is affected upon protein binding (Figure 3D). It is possible that the sequences surrounding the main motif (either GC rich as in Engrailed or AT as in TGIF1-HD complexes respectively) have a more remarkable role than previously thought. These surrounding base pairs would affect both the shape and malleability of the DNA structure upon protein interaction. The observation that many HD sequences have N-term arms with four or five Arg-Lys residues may indicate that the efficient accommodation of the N-term arm into the minor groove is sequence-dependent and that it might be one of the requirements that HD proteins need to satisfy to identify target sites in promoters and enhancers.

We also found that the C-term domain is unstructured in solution and that it interacts with the HD *in vitro*. The presence of intrinsically disordered regions in many transcription factors and repressors is very common, and these regions provide a large interaction interface, which can sample many conformations with a low energetic penalty and can be adjusted to recognizing distinct protein partners (48). In this case, the interactions of the HD with the flexible C-term domain might indicate that the full-length protein is present as an equilibrium of open and closed conformations, which can be regulated by phosphorylation.

Regarding the interaction with Smad proteins, TGIF1 contains several regions that act as potential repressors (13). In this regard, we examined potential interactions of these regions with Smad domains *in vitro*. Indeed, we found weak-medium interactions between the phosphorylated TGIF1 C-term and Smad2 linker-MH2 construct and also for TGIF1-HD and the MH1 domain of Smad proteins. We mapped the HD binding region for MH1 domains, which involves the $\alpha 1$ and $\alpha 3$ helices. The interaction of the HD with Smad proteins has been proposed for the distal-less-like HD-containing proteins DLX1 and for HOXC9-HD (37). Thus, our results might be valid in other Smad-cofactor interactions.

Since the $\alpha 3$ helix participates in DNA binding, we noted that HD-MH1 domain recognition occludes, at least in part, the DNA-binding site, thus reducing the affinity of the protein complex for DNA. The ability of HD and MH1 domains to interact with proteins, as well as with DNA, identified here reinforces the versatility of these domains and challenges the classical distinction of DNA- and protein-binding sites in Smad and in HD-containing proteins. Furthermore, the binding of TGIF1 and Smad proteins through their respective DNA-binding domains can enhance the overall affinity of the TGIF1 C-term domain for the MH2 domain of Smad2 and *vice versa*. This concerted mechanism, which involves several binding sites on both TGIF1 and Smad proteins, explains how TGIF1 efficiently represses Smad-dependent TGF- β signaling.

Biological relevance

Holoprosencephaly is a congenital disorder caused by the failure of the embryonic forebrain to divide correctly, generating a single-lobed brain structure and severe skull and facial defects. Of the eleven mutations found in TGIF1 related to this condition, three of them are localized in the HD and cannot interact with DNA, both *in vitro* and in cellular assays (24,25). Our HD structure reveals that three of these mutations clustered at the protein-DNA interface have a direct impact on DNA recognition. Moreover, the models we generated for each mutant suggested that His205Gln mutation affects the interaction of the N-term arm with the minor groove. Holoprosencephaly (HPE) patients also have mutations located at the C-term domain, at the interaction site with Smad proteins. In fact, some patients have Ser291 and Thr280 (Figure 4B) mutated (Ser191Phe and Thr280Ala), residues that we found to participate in interactions with Smads. However, previous experiments in cellular assays indicate that these mutations do not affect the capacity of TGIF1 to repress TGF β -activated transcription (26,27). Although these mutations might not be critical for the repression of TGF β signaling, they might affect the population of open/closed conformations of TGIF1, and also the interactions with Smads and other proteins, thereby altering the function of TGIF1 protein in these patients.

TGIF1 is an active repressor of TGF- β signaling, and it exerts its function through specific interactions with Smad proteins and other cofactors. We have shown how two distinct regions of TGIF1, namely the HD and the C-term domain, interact with the MH1 and MH2 domains of Smad proteins respectively. The binding of the HD to the MH1 domains of Smad proteins involves the DNA binding site of the HD and also affects the interaction of the MH1 domain with its canonical binding site. It therefore appears that the binding of TGIF1-HD to Smad proteins and to their canonical DNAs is mutually exclusive, thus revealing a mechanism through which the Smad-TGIF1 complex acts as a transcriptional repression system.

DATA AVAILABILITY

NMR assignments and chemical shifts of the TGIF1 C-term domain have been deposited in the Biological Magnetic Resonance Data Bank, BMRB entry 27461. Densities and coordinates have been deposited in the Protein Data Bank, PDB entries 6FQP and 6FQQ.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr P. Selenko and Dr F.-X. Theillet for insightful suggestions regarding the phosphorylation reactions of the TGIF1 C-term, and Dr P. Schmieder and Dr H. Oschkinat for access and support with triple resonance experiments acquired at the 750 MHz (Leibniz Institute of Molecular Pharmacology, FMP Berlin). Thanks also go to the Automated Crystallography Platform (CSIC-IRB Barcelona) for support with setting-up crystallization trials, to Dr M.

Navia for help with the preparation of the manuscript, to Dr M.J. Klein for help with the triple phosphorylated peptide synthesis, to N. de Martin for support with the EMSA experiments and to the ESRF-staff (Grenoble) for help using the synchrotron beamlines ID29, ID23-1 and ID23-2.

Author contributions: D.S., L.R., E.G., E.A. and A.K. cloned, expressed and purified all proteins. E.G., L.R., D.S., C.T. and A.K. performed the EMSA and ITC experiments. D.S. and A.R. performed the phosphorylation experiments and synthesized control peptides. E.G. screened and optimized crystallization conditions, collected X-Ray data, and analyzed the X-ray structures with M.J.M., P.M. and M.J.M. acquired and processed NMR data and assigned and analyzed the data with D.S., J.C. and A.K. M.J.M. designed and supervised the project. All authors contributed ideas to the project. E.G. and M.J.M. wrote the manuscript.

FUNDING

Spanish MINECO [BFU2014-53787-P and BFU2017-82675-P to M.J.M.]; BioMedTec Project of the Obra Social 'La Caixa', the BBVA Foundation; Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the Centres of Excellence Severo Ochoa award; CERCA Programme of the Catalan Government. Access to the ESRF synchrotron was granted as part of a BAG proposal [MX-1941]. E.G. is a postdoctoral fellow co-funded by the Marie Skłodowska-Curie COFUND actions (IRB Barcelona Interdisciplinary Postdoc Programme [IRBPostPro2.0 600404]; D.S. is a PhD fellow of La Caixa Fellowship Programme; J.C. is a fellow of the Future in Biomedicine Programme (IRB Barcelona); A.K. belongs to the Erasmus internship programme; M.J.M. is an ICREA Programme Investigator. Funding for open access charge: Spanish MINECO [BFU2014-53787-P/BFU2017-82675-P].

Conflict of interest statement. None declared.

REFERENCES

1. Massagué, J. (1998) TGF-beta signaling transduction. *Annu. Rev. Biochem.*, **67**, 753–791.
2. Massagué, J. (2012) TGF β signalling in context. *Nat. Rev. Mol. Cell Biol.*, **13**, 616–630.
3. Massagué, J. and Obenauf, A.C. (2016) Metastatic colonization by circulating tumour cells. *Nature*, **529**, 298–306.
4. Massagué, J., Seoane, J. and Wotton, D. (2005) Smad transcription factors. *Genes Dev.*, **19**, 2783–2810.
5. Macias, M.J., Martin-Malpartida, P. and Massagué, J. (2015) Structural determinants of Smad function in TGF- β signaling. *Trends Biochem. Sci.*, **40**, 296–308.
6. Shi, Y. and Massagué, J. (2003) Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell*, **113**, 685–700.
7. Gao, S., Alarcón, C., Sapkota, G., Rahman, S., Chen, P.-Y., Goerner, N., Macias, M.J., Erdjument-Bromage, H., Tempst, P. and Massagué, J. (2009) Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signaling. *Mol. Cell*, **36**, 457–468.
8. Alarcón, C., Zaromytidou, A.-I., Xi, Q., Gao, S., Yu, J., Fujisawa, S., Barlas, A., Miller, A.N., Manova-Todorova, K., Macias, M.J. *et al.* (2009) Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. *Cell*, **139**, 757–769.
9. Aragón, E., Goerner, N., Zaromytidou, A.-I., Xi, Q., Escobedo, A., Massagué, J. and Macias, M.J. (2011) A Smad action turnover switch operated by WW domain readers of a phosphoserine code. *Genes Dev.*, **25**, 1275–1288.

10. Morales, B., Ramirez-Espain, X., Shaw, A.Z., Martin-Malpartida, P., Yraola, F., Sánchez-Tilló, E., Farrera, C., Celada, A., Royo, M. and Macias, M.J. (2007) NMR structural studies of the ItchWW3 domain reveal that phosphorylation at T30 inhibits the interaction with PPxY-containing ligands. *Structure*, **15**, 473–483.
11. Fuentealba, L.C., Eivers, E., Ikeda, A., Hurtado, C., Kuroda, H., Pera, E.M. and De Robertis, E.M. (2007) Integrating patterning signals: Wnt/GSK3 regulates the duration of the BMP/Smad1 signal. *Cell*, **131**, 980–993.
12. Wotton, D., Lo, R.S., Lee, S. and Massagué, J. (1999) A Smad transcriptional corepressor. *Cell*, **97**, 29–39.
13. Wotton, D., Lo, R.S., Swaby, L.A. and Massagué, J. (1999) Multiple modes of repression by the Smad transcriptional corepressor TGIF. *J. Biol. Chem.*, **274**, 37105–37110.
14. Bertolino, E., Reimund, B., Wildt-Perinic, D. and Clerc, R.G. (1995) A novel homeobox protein which recognizes a TGT core and functionally interferes with a retinoid-responsive motif. *J. Biol. Chem.*, **270**, 31178–31188.
15. Gehring, W.J. (1992) The homeobox in perspective. *Trends Biochem. Sci.*, **17**, 277–280.
16. Bürglin, T.R. and Affolter, M. (2016) Homeodomain proteins: an update. *Chromosoma*, **125**, 497–521.
17. Bengoechea-Alonso, M.T. and Ericsson, J. (2010) Tumor suppressor Fbxw7 regulates TGF β signaling by targeting TGIF1 for degradation. *Oncogene*, **29**, 5322–5328.
18. Orekhov, V.Y. and Jaravine, V.A. (2011) Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. *Prog. Nuclear Magn. Reson. Spectrosc.*, **59**, 271–292.
19. Solyom, Z., Schwarten, M., Geist, L., Konrat, R., Willbold, D. and Brutscher, B. (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR*, **55**, 311–321.
20. Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. (2010) Features and development of coot. *Acta Crystallogr. D: Biol. Crystallogr.*, **66**, 486.
21. Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W. et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D: Biol. Crystallogr.*, **66**, 213–221.
22. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
23. Schrödinger LLC (2002) The PyMOL molecular graphics system. <http://www.pymol.org>.
24. Keaton, A.A., Solomon, B.D., Kauvar, E.F., El-Jaick, K.B., Gropman, A.L., Zafer, Y., Meck, J.M., Bale, S.J., Grange, D.K., Haddad, B.R. et al. (2010) TGIF mutations in human Holoprosencephaly: Correlation between genotype and phenotype. *Mol. Syndromol.*, **1**, 211–222.
25. Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L. et al. (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
26. El-Jaick, K.B., Powers, S.E., Bartholin, L., Myers, K.R., Hahn, J., Orioli, I.M., Ouspenskaia, M., Lacbawan, F., Roessler, E., Wotton, D. et al. (2007) Functional analysis of mutations in TGIF associated with holoprosencephaly. *Mol. Genet. Metab.*, **90**, 97–111.
27. Zhu, J., Li, S., Ramelot, T.A., Kennedy, M.A., Liu, M. and Yang, Y. (2018) Structural insights into the impact of two holoprosencephaly-related mutations on human TGIF1 homeodomain. *Biochem. Biophys. Res. Commun.*, **496**, 575–581.
28. Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.-P., Simmonett, A.C., Harrigan, M.P., Stern, C.D. et al. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.*, **13**, e1005659.
29. Draganescu, A., Levin, J.R. and Tullius, T.D. (1995) Homeodomain proteins: what governs their ability to recognize specific DNA sequences? *J. Mol. Biol.*, **250**, 595–608.
30. Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B. and Mann, R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530–543.
31. Blanchet, C., Pasi, M., Zakrzewska, K. and Lavery, R. (2011) CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.*, **39**, W68–W73.
32. Mertins, P., Qiao, J.W., Patel, J., Udeshi, N.D., Clauser, K.R., Mani, D.R., Burgess, M.W., Gillette, M.A., Jaffe, J.D. and Carr, S.A. (2013) Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat. Methods*, **10**, 634–637.
33. Stuart, S.A., Houel, S., Lee, T., Wang, N., Old, W.M. and Ahn, N.G. (2015) A phosphoproteomic comparison of B-RAFV600E and MKK1/2 inhibitors in melanoma cells. *Mol. Cell Proteomics*, **14**, 1599–1615.
34. Brill, L.M., Xiong, W., Lee, K.-B., Ficarro, S.B., Crain, A., Xu, Y., Terskikh, A., Snyder, E.Y. and Ding, S. (2009) Phosphoproteomic analysis of human embryonic stem cells. *Cell Stem Cell*, **5**, 204–213.
35. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V. and Skrzypek, E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–D520.
36. Kozłowski, L.P. and Bujnicki, J.M. (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*, **13**, 111.
37. Chiba, S., Takeshita, K., Imai, Y., Kumano, K., Kurokawa, M., Masuda, S., Shimizu, K., Nakamura, S., Ruddle, F.H. and Hirai, H. (2003) Homeoprotein DLX-1 interacts with Smad4 and blocks a signaling pathway from activin A in hematopoietic cells. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15577–15582.
38. Zhou, B., Chen, L., Wu, X., Wang, J., Yin, Y. and Zhu, G. (2008) MH1 domain of SMAD4 binds N-terminal residues of the homeodomain of Hoxc9. *Biochim. Biophys. Acta*, **1784**, 747–752.
39. Yang, Y., Hwang, C.K., D'Souza, U.M., Lee, S.H., Junn, E. and Mouradian, M.M. (2000) Three-amino acid extension loop homeodomain proteins Meis2 and TGIF differentially regulate transcription. *J. Biol. Chem.*, **275**, 20734–20741.
40. Martin-Malpartida, P., Batet, M., Kaczmarek, Z., Freier, R., Gomes, T., Aragón, E., Zou, Y., Wang, Q., Xi, Q., Ruiz, L. et al. (2017) Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors. *Nat. Commun.*, **8**, 2070.
41. Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R. and Rohs, R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
42. Zambelli, F., Pesole, G. and Pavesi, G. (2013) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief. Bioinformatics*, **14**, 225–237.
43. Dror, I., Rohs, R. and Mandel-Gutfreund, Y. (2016) How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *Bioessays*, **38**, 605–612.
44. Bobola, N. and Merabet, S. (2017) Homeodomain proteins in action: similar DNA binding preferences, highly variable connectivity. *Curr. Opin. Genet. Dev.*, **43**, 1–8.
45. Bürglin, T.R. (1997) Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. *Nucleic Acids Res.*, **25**, 4173–4180.
46. Tripp, K.W., Sternke, M., Majumdar, A. and Barrick, D. (2017) Creating a homeodomain with high stability and DNA binding affinity by sequence averaging. *J. Am. Chem. Soc.*, **139**, 5051–5060.
47. Zucchelli, C., Ferrari, E., Blasi, F., Musco, G. and Bruckmann, C. (2017) New insights into cooperative binding of homeodomain transcription factors PREP1 and PBX1 to DNA. *Sci. Rep.*, **7**, 40665.
48. Sammak, S. and Zinzalla, G. (2015) Targeting protein-protein interactions (PPIs) of transcription factors: challenges of intrinsically disordered proteins (IDPs) and regions (IDRs). *Prog. Biophys. Mol. Biol.*, **119**, 41–46.