Computational and Structural Biotechnology Journal 19 (2021) 632–646

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Unveiling the dimer/monomer propensities of Smad MH1-DNA complexes

Lidia Ruiz [a,1], Zuzanna Kaczmarska [b,c,1], Tiago Gomes [a,1], Eric Aragon [a], Carles Torner [a], Regina Freier [a], Blazej Baginski [a], Pau Martin-Malpartida [a], Natàlia de Martin Garrido [a], José. A. Marquez [b], Tiago N. Cordeiro [d], Radoslaw Pluta [a,*], Maria J. Macias [a,e,*]

[a] Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, Barcelona 08028, Spain
[b] EMBL Grenoble, 71 Avenue des Martyrs, CS 90181, Grenoble Cedex 9 38042, France
[c] International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, Warsaw 02-109, Poland
[d] Instituto de Tecnologia Química e Biológica António Xavier (ITQB), Universidade NOVA de Lisboa, Av. da República, 2780-157 Oeiras, Portugal
[e] ICREA, Passeig Lluís Companys 23, Barcelona 08010, Spain

ARTICLE INFO

Article history:
Received 9 October 2020
Received in revised form 29 December 2020
Accepted 31 December 2020
Available online 6 January 2021

Keywords:
Smad
Transcription factor
Protein-DNA
MH1 domain
BMP signaling
Smad5
Smad8
5GC-motifs
Smad chimeras
Domain-swapping

ABSTRACT

Smad transcription factors are the main downstream effectors of the Transforming growth factor β superfamily (TGFβ) signalling network. The DNA complexes determined here by X-ray crystallography for the Bone Morphogenetic Proteins (BMP) activated Smad5 and Smad8 proteins reveal that all MH1 domains bind [GGC(GC)|(CG)] motifs similarly, although TGFβ-activated Smad2/3 and Smad4 MH1 domains bind as monomers whereas Smad1/5/8 form helix-swapped dimers. Dimers and monomers are also present in solution, as revealed by NMR. To decipher the characteristics that defined these dimers, we designed chimeric MH1 domains and characterized them using X-ray crystallography. We found that swapping the loop1 between TGFβ- and BMP- activated MH1 domains switches the dimer/monomer propensities. When we scanned the distribution of Smad-bound motifs in ChIP-Seq peaks (Chromatin immunoprecipitation followed by high-throughput sequencing) in Smad-responsive genes, we observed specific site clustering and spacing depending on whether the peaks correspond to BMP- or TGFβ-responsive genes. We also identified significant correlations between site distribution and monomer or dimer propensities. We propose that the MH1 monomer or dimer propensity of Smads contributes to the distinct motif selection genome-wide and together with the MH2 domain association, help define the composition of R-Smad/Smad4 trimeric complexes.

1. Introduction

The gene responses activated by the transforming growth factor β (TGFβ) superfamily (a term that includes also the bone morphogenetic proteins (BMP), Nodal, Activin and other members) play essential roles in development, immunity, tissue regeneration/homeostasis, tissue fibrosis and neuroprotective functions [56–57,41,71]. These critical roles demand a high level of conservation and fidelity of the TGFβ signaling elements in healthy organisms [56]. The canonical TGFβ signal transduction mechanism is the Smad pathway, with Smad transcription factors (TFs) being responsible for the transmission of the signals from the membrane receptor into the nucleus [55]. Smad proteins contain a DNA-binding domain (Mad homology 1, MH1) a linker and a protein–protein interaction region (Mad homology 2 (MH2 domain) (Supplementary Fig. S1A) [52,74]. The MH1 and MH2 domains are highly conserved across Smad proteins and along evolution, whereas the linker has a higher sequence variability and function [53,62]. After being phosphorylated at the MH2 domains by TGFβ receptors, activated R-Smads interact with Smad4 and define the

Abbreviations: ChIP-seq, Chromatin immunoprecipitation followed by high-throughput sequencing; Smad proteins, Homologous proteins to the Caenorhabditis elegans SMA ("small" worm phenotype and MAD family ("Mothers Against Decapentaplegic" of genes in Drosophila MH1 domain, Mad homology 1; BEST, Band-Selective Excitation Short-Transient-type experiments; TROSY, Transverse relaxation optimized spectroscopy.

[*] Corresponding authors at: Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, Barcelona 08028, Spain.

E-mail addresses: radoslaw.pluta@irbbarcelona.org (R. Pluta), maria.macias@irbbarcelona.org (M.J. Macias).

[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.csbj.2020.12.044
2001-0370/

canonical hetero-trimeric functional unit. Once in the nucleus, and upon linker phosphorylation, the hetero-trimeric Smad complex is ready to define a new set of interactions with cofactors and with cis-regulatory elements containing Smad Binding Elements (SBE and 5GC sites, GGC(GC)|(CG) Supplementary Fig. S1B), interactions that go on to modulate the outcome of the signaling network [1,3,29].

R-Smad/Smad4 complexes have been observed with over-expressed and endogenous full-length proteins [38,44]. Crystal structures of MH2 domains and biophysical experiments in solution have revealed a conserved propensity of these domains to interact as homo- and hetero-trimers [16,61,74]. Full-length proteins are also believed to associate through the MH2 domain to define heterotrimeric complexes as observed in the MH2 complexes.

R-Smad proteins were considered to have different specificities regarding the recognition of DNA motifs and to respond to specific BMP- and TGFβ-activation inputs [82]. Initial hypotheses suggested that the TGFβ-activated Smads (Smad2/3) and Smad4 showed a preference for the GTCT site (known as the Smad Binding Element, SBE), whereas the BMP-activated Smads (Smad1/5/8) preferred GC-rich motifs. However, the very high sequence conservation of the MH1 domains and recent experimental evidence indicate that the separation between DNA binding preferences of R-Smads is subtler than initially thought (Supplementary Figs. S1B,C). For instance, combined TGFβ and BMP receptors influence Smad1/5-driven responses [70] and the MH1 domains of Smad3 and Smad4 proteins interact —efficiently and specifically— with GC-rich motifs grouped in the 5GC consensus [54]. This 5GC consensus is functionally relevant for TGFβ-activated Smads and for Smad4, and it overlaps with the palindromic BRE site GGCGCC, previously defined as the GC-rich target sequence of BMP-activated Smads [42]. Complexes of MH1 domains bound to different DNA motifs have revealed that MH1 domains are able to interact with specific DNA sites using a distinctive binding site [6,7,17,54,75]. Only the long isoform of Smad2 displayed additional contacts from residues in the E3 insert, exclusively present in this specific isoform [5]. Notably, while keeping the same apparent fold and DNA binding, these crystal structures showed different domain architectures. Complexes of Smad2, Smad3, and Smad4 MH1 domains with both SBE and 5GC DNAs adopt monomeric conformations [6,54,75], whereas Smad1 and Smad5 interact with the SBE site as homo-dimers, with the α1 helix being swapped between the two monomers [7,17]. The association between transcription factors (TFs) to form homo- and hetero-dimers is a common feature employed by many TF families in eukaryotes [2,36,37] and domain-swapped dimers have been detected in members of the Forkhead family of transcription factors [60]. In many TFs, the capacity of DNA-binding domains to dimerize has implications in the regulation of specific cellular responses, in the stability of the proteins, and in the optimal selection of DNA binding sites in native contexts [2,36,37]. Remarkably, in Smad proteins, the association through their MH1 domains, and its potential function, has been somehow overlooked, relegated by the interactions occurring via the MH2 domains. However, this association through MH2 domains does not fully explain all current evidence as to why only some trimeric complexes have been experimentally identified. For instance, trimeric complexes containing BMP-Smad homotrimers or Smad4/TGFβ-Smad/BMP-Smad heterotrimers have never been detected in cells [21,27,32]. It seems that there might be other restrictions favoring the composition of some complexes over others, suggesting that a second layer of selection might exist, perhaps encoded within the different dimer/monomer propensities of the MH1 domains.

In the search for new clues to clarify how BMP-activated Smad proteins interact with the GC sites (currently not fully characterized), and to decipher the characteristics that define monomers and dimers of MH1 domains, we used several biophysical techniques to study the interaction of Smad5 and Smad8 MH1 domains with these motifs. Our X-ray structures revealed the specific protein-DNA contacts and that these Smad proteins interact as dimers, in contrast to Smad3 and Smad4 that interact similarly with the same sites as monomers. Moreover, we also used Nuclear Magnetic Resonance, and other biophysical techniques to study the conformational ensemble of Smad5 MH1 domain in solution and in the gas phase. NMR relaxation experiments and IM-MS reveal the presence of dimeric conformations in Smad5 MH1 domain even in the absence of DNA, thereby indicating that dimers are also present in non-crystallographic conditions. We also found that swapping the loop1 sequence of Smad5 for that of Smad3 (or vice versa), reversed the dimer/monomer propensities of the chimeric constructs while retaining the DNA binding capacity. To correlate these structural properties with Smad function, we have scanned the distribution of Smad1/5- and Smad3-bound DNA motifs in ChIP-Seq peaks. In this analysis, we observed specific site clustering and motif spacing, depending on whether the regions were BMP- or TGFβ-responsive, suggesting a positive correlation between the monomer/dimer propensities and the motif distribution.

Based on our results, we propose that the MH1 domains' capacity to form monomers or dimers may help define the Smad components for a given R-Smad/Smad4 ternary complex, as well as the selection of binding sites in promoters and enhancers.

2. Materials and methods

2.1. Protein production and cloning

For the Smad5/8 constructs, we used the domain boundaries described in the Smad1-GTCT structures (Uniprot: P70340, Phe9-Ser132) [7,17]. The Smad5 (Uniprot: Q99717-1, Ser9-Arg143), Smad8 (O15198-1, Thr14-Pro144) and the three chimeric domains (Table 1) were cloned using an 'In Fusion Cloning strategy' [67]. Inserts were synthesized by Thermo Fisher Scientific, amplified by PCR (oligos shown in Table 1) and confirmed by DNA sequencing (GATC Biotech). Labeled and unlabeled proteins were expressed and purified following standard procedures essentially as described [54]. Proteins were verified by Mass Spectrometry. Theoretical masses for unlabeled samples are Smad5WT Mw: 15079.56, Smad8WT Mw: 16122.75, Smad5Gly Mw: 15183.62, Smad5_3 Mw: 15003.46, and Smad3_5 Mw: 14645.06. In all cases, the elution buffer was 20 mM Tris-HCl buffer (pH 7.2), 80 mM NaCl and 2 mM TCEP (buffer 1) to facilitate the comparison to other MH1 domains previously studied [54,33,5]. Aliquots were kept frozen at −80 °C. Oligonucleotides were purchased at Biomers and/or at Metabion, Germany HPLC-purified. Resulting dsDNA molecules were dissolved in the protein buffer at 2 mM concentration and annealed as described in [33].

2.2. Crystallization

Crystallization experiments were performed at the HTX facility of the EMBL Grenoble Outstation [84] and at IBMB-IRB Barcelona Crystallography Platform. All crystals were grown by sitting-drop vapor diffusion at 4 °C. The final protein concentration in the complexes was 4.2 mg/mL (buffer 1). Proteins and DNAs were mixed at 2:1 (Molar ratio), except the Smad5_3 sample (1:1 ratio).

Optimized crystallization conditions:

– Smad5 (PDB:6FZS): 100 nL drop volume, 30 μL reservoir solution: 0.1 M bis-Tris propane pH 7.5, 20% PEG 3350, 0.2 M NaF.

Table 1
Cloning and sequence information.

	Smad5 PDB: 6FZS	Smad8PDB: 6FZT
Source organism	Homo sapiens	Homo sapiens
DNA source	Synthesis	Synthesis
Forward primer	GAAGTTCTGTTTCAGGGCCCGTCTTTTACTAGTCCAGCA	GAAGTTCTGTTTCAGGGCCCGGCAGTGAAGAGACTGCTA
Reverse primer	AAACTGGTCTAGAAAGCTTCATGGAGGTAAGACTGGACT	AAACTGGTCTAGAAAGCTTCAAGGAGGCAGTACTGGAGT
Cloning vector	PopIn F	PopIn F
Expression vector	PopIn F	PopIn F
Expression host	E. Coli	E. Coli
Complete amino acid sequence of the construct produced	Mothers against decapentaplegic homolog 5 GPSFTSPAVKRLLGWKQGDEEEKWAEKAVDALVKKLKKKKGAMEELEKALSSPGQPSKCVTIPRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLDICEFPFGSKQKEVCINPYHYKRVESPVLPP DNA (5′-D(P*TP*GP*CP*AP*GP*GP*CP*GP*CP*GP*CP*CP*TP*GP*CP*A)-3′) TGCAGGCGCGCCTGCA	Mothers against decapentaplegic homolog 9 MHSTTPISSLFSFTSPAVKRLLGWKQGDEEEKWAEKAVDSLVKKLKKKKGAMDELERALSCPGQPSKCVTIPRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLE CCEFPFGSKQKEVCINPYHYRRVETPVLP DNA (5′-D(P*TP*GP*CP*AP*GP*GP*CP*GP*CP*GP*CP*CP*TP*GP*CP*A)-3′) TGCAGGCGCGCCTGCA

	Smad5_3 PDB:6TBZ	Smad5_gly PDB:6TCE	Smad3_5 PDB:6ZMN
Source organism	Homo sapiens	Homo sapiens	Homo sapiens
DNA source	Synthesis	Synthesis	Synthesis
Forward primer	CGCGAACAGATCGGTGGTTCCTTTACCAGCCCGGCAGTA	CGCGAACAGATCGGTGGTTCCTTTACCAGCCCGGCAGTA	GAAGTTCTGTTTCAGGGCCCGGCGGTTAAACGCTTATTGGGCTGGAAG
Reverse primer	TGGTCTAGAAAGCTTTATGGCGGTAAGACGGGACT	TGGTCTAGAAAGCTTTATGGCGGTAAGACGGGACT	CACCAGGCTTTTCACCGCCTTCTCGGCCCATTTTTCCTCCTCATCACCTTG
Cloning vector	PopIn S	PopIn S	PopIn F
Expression vector	PopIn S	PopIn S	PopIn F
Expression host	E. Coli	E. Coli	E. Coli
Complete amino acid sequence of the construct produced	Mothers against decapentaplegic homolog 5 TSPAVKRLLGWKQGEQNGQEEKWAEKAVDALVKKLKKKKGAMEELEKALSSPGQPSKCVTIPRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLDICEFPFGSKQKEVCINPYHYKRVESPVLPP DNA (5′-D(P*TP*GP*CP*AP*GP*GP*CP*GP*CP*GP*CP*CP*TP*GP*CP*A)-3′) TGCAGGCGCGCCTGCA	Mothers against decapentaplegic homolog 5 SFTSPAVKRLLGWKGGGSQGDEEEKWAEKAVDALVKKLKKKKGAMEELEKALSSPGQPSKCVTIPRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLDICEFPFGSKQKEVCINPYHYKRVESPVLPP DNA (5′-D(P*TP*GP*CP*AP*GP*GP*CP*TP*AP*GP*CP*CP*TP*GP*CP*A)-3′) TGCAGGCTAGCCTGCA	Mothers against decapentaplegic homolog 3 GPAVKRLLGWKQGDEEEKWCEKAVKSLVKKLKKTGQLDELEKAITTQNVNTKCITIPRSLDGRLQVSHRKGLPHVIYCRLWRWPDLHSHHELRAMELCEFAFNMKKDEVCVNPYHYQRVETPVLP DNA(5′-D(P*TP*GP*CP*AP*GP*GP*CP*GP*CP*GP*CP*CP*TP*GP*CP*A)-3′) TGCAGGCGCGCCTGCA

- Smad8 (PDB:6FZT): 100 nL drop volume, 30 μL reservoir solution: 0.1 M bis-Tris propane pH 8.5, 20% PEG 3350, 0.2 M NaF.
- Smad5_3 (PDB:6TBZ): 200 nL drop volume, 30 μL reservoir solution: 50 mM sodium citrate pH 5.5, 22% PEG 3350.
- Smad5_gly (PDB:6TCE): 200 nL drop volume, 30 μL reservoir solution: 50 mM HEPES pH 7.0, 21% PEG Smear Medium (PEG 2000, 3350, 4000, 5000 MME).
- Smad3_5 (PDB:6ZMN): 150 nL drop volume, 30 μL reservoir solution: 20% PEG 3350, 0.2 M sodium acetate.

Crystals were cryo-protected in mother liquid supplemented with glycerol.

Diffraction data were recorded at the ESRF beamline ID30a3 (Grenoble, France) and at the ALBA beamline BL13-XALOC (Barcelona, Spain). The data were processed, scaled and merged with autoPROC [81] applying the anisotropy correction by STARANISO [79]. The $CC_{1/2}$ criterion was used for selecting the diffraction resolution cut-off [23]. Initial phases were obtained by molecular replacement using PHASER [58,59] from the CCP4 and PHENIX suites [49,83] (search model PDB code: 3KMP) with anisotropic correction. REFMAC [65] phenix.refine [49] and BUSTER [76] 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 cutoff (paired-refinement) and for the structure model optimization [39] Table 2. Specifically bound water molecules at the protein-DNA interface were collected in Supplementary Table S1. UCSF Chimera [69] was used to prepare figures and calculate RMSD values for structural comparisons (Supplementary Table S2).

2.3. Triple resonance backbone and relaxation experiments

NMR data were acquired on a Bruker Avance III 600-MHz spectrometer equipped with a cryogenic probe head and a z-pulse field gradient unit at 298 K using Non-Uniform Sampling (NUS) and BEST-TROSY backbone experiments [8,48,66,68,77]. Proline residues were connected using a set of specific experiments [13]. T_1 and T_2 relaxation measurements were acquired using standard pulse sequences [8]. The rotational correlation times (τ_c) of the Smad5, Smad5-gly as well as of their complexes with DNA were calculated essentially as described [73,54,5] using several protein concentrations in buffer 1 supplemented with 10% D_2O. Spectra were processed with NMRPipe [20] and MddNMR [66]. Backbone assignment was performed with CARA [10] as previously described [54,5]. T_1, T_2 and hetNOE data were processed and integrated with TopSpin3.5, Bruker BioSpin Corp. (https://www.bruker.com). K_D fittings from R_2 values were calculated with GraphPad Prism following established procedures [11].

Molecular Weight determinations from τ_c values were performed using the correlations determined by the Northeast Structural Genomics Consortium and collected in the literature [72], with the following relation: MW = 1.569*τ_c + 0.4972.

2.4. SAXS data

Data were collected at Beamline 29 (BM29) at the European Synchrotron Radiation Facility (ESRF; Grenoble, France) on samples of Smad5 MH1 domain at protein concentrations ranging from 0.96 to 10 mg/mL dissolved in buffer 1. Protein samples were prepared as previously described [5]. Small-angle scattering data were deposited in the SASBDB database under the entry code SASDE32. Acquisition parameters are described in Supplementary Table S3.

2.5. ChIP-Seq statistical analysis

Chip-Seq datasets (SRP179614 for Smad3 and GSM1810980 for Smad1/5) were downloaded from the NCBI SRA [47] and GEO [9] data repositories. The fastq format data was extracted with

Table 2
Data collection and processing. Values for the outer shell are given in parentheses.

Protein	Smad5	Smad8	Smad5_3	Smad5_gly	Smad3_5
PDB code	6FZS	6FZT	6TBZ	6TCE	6ZMN
Data collection	ESRF-ID30a3	ESRF-ID30a3	ALBA BL13	ALBA BL13	ALBA BL13
Space group	P212121	P212121	$P3_2$	$I4_122$	$P2_12_12_1$
a, *b*, *c* (Å)	67.76, 73.37, 89.19	75.43, 79.51, 88.37	53.14, 53.14, 83.15	106.12, 106.12, 82.40	54.50, 73.42, 111.15
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 120.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Resolution (Å)*	53.96–2.31 (2.48–2.31)	59.11–2.46 (2.63–2.46)	46.02–1.82 (2.08–1.82)	53.06–2.92 (3.06–2.92)	48.94–2.37 (2.64–2.37)
$R_{r.i.m}$	0.130 (1.413)	0.097 (1.403)	0.087 (0.885)	0.049 (3.208)	0.179 (0.965)
$R_{p.i.m}$	0.052 (0.555)	0.041(0.550)	0.038 (0.578)	0.017 (1.129)	0.069 (0.531)
$I/\sigma(I)$	11.4 (1.5)	14.3 (1.3)	10.9 (1.9)	21.1 (0.6)	8.9 (1.2)
$CC_{1/2}$Completeness (%):	0.998 (0.529)	0.998 (0.572)	0.999 (0.578)	1.000 (0.315)	0.998 (0.552)
Spherical ellipsoidal#	83.0 (22.0)	79.7 (22.4)	48.2 (7.3)	90.1 (34.6)	69.0 (12.7)
	92.4 (44.6)	92.0 (54.3)	88.8 (68.2)	91.9 (40.6)	91.8 (57.9)
Redundancy	9.3 (9.9)	6.0 (6.3)	5.0 (3.9)	7.9 (7.9)	6.6 (2.9)
Refinement					
Resolution (Å)	29.77–2.31	59.11–2.46	46.00–1.82	53.11–2.92	26–2.33
Number of unique refl.	16,665	15,793	11,400	4823	12,973
R_{work} / R_{free}	0.193 / 0.241	0.193 / 0.237	0.193 / 0.226	0.210 / 0.253	0.210 / 0.252
No. of atoms	2815	2852	1702	1167	2670
Protein	2010	2057	988	842	1978
DNA	656	656	656	324	656
Zinc ions	2	2	1	1	2
Water	143	123	57	0	16
B factors					
Protein	48	59	51	142	42
DNA	89	116	86	116	83
Zinc ions	41	53	27	157	30
Water	48	56	36	NA	21
R.M.S.D.					
Bond lengths (Å)	0.010	0.010	0.008	0.010	0.008
Bond angles (°)	1.01	1.04	0.92	0.97	0.92
Ramachandran (%):	98.4	97.3	98.4	96.8	95.3
Favored	0.0	0.0	0.0	1.7	0.0
Outliers					

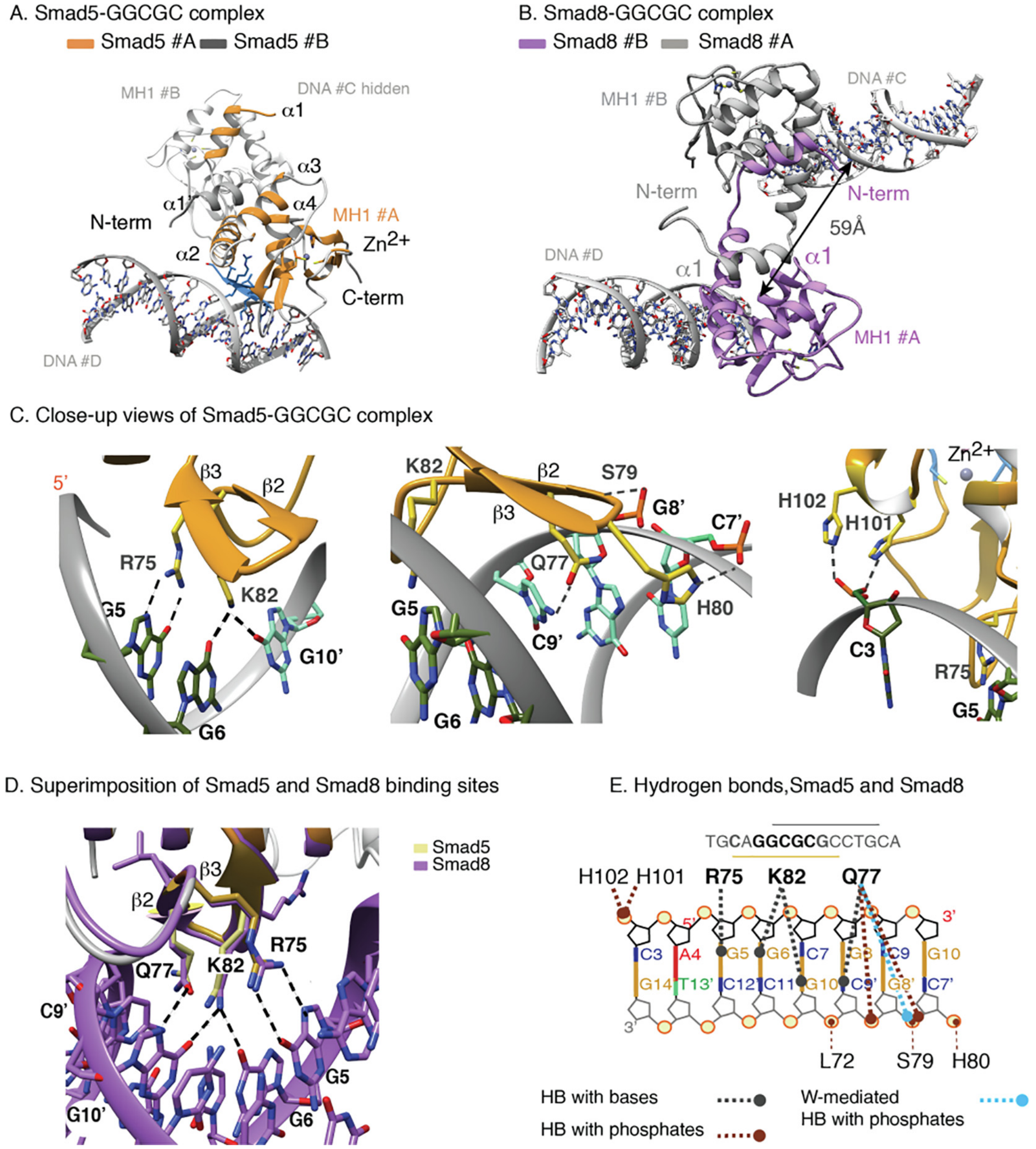

Fig. 1. Structures of Smad5 and Smad8 bound to the GGCGC site. A. One of the possible ASU representations. Ribbon diagram of the Smad5 MH1 domain homodimers in complex with the GGCGC DNA motif. Protein chains are shown in orange and grey, and the DNA-binding region is shown in blue. The Zn-binding site is indicated. The elements of secondary structure are labelled. B. Biological assembly representation. Ribbon diagram of the homodimeric Smad8 MH1 in complex with the GGCGC DNA motif. Protein chains are shown in purple and grey, and the distance between DNA binding sites is indicated. C. Close-up of the Smad5 protein-DNA binding interface. Interacting residues are shown as blue sticks and hydrogen bonds (HBs) as dotted lines. The residues and bases involved in the interaction are labelled. D. Superposition of protein-DNA binding interface for both Smad5 and Smad8. E. Schematic representation of the protein-DNA contacts observed for Smad5 and Smad8 complexes. Grey lines indicate HBs between protein residues and DNA bases whereas brown lines indicate HBs with the DNA phosphates. Blue lines indicate water-mediated HBs with DNA phosphates observed in the Smad5 complex. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

fastq-dump (**SRA Toolkit**, SRA toolkit, SRa Toolkit Development Team, http://ncbi.github.io/sra-tools/). Bowtie2 [46] was used to align the fastq reads against the mouse mm10 assembly, and sambamba [78] was used to sort and remove duplicate and unmapped reads. Peak calling was performed with macs1.4, for consistency with the Smad1/5 dataset [26].

The Smad1/5 data was downloaded from the GEO database. As the original data was aligned against the mm9 genome assembly, the UCSC liftOver software [34] was used to convert the coordinates to the mm10 assembly. All ChIP-seq peaks were normalized to 200 bp centered with respect to the peak center, using a custom python script and all regions were scanned for SBE + 5GC motifs (SBE: GTCTG, 5GC: GGCTG, GGCGC and GGCCG) and the number of motifs per Kb as well as the distance of each motif to the nearest one in the same band downstream the genome were determined. By doing this we obtained a number of distances per band which is equal to the number of detected motifs minus 1. For promoter analysis, we assigned each ChIP-Seq peak to the nearest annotated Transcription Start Site, and then extracted the ones related to the genes of interest. The gene names and the coordinates for the bands are included in Supplementary Table S4. The Smad3 ChIP-seq was used to extract bands for the TGF-β regulated genes and the Smad1/5 ChIP-Seq was used for the BMP regulated ones.

The Statistical analysis was performed using R language, version 3.6.3 (R Core Team (2017) R: A Language and Environment for Statistical Computing) using the R built-in functions *ad.test* for the Anderson-Darling normality test and *wilcox.test* for the Wilcoxon rank sum test. Plots displayed as Fig. 6A and B were generated with the ggplot2 package.

2.6. TWIM-MS experiments

Experiments were performed using a Synapt G1 HDMS mass spectrometer (Waters UK Ltd., Manchester, UK) and essentially as described, [4,73]. Mass spectra were acquired by positive nano-electrospray ionization (ESI) using a Nanospray Triversa (Advion Biosciences Corpn., Ithaca, NY, USA) interface. To optimize the separation of the different conformers, traveling-wave drift times of selected ions corresponding to monomers and dimers of Smad MH1 domains (in 150 mM ammonium acetate buffer) were measured at wave heights of 7 V, 8 V, 9 V, and 10 V and at a velocity of 300 m/s. Data acquisition and processing were carried out using MassLynx (v4.1) software. Drift time calibration of the T-Wave cell was performed using β-lactoglobulin (monomer, 18 kDa, and dimer, 37 kDa) from bovine milk. Reduced cross-sections (Ω') were calculated from published cross-sections [14] and subsequently plotted against final corrected drift times (tD). Calibration coefficients were determined applying an allometric y = AxB fit. Experimental cross-sections were determined by measuring the drift time centroid for the molecular-related ions by means of Gaussian fitting to the drift time distribution (Prism v6, GraphPad Software Inc., California, USA). Acquisition parameters are described in Supplementary Table S5.

An extended description of the methods is provided in https://www.biorxiv.org/content/10.1101/833319v3

3. Results

3.1. Smad5 and Smad8 complexes with the 5GC DNA site

Before setting up the crystallization screenings, we studied the protein-DNA interactions by EMSA and NMR (Supplementary Fig. S1B), and observed that the interaction with 5GC motifs was

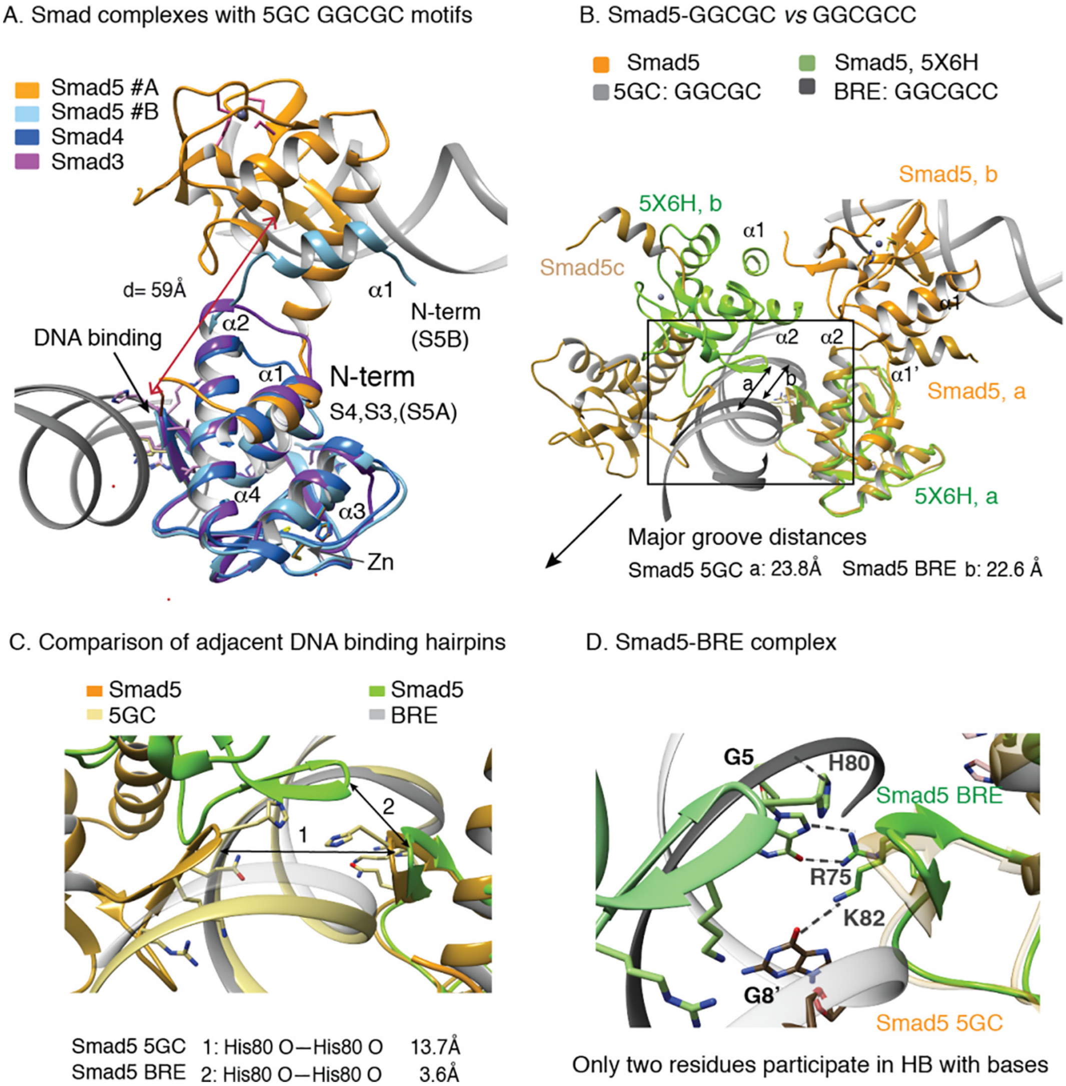

Fig. 2. Structural comparison of different Smads bound to DNA. A. Overlay of dimeric Smad5 MH1 domains (gold and light blue) and the monomeric Smad4 (royal blue, PDB:5MEY) and Smad3 (purple, PDB: 5OD6). All backbones are shown as ribbons. Some secondary structural elements are labeled. N-term of all proteins are indicated. For simplicity only the DNA bound by Smad5 is displayed. B. Comparison of the Smad5-GGCGC structure (orange) to that of Smad5-BRE (green, PDB:5X6H). Major groove distances are indicated to highlight the DNA compression observed in the Smad5-BRE complex. The first helix of the Smad5-BRE is drawn as described [7] although the electron-density map for this region is not well defined. C. Comparison of adjacent DNA binding hairpins in both Smad5 complexes. Distances between the hairpins are indicated. D. Close-up view of the comparison of the DNA binding site. Only two residues, R75 and K82, display specific HB with DNA in the 5X6H complex. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

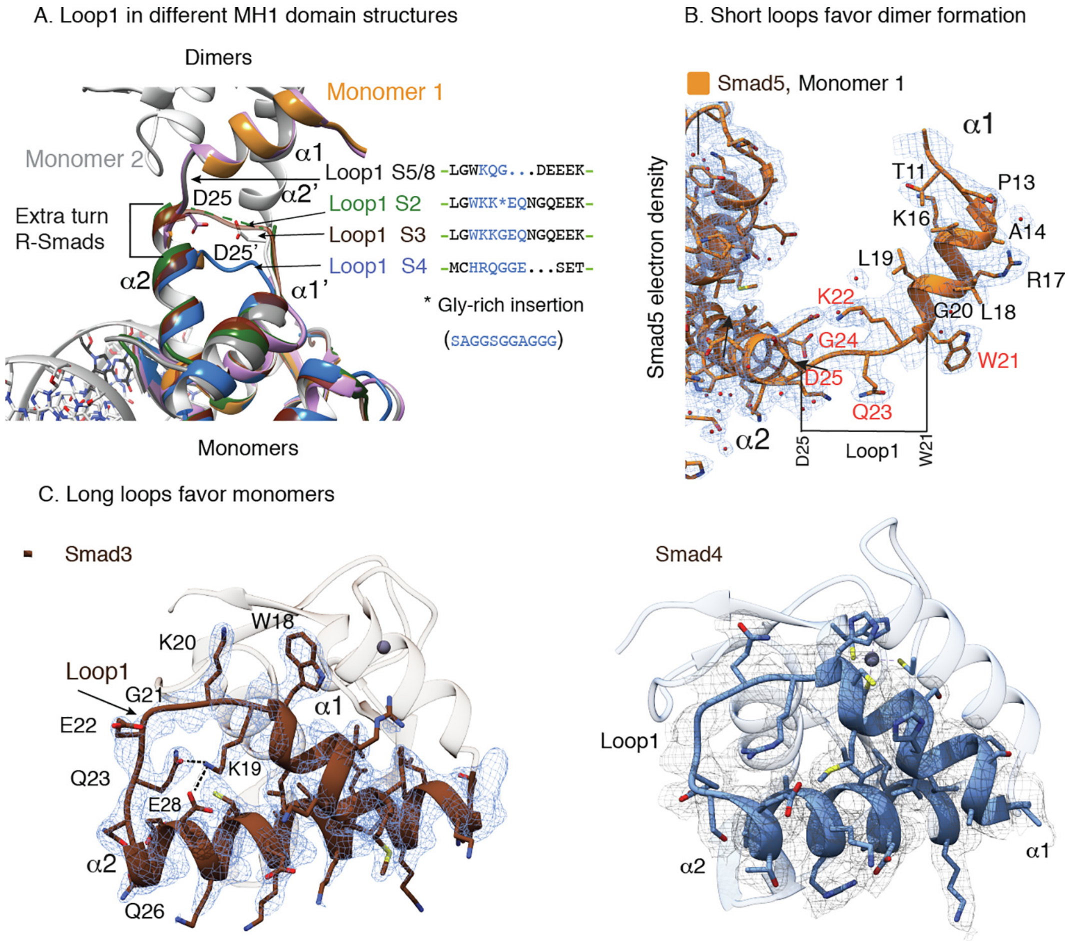

Fig. 3. Key features that define the dimer/monomer propensities. A. Overlay of the complexes of Smad5 MH1 (dimer, gold, and light blue) and the monomers corresponding to Smad4 (royal blue, PDB:5MEY) and Smad3 (purple, PDB: 5OD6) complexes. The view is focused on loop1. All backbones are shown as ribbons. Some secondary structural elements are labeled. The N-terminus of all proteins is indicated. Loop1 sequences are shown in blue. Differences in loop1 and helix α2, which is longer in R-Smads than in Smad4, are indicated with arrows and as a bracket, respectively. The positions of D25 (the first residue of the helix α2) for both Smad5 monomers are indicated, to highlight the inability of BMP-Smads to bridge the distance between helix α1 and α2 with a short loop. Loop1- residues located at the Smad5 dimer interface are labeled in blue and black. These residues are also conserved in Smad1 and Smad8. B. A section of the electron-density map of the Smad5 dimer (this work), Smad3 monomer (PDB: 5OD6) C (left) and Smad4 (PDB: 5MEY) C (right) are shown. In all cases, maps are contoured at 1.0σ and are shown for helixes 1 and 2 and loop1. Some side chains and HB are indicated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in the same affinity range as that observed for the SBE sequence. Crystals were obtained in several conditions, but the best diffracting ones were crystalized with a 16 bp dsDNA containing the 5GC sequence (underlined) TGCAGGCGCGCCTGCA (note that the internal part of the oligo, GGCGCGCC, is palindromic and therefore the dsDNA contains two 5GC sites).

We solved the Smad5-5GC complex (space group $P2_12_12_1$) by molecular replacement using a model derived from the Smad1/GTCT complex (PDB: 3KMP) and then used the Smad5/5GC complex to solve that of Smad8 bound to the same DNA.

In both complexes, the asymmetric unit contains two MH1 domains and two DNA strands. These models were refined at 2.31 Å and 2.46 Å for Smad5 and Smad8, respectively. The biological assembly is defined as a protein dimer, with the α1 helix being swapped between monomers, and with each monomer bound to a ds-DNA molecule (Fig. 1A, B). Data collection and statistics are shown in Table 2. The electron densities for the Smad5 and Smad8 proteins and the bound DNA are well-defined for the entire complexes (Supplementary Fig. S1C,D) and the Smad5 and Smad8 MH1 structures display all the characteristic features of MH1 domains including the presence of a Zn^{2+} coordinated by three cysteines and one histidine [54,74]. These domains are composed of four helices (arranged as a four-helical bundle) and six strands arranged as three anti-parallel pairs (β1-β5, β2-β3, and β4-β6). The protein DNA binding region comprises the loop following the β1 strand, and the β2–β3 hairpin (residues 70–83). This hairpin contains Arg75, Gln77 and Lys82 residues, which are strictly conserved in all MH1 domains. These residues interact directly with the major groove through a set of hydrogen bonds (HBs) with the GGCGCg motif (Fig. 1C). An additional network of HB interactions between Ser79, Leu72, Gln77, (backbone atoms) and His101 and His102 (side chain) with Gua8′, Gua10′ and Cyt3 bases reinforced the complex stability (Fig. 1C, middle and right). There is also a set of 10 well-ordered water molecules bound at the protein-DNA-binding interface that contribute to the stability of the complex (Supplementary Table S1). Similar interactions are observed for the Smad8/5GC complex (Fig. 1D). When superimposed, the Smad5 and the Smad8 MH1 domains are nearly identical (Cα RMSD of 0.25 Å for 124 aligned residues) and the complexes are very similar to that of the dimeric Smad1 bound to the SBE GTCT site (PDB 3KMP; Cα RMSD of 0.30 Å for 123 aligned residues, Supplementary Table S2). The observed contacts are represented as a cartoon in Fig. 1E, showing that one bound MH1 domain covers the 3-CAGGCGC-9 area.

Overall, these results show that homodimers of Smad5 and Smad8 MH1 domains interact with DNA using the conserved binding mode displayed by all Smad proteins. Given the sequence conservation of Smad1/5/8 at the MH1 dimer interface (Supplementary Figs. S1E,F,G), we hypothesize that heterodimers of Smad1/5/8 might also be formed. Similar homo- and heterodimeric associations have been observed with Myc-Max and Mad-Max transcription factors, where each specific complex determines whether the gene targets are activated or silenced, respectively [31].

3.2. SBE and 5GC DNA sites: One binding mode for all Smads

Except for the MH1 domains monomer/dimer arrangement and regardless of the SBE or 5GC DNA motif type, the protein-DNA binding interface of all R-Smads and Smad4 is very similar; see Fig. 2A for Smads bound to the 5GC GGCGC motif [54]. The similarity is reflected by the conserved pattern of interactions between the protein and the DNA and by the RMSD value of their Cα superimposition (Supplementary Table S2). Even the general DNA topology of the major groove (the principal binding site of all complexes) is conserved between the different bound 5GC DNAs (Supplementary Fig. S2), as characterized using Curves [12]. These complexes also revealed that one MH1 domain is efficiently accommodated on one full DNA major groove, with a clear distinction of minor and major grooves in all SBE and 5GC Smad complexes, without introducing protein-DNA structural clashes or distortions.

3.3. Smad5-5GC complex comparison to the Smad5-BRE complex

The interaction of Smad5 with a palindromic sequence of six bp (GGCGCC) named BRE-GC (5X6H) has been reported [17]. In this complex (Fig. 2B, C), the electron density for the residues in the loop1 and for most residues present in the helix α1 is absent and it is unclear how the N-terminal part is arranged in the complex. Moreover, the interaction with the DNA is distorted, with both DNA grooves showing similar depth and width (Supplementary Fig. S2) and varying from all other Smad-DNA structures solved to date [7,54], including the structures determined here. This binding interface is highly distorted and shows the fewest specific hydrogen bonds between protein and DNA of all Smad complexes determined to date.

At first glance, the sequences of BRE-GC (GGCGCC) and 5GC (GGCGCG) appear to be remarkably similar. However, the BRE motif is a 3 bp palindrome, and two MH1 domains were modeled to interact with the 6 bp BRE-GC site in the 5X6H structure (one MH1 domain is bound to each half of the palindrome, Fig. 2B, C). This effect causes both a huge geometrical perturbation in the B-DNA and a reduction of specific HBs (Fig. 2D) with the protein due to steric hindrance when compared to the 5GC complexes determined here.

Considering that Smad proteins bind to cis-regulatory elements containing clusters of consecutive motifs [54], we believe that the most probable binding mode *in vivo* is that observed in the 5GC and SBE complexes. It seems very unlikely that two MH1 domains would interact as in the BRE complex —using half of their protein binding site and causing a high distortion to the DNA structure— if there is the possibility to interact with two neighboring sites using the full protein binding interface and a perfect accommodation to the DNA.

3.4. Molecular bases of MH1 domain dimer/monomer propensities

In Smads, the differences in sequence are higher between Smad4 and the R-Smads than between BMP- and TGFβ-activated R-Smads themselves. The most substantial differences are detected either at the linker connecting the MH1 and MH2 domains or at loops within these two domains [52]). If we focus on the differences observed in the MH1 domains, Smad1/5/8 have four residues in loop1, whereas the same loop has six residues in Smad3 and in Smad4, and sixteen in Smad2 (Supplementary Figs. S1E and S3) [5]). The different lengths of loop1 and helix2 seem to have an impact on Smads dimerization state. The superposition of Smad protein-DNA structures (Fig. 3A) shows that loop1 of Smad2/3/4 is long enough to bridge the distance between helices α1 and α2 in one monomer (even though Smad2/3 have helix α2 one turn longer than Smad4). In contrast, for Smad1/5/8 the combination of a longer helix α2 (as long as in Smad2/3) with a short loop1 seems to make impossible such compact packing of helices in one monomer. Instead, the loop1 and the helix α1 protrude away and are swapped between two monomers to form a dimer as observed here for Smad5/8-5GC complexes (Fig. 3B) and by others for Smad1 (3KMP) and Smad5 (5X6G) bound to the SBE site [7,17]. In the case of Smad3 (5OD6), the turn in loop1 is stabilized by internal HBs (Fig. 3C left) [54,75]. Regarding other monomeric Smads, Smad2 (6H3R) has a long Gly-rich insertion (indicated as an asterisk Fig. 3A) and its loop1 is mostly disordered [5], whereas in Smad4 (5MEY), the loop1 is well-defined without the presence of internal HBs (Fig. 3C right) [54].

Overall these observations suggest that the differences in sequence observed at the N-terminal region of the MH1 domains can condition the monomer or dimer conformations adopted by the domains.

3.5. Swapping the loop1 sequence between Smads is enough to switch the dimerization propensities of MH1 domains in crystals

Upon the observation that different R-Smads have specific propensities to form monomer or swapped dimers when bound to SBE and GC-rich DNA sequences, we set to investigate if modifying the loop1 length, might condition these structural propensities. Correlations between loop length, domain swapping and protein folding mechanisms have been reported in the literature [15,18,22,30,40,50,80]

We started by increasing the length and flexibility of the Smad5 loop1 by inserting a GGGS sequence into the loop (Smad5_gly mutant). In the resulting structure the asymmetric unit (ASU) contains one protein chain and one DNA strand, and the biological assembly is represented by two MH1 domains bound to each GGCT site of the palindromic dsDNA. Although, in this structure, the crystal packing resembles that of the monomeric Smad3 structure (5ODG) and not that of the dimeric wild-type Smad5 structure, the loop1 could not be fully traced in the electron density map due to the flexible nature of the GGGS insertion (Supplementary Fig. S4A). For the second mutant, we replaced the Smad5 loop1 by that of Smad3 (Smad5_3 chimera). In this case, we obtained well-diffracting crystals of the complex with the 5GC DNA, which allowed us to solve the structure at 1.8 Å (the highest resolution of all MH1-DNA complexes to date) (Table 2). Here, ASU and the biological assembly are the same and contain one MH1 domain bound to one dsDNA molecule. The structure includes specific side chain-nucleobase contacts with all five 5GC bp. Moreover, the loop1 is ordered as in the monomeric Smad3 structure. The loop position is well-supported by the electron density 2Fo-Fc map and by the Fo-Fc omit map when refining against a structure lacking the loop1 residues (Fig. 4A, Supplementary Fig. S4B). Regarding the DNA recognition, the binding mode is conserved with respect to the Smad5 homodimer (and other Smads), with the presence of specific HBs being formed between the highly-conserved R75, K82, and Q77 residues, and the GGCGC motif (Fig. 4B). Additionally, Q77 and H83 side chains make HBs with phosphates of Gua8′ and

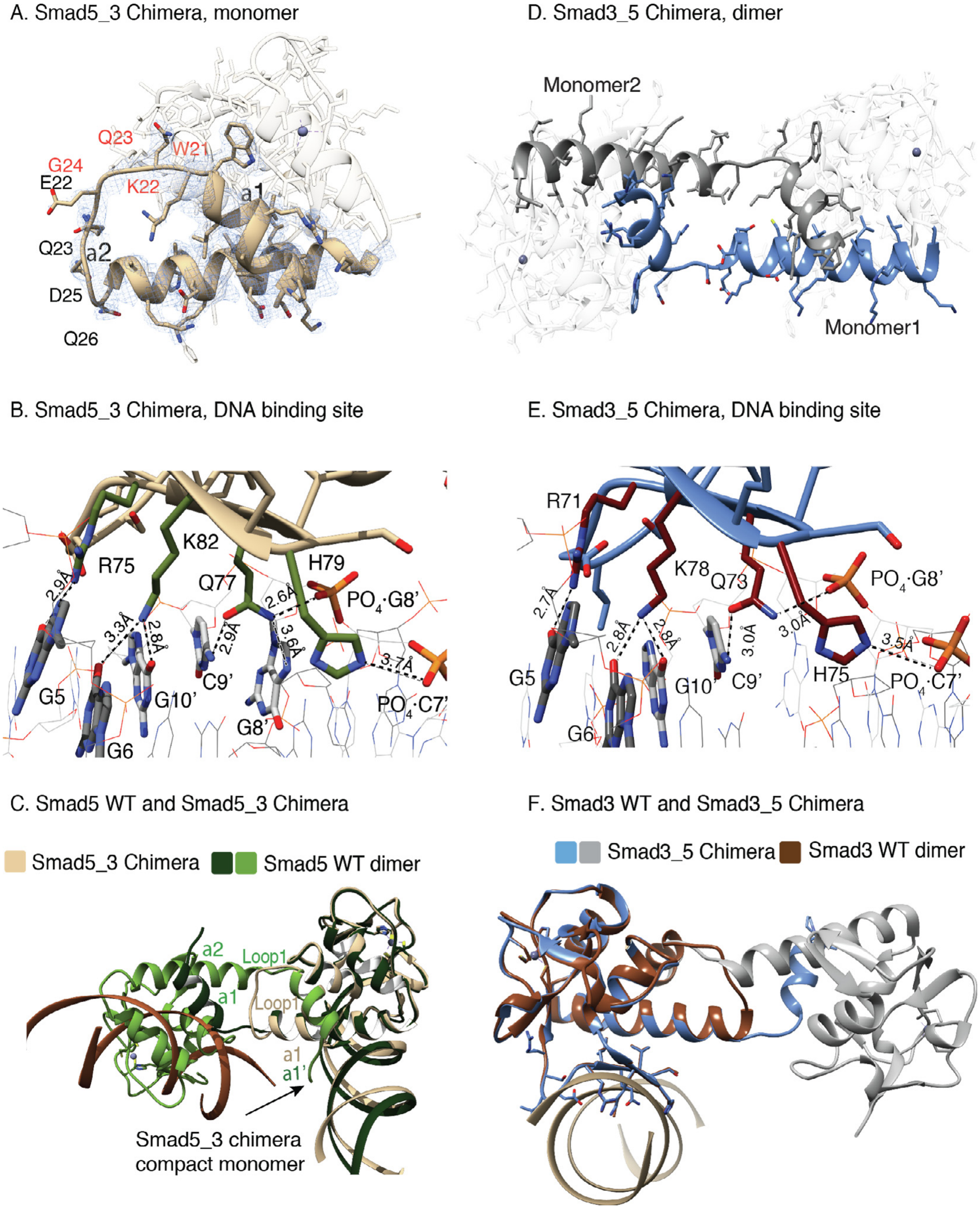

Fig. 4. Structures of the three Smad5 chimeras. A. Smad5_3 chimera folded as a monomer. The omit map corresponding to the loop1 is shown as Supplementary Fig. 5B right. Some elements of the secondary structure and residues in and around loop1 are labeled. B. Close-up view of the Smad5_3-DNA binding interface. Residues located at the protein-DNA binding site are shown in green. HBs are indicated as dotted lines, and distances are indicated in Å. DNA backbone phosphates involved in protein interaction are shown as sticks and labeled. C. Superposition of Smad5 dimer and monomer (green and beige). The loop1 in the monomer (chimeric construct) and the dimer interface (Smad5 WT) are labeled. The arrow indicates the compact fold of the chimeric construct. D. Smad3_5 chimera folded as a dimer. Loop1 and α1α2 helices are indicated and colored in gray and blue. The omit map corresponding to the loop1 is shown as Supplementary Fig. 5B (right). The rest of the dimer is shown as a semi-transparent ribbon. E. Close-up view of the DNA binding interface. Smad3_5 residues participating in HBs are shown in dark red and the DNA in gray. HBs are indicated as dotted lines and distances are indicated in Å. F. Superposition of Smad3 monomer (sienna) to the Smad3_5 dimer (blue and gray). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Cyt7′, respectively. The comparison of the chimeric construct and the wild type dimer is shown in Fig. 4C.

Prompted by this result, we also set out to investigate the reversed effect and determine whether shortening the loop1 will favor the helix-swapped dimer of Smad3 over the monomer. For that, we generated the Smad3_5 chimera (Smad3 MH1 domain with Smad5 loop1 and a Smad5-like I11A mutation to mimic the Smad5 helix α1 sequence and lower the helix hydrophobicity). In this case, we solved the Smad3_5 chimera-DNA complex structure at 2.3 Å resolution (ASU contains two protein chains and two DNA strands). The biological assembly is the same as for other dimeric MH1-DNA complexes (MH1 dimer bound to two DNA molecules) and the loop1 is well-structured and supported by the electron density (2Fo-Fc and Fo-Fc omit) maps in both chains of the dimer (Supplementary Fig. S4C). Other than loop1, the rest of the protein and the DNA interface are mostly identical to those observed in the wild-type Smad3 MH1 domain bound to this DNA motif (specific R71, K78, and Q73 side chain-nucleobase contacts with 4 bp of

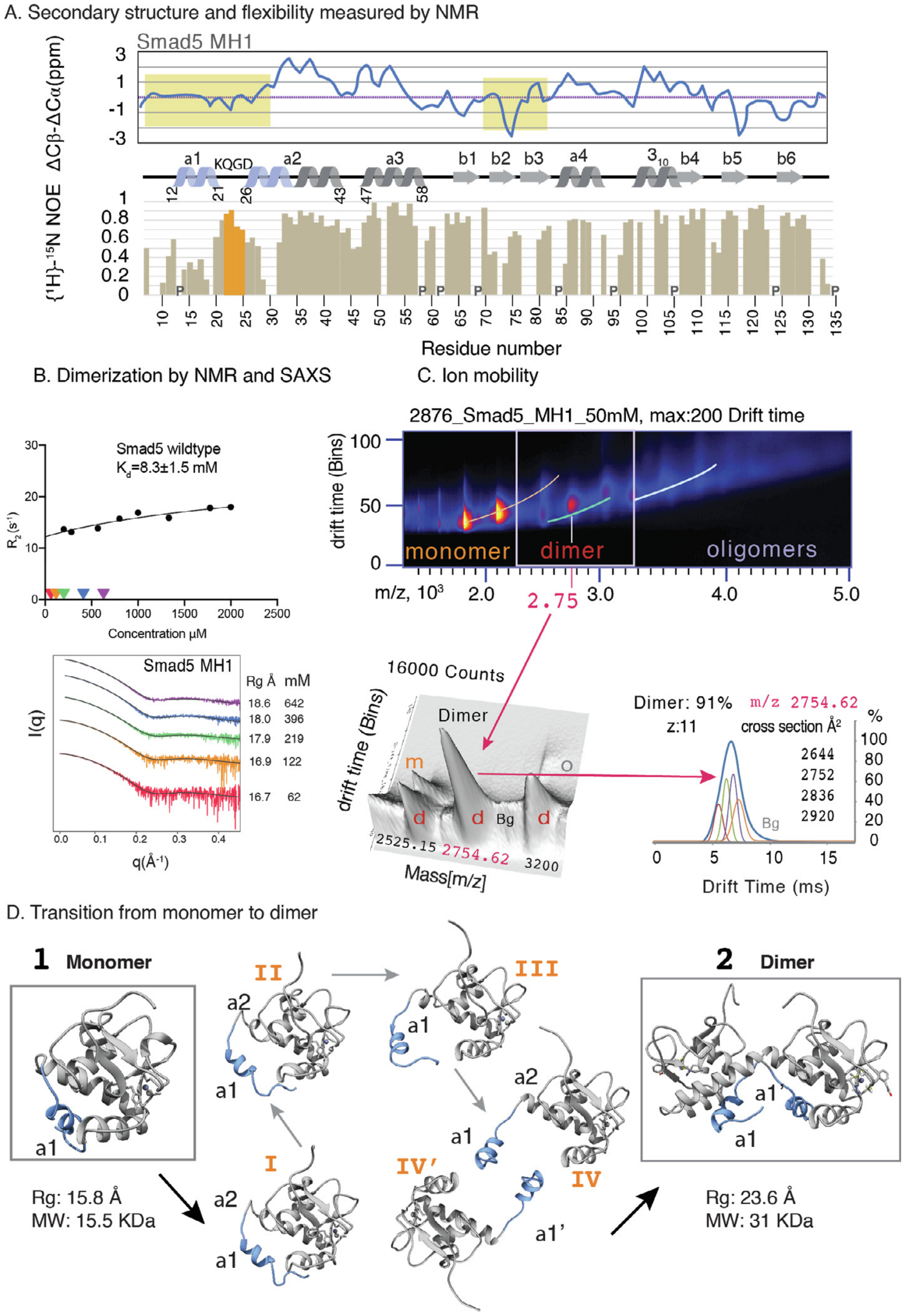

Fig. 5. Smad5 Monomers and Dimers in solution and in gas phase. A. Secondary structure elements of the Smad5 MH1 domain based on $^{13}C\alpha$-$^{13}C\beta$ chemical shift differences with respect to random coil values and NOE patterns. Positive values indicate α helices whereas negative ones correspond to β sheet structures. The region highlighted in yellow represents values close to zero, an indication of undefined secondary structure (first helix and loop1 as well as the β1- β2 DNA binding hairpin). Below, hetNOE values at 800 MHz ^1H Larmor frequency ordered according to the residue number. Proline (P) residues indicated along the sequence. B. Top: Dimerization K_d estimation using NMR (R_2 vs. concentration). The arrow heads indicate the concentrations used in SAXS experiments. Bottom: Small-angle X-ray scattering (SAXS) data at five protein concentrations and derived Radius of Gyration in Å (Rg). C. Plot of the mobility drift time versus m/z for the Smad5 MH1 domain, with the peaks corresponding to monomer, dimer and oligomers (minority) labeled. As depicted in the figure, monomer and dimer forms (more compact) travel faster than oligomeric ones (instrumental settings and analysis are provided in Supplementary Table S5 and in methods). The region containing the dimers is shown as intensities and the cross-section analysis is indicated for the m/z 2754.62 peak, which is fully described as contributions from dimeric conformations. The mass spectra from all ions and for Smad3 and Smad4 MH1 domains are shown as Supplementary Fig. S4. D. Structural models generated to illustrate the dynamic transition from the compact monomer towards the dimer and vice versa. This transition requires the flexibility of the α1 and α2 helices and a reorientation of loop1, with a concomitant separation of the α1 helix from the protein core, (models I, II, III). Model IV adopts an extended conformation of the N-terminus, suited to interact with a second monomer and to define a dimer. Compact monomers were generated using the Smad4 MH1 structure (PDB: 5MEY) as the template and the Smad5 sequence whereas the dimers correspond to the Smad5 structures determined in this work. The transition from monomer to dimer was generated using Chimera. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the 5GC site; Fig. 4D, E) [54]. A comparison of the chimeric and the wild-type Smad3 structures is shown as Fig. 4F.

In summary, our three structures confirm that the propensity of MH1 domains to form dimers or monomers is mainly encoded within the loop1 length and observed in crystals of complexes with DNA. Of note, whereas the longer loop1 of the TGFβ-responsive Smads (monomers) tend to be more variable in length, the short loop1 of BMP-responsive Smads (dimers) has been conserved

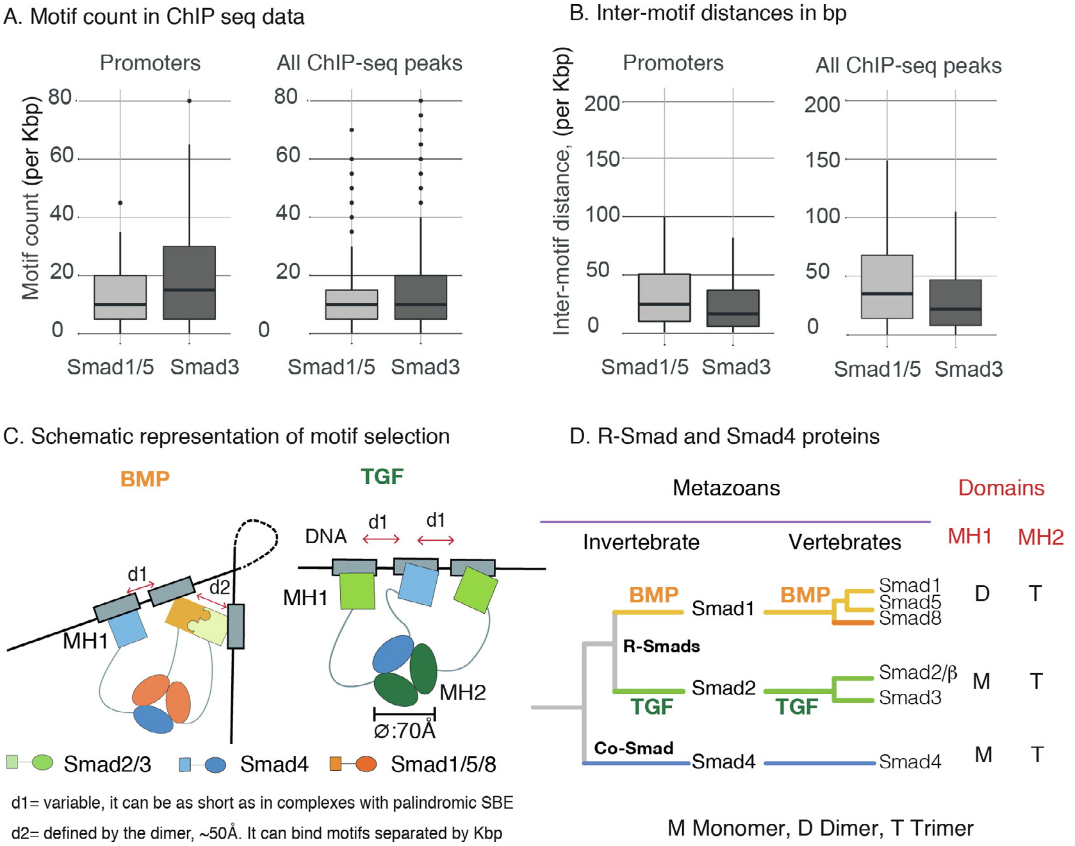

Fig. 6. A. Box-plot representing the motif count (SBE + 5GC) for Smad1/5 and Smad3 ChIP-Seq datasets, for BMP- (Smad1/5) and TGFβ- (Smad3) regulated promoters. Right. Box-plot of motif counting for all detected peaks for Smad1/5 and Smad3 ChIP-Seqs. The list of promoters is available as Supplementary Table S4. B. Box-plot of the inter-motif distance, in bp, for the motifs (SBE + 5GC) for Smad1/5 and Smad3 ChIP-Seqs, for BMP- (Smad5) and TGFβ- (Smad3) regulated promoters. Right. Box-plot of the inter-motif distance for all detected peaks for Smad1/5 and Smad3 ChIP-Seqs C. Schematic representation of Smad1/5-Smad4 and Smad2/3-Smad4 heterotrimers binding to SBE/5GC sites. The dimers of Smad1/5/8 MH1 domains would select DNA sites separated by the distance of the two DNA-binding hairpins in the dimer (a distance of approx. 50 Å) and with the Smad4 bound without restrictions. The motifs recognized by the dimer can be located far away as observed in the analysis of motif distribution in BMP bound peaks. Conversely, MH1 monomers of Smad2/3-Smad4 heterotrimers are not restricted by the location of their binding sites and can bind to both adjacent or distant motifs, although the motif distribution in the clusters seemed to indicate a preference for local clusters. D. Schematic representation of the different Smad proteins in metazoans, including the duplications observed in vertebrates, adapted from [52]. For each R-Smad protein we indicated the propensity of their MH1 domains to fold as monomers and dimers (M, D) and of their MH2 domains to associate as heterotrimers with Smad4 (T).

during metazoan evolution (Fig. S3B), indicating a potential connection between the MH1 domain dimerization propensity and different biological functions of TGFβ- and BMP-activated Smads.

3.6. Dimers and monomers in solution and in gas phase

To clarify whether there is an intrinsic propensity of Smad5 MH1 domains to associate as dimers in non-crystallographic conditions and in the absence of DNA, we sought to explore the protein behavior in solution by NMR, under similar conditions to those reported for Smad2/3/4 in the literature [5,54]. To this end, we first acquired sets of backbone triple resonance experiments to facilitate the assignment of the Smad5 protein resonances. We observed a good agreement of the ^{13}C chemical shift values (CSV) for the elements of the secondary structure, except for the N-terminus (residues located at the α1 helix and at the beginning of α2 helix), whose CSV were close to random coil. These residues also displayed low heteronuclear NOE values, thereby suggesting the presence of conformational flexibility of these two helical regions (Fig. 5A). Unexpectedly, residues in the loop1, (connecting both helices) do not adopt secondary structure but are less flexible than the α1 and α2 helices themselves, perhaps reflecting the presence of an extended conformation in the loop, which could facilitate the inter-domain association. The overall flexibility at the N-terminus was not observed for Smad2/3/4 MH1 domains studied under similar conditions [5,54]. We also observed low heteronuclear NOE values in and around the β2- β3 hairpin (DNA binding site), as previously observed for other Smads [5,54]. Furthermore, the T_1 and T_2 values corresponding to non-overlapped residues of the Smad5 protein (MW of 15.1 kDa) yielded an average correlation time (τ_c) of 12.1 ns (at 0.6 mM) measured at 850 and 600 MHz respectively. Such τ_c values are in agreement with a sample close to ~20–22 kDa [72], which seems to contain monomers and dimers in equilibrium. When the Smad5_Gly monomeric chimera was measured under the same conditions, we obtained an average correlation time (τ_c) of 8.13 ns, corresponding to a compact and monomeric sample with a MW of ~13.25 KDa [72]. In order to estimate the dimerization constant of the Smad5 association (K_d) we measured the overall R_2 value of the protein at different concentrations, ranging from 200 µM to 2.0 mM), following the approach described in the literature [11] and obtained a K_d = 8.3 ± 1.5 mM (Fig. 5B).

We also compared the correlation times of Smad5 in complex with a dsDNA containing a single Smad binding site, to facilitate the interpretation of the results. We found that upon addition of 1 equiv of DNA (7.3 KDa) the τ_c of the solution increased up to 20.8 ns, suggesting a MW of ~33 KDa, larger than 22.4 KDa expected for a monomeric 1:1 protein:DNA complex. In contrast, a titration of Smad5-GlyLoop chimera with 1 equiv of DNA yields a τ_c value of 16.2 ns, in agreement with a MW of ~21 KDa monomeric 1:1 compact complex. Unfortunately, the presence of DNA induces protein precipitation in both cases, precluding the determination of the dissociation constant and relaxation properties.

The presence of several conformations in solution that depend on the protein concentration was also corroborated by SAXS, using five concentrations, from 62 to 642 μM (more concentrated samples precipitated after freezing in liquid nitrogen). The data obtained indicated an interval of the radius of gyration between 16.7 and 18.6 Å, dependent on the concentration, which is between 15.8 Å corresponding to a compact monomeric form [5,54] and 23.6 Å for the fully formed dimer (Fig. 5B, Supplementary Table S3). These values are in agreement with the K_d constant determined by NMR.

Finally, dimers of Smad5 were also detected by ion mobility mass spectrometry (IM-MS) in the gas phase. This technique separates ions on the basis of their differential mobility through a gas buffer, without disrupting native structures or oligomeric associations [27,60]. We measured various micromolar concentrations of Smad5 and Smad3/4, as controls of monomeric MH1 domains. In all cases we monitored the presence of bound Zn^{2+} as proof of folded samples after buffer exchange [32]. The analyses of m/z and drift time values revealed that, in the gas in gas phase, Smad5 populates monomeric, dimeric and even larger oligomeric conformations, with all these distinct species resolved at different drift times (Fig. 5C). Remarkably, the presence of dimer/monomer species was detected at all protein concentrations evaluated (from 20 up to 150 μM) and under slightly different experimental conditions (Supplementary Table S5). The analysis of the collision cross-section (CCS) area of two selected peaks containing the dimer forms (m/z 2754.6 and 3029.98 (Fig. 5C and Supplementary Fig. S5-A-C) confirmed the presence of several dimeric conformations in Smad5 samples though monomer and dimer conformations were detected in other peaks (such as m/z 2525.15). It is worth noting that, under the same conditions, the most abundant species of Smad3 and Smad4 MH1 domains were monomeric forms, while signals arising from the dimer conformations were detected as minority species (Supplementary Fig. S5D-F).

Altogether, these results indicate that Smad5 MH1 domains present a high degree of plasticity in non-crystallographic conditions consistent with the presence of monomers in equilibrium with dimers, whose presence is also observed in the presence of DNA.

3.7. Distribution of Smad-binding motifs in BMP- and TGFβ-responsive elements

Smad-binding sites tend to occur in clusters of three or more Smad-binding motifs, thereby facilitating the interaction of Smad complexes in regulatory elements [54]. To assess whether the distribution of Smad binding sites could reflect the monomer or dimer preferences, we sought to analyze ChIP-Seq data available in public databases for both BMP and TGFβ-activated Smad proteins performed under similar experimental conditions. We found two datasets for Smad1/5/8 (regulated by BMP) and Smad2/3 proteins (regulated by TGFβ) that fulfil these requirements (mESC, E14 cell lines) [5,64]. For motif counting, we divided the study into two separate analyses, either peaks found in promoters or enhancers of genes or all regions with Smad-bound peaks (without specific loci localization for each dataset, collected in Supplementary Table S4) [25,35,85]. We normalized all ChIP-Seq peaks to be 200 bp and scanned the set of known 5-bp long Smad-binding motifs (SBE and 5GC) to obtain their frequencies in these regulatory regions. For comparison, we also studied the distribution of the 6-base GC-BRE site (GGCGCC), a GGC palindrome that overlaps with the 5GC GGCGC sequence also described as a Smad1/5/8 binder [45].

When we scanned the set of specific genes, we found that TGFβ-regulated promoters have a higher average count of Smad-binding motifs than the BMP ones (**20** motifs/Kb and **13** motifs/Kb, respectively). When the analysis was extended to all peaks, we observed a similar trend, although the values were slightly smaller than before (**16** motifs/Kb and **9** motifs/Kb respectively), (Fig. 6A). Our analysis showed that the GC-BRE motif is not significantly enriched in the Smad1/5 and Smad3 datasets (29% of the 5GC GGCGC motifs are followed by C to form the GC-BRE in comparison to 25% value in case of the uniform distribution). The motif distribution was analysed by the Anderson-Darling normality test, yielding non-normally distributed data as indicated by p-values < 0.01, well below the 0.05 threshold value. We also compared both non-normal distributions using the Wilcoxon rank sum test and obtained a p-value = 2.2e−16, (much lower than 0.05 threshold), confirming that the differences in motif distribution were statistically significant.

Further, we measured the inter-motif distances for each cluster, calculated with respect to the ChIP-Seq peak-center and observed that the motifs showed no particular preference for being localized near the center or at the boundaries.

On average, we observed that the Smad3 data has a higher number of motifs/Kb than that of Smad1/5, with the Smad-binding motifs being separated by ~**46 bp** in BMP-activated regions and by ~**33 bp** in TGFβ ones (16 bp are equivalent to 50 Å distance), (Fig. 6B). For comparison, in uniform distributions, the expected values would have been 83 bp for BMP (1000/(13-1 motifs/Kb)) = 83 bp) and 52 bp for TGFβ ones (1000/(20-1)) = 52 bp. Again, the data were non-normally distributed (Anderson-Darling test), and the Wilcoxon test showed that the differences in inter-motif distances were significant (p-value under 2.2e−16 ≪ 0.05).

4. Discussion

Our results confirm that BMP-activated Smads form MH1 dimers by exchanging the α1 helix between two monomers in solution and in crystals of 5GC DNA complexes. We also observed that, in the absence of DNA, the Smad5 MH1 domain has a tendency to populate an ensemble of conformations in solution including monomers and dimers. We could also correlate the dimeric propensity to the loop1 because swapping this loop between Smad3 and Smad5 is enough to revert their native propensities as revealed by the structures we have determined. Moreover, our results confirm the hypothesis that all R-Smads and Smad4 (monomers or dimers) are able to interact specifically with 5GC and SBE sites by means of a conserved binding mode, mostly using the β2-β3 hairpin.

It is now well accepted that R-Smad/Smad4 heterotrimeric associations are driven *via* interactions of the conserved MH2 domains of two R-Smads and a single Smad4 protein, as visualized in the crystal structures of various complexes of MH2 domains [21,27,32,52]. The high level of MH2 domain conservation among different R-Smads would, in theory, allow for virtually all combinations of Smad proteins in a native context.

However, only complexes with the presence of Smad1/5/8 and either Smad2/3 or Smad4 proteins have been experimentally detected using full-length proteins [19,32]. If these complexes are trimers, the specific composition of these ternary complexes seems to require a second layer of selection rules to favor some complexes over others. Some of these rules could include holding specific combinations of Smad proteins whose MH1 domains form either dimers or monomers. For instance, either a dimer of Smad1/5/8 and one monomer of either Smad2/3/4 (whose MH1 domains cannot form dimers) or a trimer of Smad2/3/4 (all monomers). The formation of these complexes in the full-length protein context would be facilitated by the flexibility provided by the long linkers (80 residues) connecting the domains.

The ChIP-Seq analysis revealed clusters containing a few adjacent Smad-binding sites in the peaks recognized by BMP-activated Smads and a higher frequency of such sites in the TGFβ-activated ones [54]. These differences in motif distribution could be rationalized based on the distinct structural features of the MH1 domains that interact with them. For monomers, the higher frequency of sites can correlate with inter-motif distances being as small as two consecutive DNA sites that enable binding of two MH1 domains without steric hindrance and, theoretically, as big as the length of the extended long linker loop connecting the MH1 and MH2 domains. In the case of dimers, we detected a more spread motif distribution, since the latter might need to fulfill the dimer-specific spatial requirements (approx. 60 Å distance, Fig. 6C) if both monomers bind to DNA. The above implies that not all theoretically available DNA sites can be recognized by all kinds of Smad trimers and that, given the DNA looping and the flexibility of long MH1-MH2 linkers, a given Smad complex could bind DNA sites separated by variable distances. This versatility would explain the experimental evidence showing how a given Smad complex can recognize different regions in promoters [51,63].

Overall, all findings available till now support the hypothesis that the selection of optimal DNA targets results from a collaborative work of bound cofactors and the Smad trimers, whose composition is finely adjusted by the presence or absence of MH1 domain interactions. In these scenarios, all components will fit to tune the context-dependent action of BMP and TGFβ signals and final transcriptional outcomes. Certainly, additional experiments, as well as structures of the full-length Smad complexes bound to DNA, will bring us closer to a complete picture on how these different layers of interactions are defined.

5. Conclusions

Our findings suggest that the composition of Smad heterotrimeric complexes may be modulated by the association through MH1 dimers, and not only through MH2 domain interactions, explaining why not all combinations of Smad complexes are detected in cellular experiments. We propose that this characteristic has been among the keys to shaping two classes of R-Smad proteins since the origin of metazoans (Fig. 6D). MH1 domain dimerization of BMP-activated Smads could also play a role in the recognition of DNA sites genome-wide. For a given Smad heterotrimer, finding the optimal DNA sites must fulfill certain specific spatial requirements dictated by the MH1 domain structures. However, these spatial requirements allow for some freedom to recognize a range of motif separation, which could explain why, for a given promoter, the distances between motifs are not strictly conserved among vertebrates.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dr. N. Berrow (IRB Barcelona, protein expression unit) for help with some DNA constructs and with protein purification. We also thank the EMBL staff for assistance at the HTX facility (Grenoble), the joint EMBL and ESRF group for access to synchrotron beamlines ID29, ID23-1 and ID23-2 and the staff at The Automated Crystallography Platform staff (IRB Barcelona-CSIC) and at the ALBA synchrotron (Barcelona) for access to the BL13-XALOC beamline. Thanks also go to Dr. J. Massagué for insightful suggestions and discussions, Drs. M. Díaz and M. Vilaseca (Mass Spectrometry Core Facility, IRB Barcelona) for support with the IM-MS data, Dr. M. Navia for suggestions on binding assays, and Dr. B. Brutscher, (Institut de Biologie Structurale, Biomolecular NMR Spectroscopy Group, Grenoble, France) for help with the acquisition of the NMR data at 850 MHz. We also thank J. Cordero for some preliminary experiments.

Funding

T.G. was a PhD student funded by a Severo Ochoa and by the BBVA. Z.K., R.F, R.P., and B.B were co-funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie COFUND actions of the EMBL, IRB Barcelona and the PROBIST and PREBIST Postdoc and Predoc Programmes (agreements EMBL_291772, IRBPostPro2.0_600404 and PROBIST_754510, and PREBIST_754558). M.J.M is an ICREA Programme Investigator. This work was supported by the Spanish MINECO program (BFU2014-53787-P and BFU2017-82675-P, M.J.M), IRB Barcelona and the BBVA Foundation. Access to the HTX facility at EMBL (Grenoble) was granted by the Horizon 2020 Programme i*NEXT* of the European Commission (grant 653706, title: Smad complexes), and to the NMR facility (Grenoble) by the Instruct Integrating Biology program (grant 2520, title: Monomer-dimer equilibrium in Smad proteins). Access to BioSAXS BM29 was part of the MX-1941 BAG proposal and to ALBA through the BAG proposal 2018092972. We gratefully acknowledge institutional funding from the CERCA Programme of the Catalan Government and from the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the Centres of Excellence Severo Ochoa award.

Author contributions

L.R. and E.A. cloned, expressed and purified all proteins, L.R., R.F., C.T., and N.M. performed EMSA experiments. L.R., T.G., T.N.C., performed and analyzed the SAXS measurements and P.M.M., L.R., and M.J.M. acquired and analyzed the NMR data. L.R., and T.G. analyzed the IM-MS data. P.M.M. analyzed the clustering of DNA motifs in ChIP-Seq data. Z.K. B.B. and R.P. screened crystallization conditions, collected X-ray data, determined the structures and analyzed them with J.A.M., and M.J.M. All authors contributed ideas to the project. M.J.M. and R.P. supervised the project. M.J.M. wrote the manuscript with contributions from all other authors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.12.044.

References

[1] Alarcon C, Zaromytidou AI, Xi Q, Gao S, Yu J, Fujisawa S, et al. Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. Cell 2009;139:757–69.

[2] Amoutzias GD, Robertson DL, Van de Peer Y, Oliver SG. Choose your partners: dimerization in eukaryotic transcription factors. Trends Biochem Sci 2008;33:220–9.

[3] Aragon E, Goerner N, Zaromytidou A-I, Xi Q, Escobedo A, Massague J, Macias MJ. A Smad action turnover switch operated by WW domain readers of a phosphoserine code. Genes Dev 2011;25:1275–88.

[4] Aragón E, Goerner N, Xi Q, Gomes T, Gao S, Massagué J, Macias M. Structural basis for the versatile interactions of Smad7 with regulator WW domains in TGF-β pathways. Structure 2012;20:1726–36.

[5] Aragon E, Wang Q, Zou Y, Morgani SM, Ruiz L, Kaczmarska Z, et al. Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-beta signaling. Genes Dev 2019;33:1506–24.

[6] BabuRajendran N, Jauch R, Tan CY, Narasimhan K, and Kolatkar P. Structural basis for the cooperative DNA recognition by Smad4 MH1 dimers. Nucleic Acids Res (2011) 39, 8213-8222.

[7] BabuRajendran N, Palasingam P, Narasimhan K, Sun W, Prabhakar S, Jauch R, et al. Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-beta effectors. Nucleic Acids Res 2010;38:3477–88.

[8] Barbato G, Ikura M, Kay LE, Pastor RW, Bax A. Backbone dynamics of calmodulin studied by 15N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. Biochemistry 1992;31:5269–78.

[9] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 2013;41:D991–5.

[10] Bartels C, Xia TH, Billeter M, Guntert P, and Wuthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR (1995) 6, 1-10.

[11] Baryshnikova OK, Sykes BD. Backbone dynamics of SDF-1α determined by NMR: interpretation in the presence of monomer–dimer equilibrium. Protein Sci 2006;15:2568–78.

[12] 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 2011;39:W68–73.

[13] Bottomley MJ, Macias MJ, Liu Z, Sattler M. A novel NMR experiment for the sequential assignment of proline residues and proline stretches in 13C/15N-labeled proteins. J Biomol NMR 1999;13:381–5.

[14] Bush MF, Hall Z, Giles K, Hoyes J, Robinson CV, Ruotolo BT. Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. Anal Chem 2010;82:9557–65.

[15] Cafaro V, De Lorenzo C, Piccoli R, Bracale A, Mastronicola MR, Di Donato A, and D'Alessio G. The antitumor action of seminal ribonuclease and its quaternary conformations. FEBS Lett (1995) 359, 31-34.

[16] Chacko BM, Qin BY, Tiwari A, Shi G, Lam S, Hayward LJ, De Caestecker M, and Lin K. Structural basis of heteromeric smad protein assembly in TGFbeta signaling. Mol Cell (2004) 15, 813-823.

[17] Chai N, Li WX, Wang J, Wang ZX, Yang SM, and Wu JW. Structural basis for the Smad5 MH1 domain to recognize different DNA sequences. Nucleic Acids Res (2017) 45, 6255-6257.

[18] Cregut D, Civera C, Macias MJ, Wallon G, Serrano L. A tale of two secondary structure elements: when a beta-hairpin becomes an alpha-helix. J Mol Biol 1999;292:389–401.

[19] Daly AC, Randall RA, Hill CS. Transforming growth factor beta-induced Smad1/5 phosphorylation in epithelial cells is mediated by novel receptor complexes and is essential for anchorage-independent growth. Mol Cell Biol 2008;28:6889–902.

[20] Delaglio F, Grzesiek S, Vuister GeertenW, Zhu G, Pfeifer J, Bax Ad. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6.

[21] Derynck R, Budi EH. Specificity, versatility, and control of TGF-beta family signaling. Sci Signal 2019;12.

[22] Donato AD, Cafaro V, Romeo I, D'Alessio G. Hints on the evolutionary design of a dimeric RNase with special bioactions. Protein Sci 1995;4:1470–7.

[23] Diederichs K, Karplus PA. Better models by discarding data?. Acta Crystallogr D Biol Crystallogr 2013;69:1215–22.

[24] Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr 2010;66:486.

[25] Fei T, Xia K, Li Z, Zhou B, Zhu S, Chen H, Zhang J, Chen Z, Xiao H, Han JD, et al. Genome-wide mapping of SMAD target genes reveals the role of BMP signaling in embryonic stem cell fate determination. Genome Res 2010;20:36–44.

[26] Feng J, Liu T, Zhang Y. Using MACS to identify peaks from ChIP-Seq data. Curr Protocols Bioinf 2011;34.

[27] Flanders KC, Heger CD, Conway C, Tang B, Sato M, Dengler SL, et al. Brightfield proximity ligation assay reveals both canonical and mixed transforming growth factor-beta/bone morphogenetic protein Smad signaling complexes in tissue sections. J Histochem Cytochem 2014;62:846–63.

[29] Gao S, Alarcon C, Sapkota G, Rahman S, Chen PY, Goerner N, et al. Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signaling. Mol Cell 2009;36:457–68.

[30] Gotte G, Mahmoud Helmy A, Ercole C, Spadaccini R, Laurents DV, Donadelli M, and Picone D. Double domain swapping in bovine seminal RNase: formation of distinct N- and C-swapped tetramers and multimers with increasing biological activities. PLoS One (2012) 7, e46804.

[31] Grandori C, Cowley SM, James LP, Eisenman RN. The Myc/Max/Mad network and the transcriptional control of cell behavior. Annu Rev Cell Dev Biol 2000;16:653–99.

[32] Grönroos E, Kingston IJ, Ramachandran A, Randall RA, Vizan P, Hill CS. Transforming growth factor beta inhibits bone morphogenetic protein-induced transcription through novel phosphorylated Smad1/5-Smad3 complexes. Mol Cell Biol 2012;32:2904–16.

[33] Guca E, Suñol D, Ruiz L, Konkol A, Cordero J, Torner C, Aragon E, Martin-Malpartida P, Riera A, Macias MJ. TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF-β signaling. Nucleic Acids Res 2018;46(17):9220–35.

[34] Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 2019;47:D853–8.

[35] Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 2018;46:D380–6.

[36] Hou C, Tsodikov OV. Structural basis for dimerization and DNA binding of transcription factor FLI1. Biochemistry 2015;54:7365–74.

[37] Huang Y-H, Jankowski A, Cheah KSE, Prabhakar S, Jauch R. SOXE transcription factors form selective dimers on non-compact DNA motifs through multifaceted interactions between dimerization and high-mobility group domains. Sci Rep 2015;5(1):10398.

[38] Jayaraman L, Massagué J. Distinct oligomeric states of SMAD proteins in the transforming growth factor-beta pathway. J Biol Chem 2000;275:40710–7.

[39] Joosten RP, Long F, Murshudov GN, Perrakis A. The PDB_REDO server for macromolecular structure model optimization. IUCrJ 2014;1:213–20.

[40] Josephson K, Logsdon NJ, Walter MR. Crystal Structure of the IL-10/IL-10R1 complex reveals a shared receptor binding site. Immunity 2001;15:35–46.

[41] Kashima R, Hata A. The role of TGF-β superfamily signaling in neurological disorders. Acta Biochim Biophys Sin (Shanghai) 2018;50:106–20.

[42] Katagiri T, Imada M, Yanai T, Suda T, Takahashi N, Kamijo R. Identification of a BMP-responsive element in Id1, the gene for inhibition of myogenesis. Genes Cells 2002;7:949–60.

[44] Kawabata M, Inoue H, Hanyu A, Imamura T, Miyazono K. Smad proteins exist as monomers in vivo and undergo homo- and hetero-oligomerization upon activation by serine/threonine kinase receptors. EMBO J 1998;17:4056–65.

[45] Kusanagi K, Inoue H, Ishidou Y, Mishima HK, Kawabata M, Miyazono K, Heldin C-H. Characterization of a bone morphogenetic protein-responsive Smad-binding element. MBoC 2000;11:555–65.

[46] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.

[47] Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res 2011;39:D19–21.

[48] Lescop E, Schanda P, Brutscher B. A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. J Magn Resonance 2007;187:163–9.

[49] Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, Hintze B, Hung L-W, Jain S, McCoy AJ, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol 2019;75:861–77.

[50] Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11:1285–99.

[51] López-Rovira T, Chalaux E, Massagué J, Rosa JL, Ventura F. Direct binding of Smad1 and Smad4 to two distinct motifs mediates bone morphogenetic protein-specific transcriptional activation of Id1 Gene. J Biol Chem 2002;277:3176–85.

[52] Macias MJ, Martin-Malpartida P, Massagué J. Structural determinants of Smad function in TGF-β signaling. Trends Biochem Sci 2015;40:296–308.

[53] Macias MJ, Wiesner S, Sudol M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett 2002;513:30–7.

[54] Martin-Malpartida P, Batet M, Kaczmarska Z, Freier R, Gomes T, Aragon E, Zou Y, Wang Q, Xi Q, Ruiz L, et al. Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors. Nat Commun 2017;8:2070.

[55] Massagué J. TGF-beta signal transduction. Annu Rev Biochem 1998;67:753–91.

[56] Massagué J. How cells read TGF-beta signals. Nat Rev Mol Cell Biol 2000;1:169–78.

[57] Massagué J. TGF-beta signalling in context. Nat Rev Mol Cell Biol 2012;13:616–30.

[58] McCoy AJ. Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallogr D Biol Crystallogr 2007;63:32–41.

[59] McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. J Appl Crystallogr 2007;40:658–74.

[60] Medina E, Córdova C, Villalobos P, Reyes J, Komives E, Ramírez-Sarmiento César A, Babul J. Three-dimensional domain swapping changes the folding mechanism of the Forkhead Domain of FoxP1. Biophys J 2016;110:2349–60.

[61] Miyazono K-I, Moriwaki S, Ito T, Kurisaki A, Asashima M, Tanokura M. Hydrophobic patches on SMAD2 and SMAD3 determine selective binding to cofactors. Sci Signal 2018;11:eaao7227.

[62] Morales B, Ramirez-Espain X, Shaw AZ, Martin-Malpartida P, Yraola F, Sánchez-Tilló E, Farrera C, Celada A, Royo M, Macias MJ. NMR Structural Studies of the ItchWW3 domain reveal that phosphorylation at T30 inhibits the interaction with PPxY-containing ligands. Structure 2007;15:473–83.

[63] Morikawa M, Koinuma D, Miyazono K, Heldin C-H. Genome-wide mechanisms of Smad binding. Oncogene 2013;32:1609–15.

[64] Morikawa M, Koinuma D, Mizutani A, Kawasaki N, Holmborn K, Sundqvist A, et al. BMP sustains embryonic stem cell self-renewal through distinct functions of different kruppel-like factors. Stem Cell Rep 2016;6:64–73.

[65] Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 1997;53:240–55.

[66] Orekhov VY, Jaravine VA. Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. Prog Nucl Magn Reson Spectrosc 2011;59:271–92.

[67] Park J, Throop AL, LaBaer J. Site-specific recombinational cloning using gateway and in-fusion cloning schemes. Curr Protocols Mol Biol 2015;110.

[68] Pervushin K, Riek R, Wider G, Wuthrich K. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci 1997;94:12366–71.

[69] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera? A visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–12.

[70] Ramachandran A, Vizan P, Das D, Chakravarty P, Vogt J, Rogers KW, et al. TGF-beta uses a novel mode of receptor activation to phosphorylate SMAD1/5 and induce epithelial-to-mesenchymal transition. Elife 2018;7.

[71] Roberts AB, Tian F, Byfield SD, Stuelten C, Ooshima A, Saika S, et al. Smad3 is key to TGF-beta-mediated epithelial-to-mesenchymal transition, fibrosis, tumor suppression and metastasis. Cytokine Growth Factor Rev 2006;17:19–27.

[72] Rossi P, Swapna GVT, Huang YJ, Aramini JM, Anklin C, Conover K, Hamilton K, Xiao R, Acton TB, Ertekin A, Everett JK, Montelione GT. A microscale protein NMR sample screening pipeline. J Biomol NMR 2010;46:11–22.

[73] Schelhorn C, Gordon JMB, Ruiz L, Alguacil J, Pedroso E, Macias MJ. RNA recognition and self-association of CPEB4 is mediated by its tandem RRM domains. Nucleic Acids Res. 2014;42:10185–95.

[74] Shi Y, Massagué J. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. Cell 2003;113:685–700.

[75] Shi Y, Wang YF, Jayaraman L, Yang H, Massagué J, Pavletich NP. Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. Cell 1998;94:585–94.

[76] Smart OS, Womack TO, Flensburg C, Keller P, Paciorek W, Sharff A, Vonrhein C, Bricogne G. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. Acta Crystallogr D Biol Crystallogr 2012;68:368–80.

[77] Solyom Z, Schwarten M, Geist L, Konrat R, Willbold D, Brutscher B. BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. J Biomol NMR 2013;55:311–21.

[78] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;31:2032–4.

[79] Tickle IJ, Flensburg C, Keller P, Paciorek W, Sharff A, Vonrhein C, Bricogne G. STARANISO. Cambridge, UK: Global Phasing Ltd.; 2018.

[80] Tsitsanou KE, Drakou CE, Thireou T, Vitlin Gruber A, Kythreoti G, Azem A, Fessas D, Eliopoulos E, Iatrou K, Zographos SE. Crystal and Solution Studies of the "Plus-C" Odorant-binding Protein 48 from Anopheles gambiae: control of binding specificity through three-dimensional domain swapping. J Biol Chem 2013;288:33427–38.

[81] Vonrhein C, Flensburg C, Keller P, Sharff A, Smart O, Paciorek W, Womack T, Bricogne G. Data processing and analysis with the autoPROC toolbox. Acta Crystallogr D Biol Crystallogr 2011;67:293–302.

[82] Wang RN, Green J, Wang Z, Deng Y, Qiao M, Peabody M, Zhang Q, Ye J, Yan Z, Denduluri S, Idowu O, Li M, Shen C, Hu A, Haydon RC, Kang R, Mok J, Lee MJ, Luu HL, Shi LL. Bone Morphogenetic Protein (BMP) signaling in development and human diseases. Genes Dis 2014;1:87–105.

[83] Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS. Overview of the CCP 4 suite and current developments. Acta Crystallogr D Biol Crystallogr 2011;67:235–42.

[84] Zander U, Hoffmann G, Cornaciu I, Marquette J-P, Papp G, Landret C, Seroul G, Sinoir J, Röwer M, Felisaz F, Rodriguez-Puente S, Mariaule V, Murphy P, Mathieu M, Cipriani F, Márquez JA. Automated harvesting and processing of protein crystals through laser photoablation. Acta Crystallogr D Struct Biol 2016;72:454–66.

[85] Zhang Y, Handley D, Kaplan T, Yu H, Bais AS, Richards T, et al. High throughput determination of TGFbeta1/SMAD3 targets in A549 lung epithelial cells. PLoS One 2011;6:e20319.