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

Authors:

Lidia Ruiz[a,1], Zuzanna Kaczmarska[b,c,1], Tiago Gomes[a,1], Eric Aragón[a], Carles Torner[a], Regina Freier[a], Błażej Bagiński[a], Pau Martin-Malpartida[a], Natàlia de Martin Garrido[a], José A. Márquez[b], Tiago N. Cordeiro[d], Radoslaw Pluta[a,2] and Maria J. Macias[a,e,2]

Affiliations:

[a]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST). Baldiri Reixac, 10, Barcelona 08028, Spain;

[b] EMBL Grenoble, 71 Avenue des Martyrs, CS 90181, 38042 Grenoble Cedex 9, France;

[c] Present address: International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109, Warsaw, 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, 08010-Barcelona, Spain.

[1]These authors contributed equally to the work

[2]To whom correspondence should be addressed:

radoslaw.pluta@irbbarcelona.org

maria.macias@irbbarcelona.org

Running title

Dimer/monomer propensities of Smad MH1 domains

Keywords 10

Smad structures, MH1 domain, BMP signaling, Smad5 protein, Smad8 protein GC-motifs, Smad DNA complex, Smad conservation and evolution

Footnotes

Supplemental information is available for this article.

Summary

R-Smads are effectors of the transforming growth factor β (TGFβ) superfamily and along with Smad4 form trimers to interact with other transcription factors and with DNA. The 5GC-DNA complexes determined here by X-ray crystallography for Smad5 and Smad8 proteins corroborate that all MH1 domains bind SBE and 5GC sites similarly, although Smad2/3/4 MH1 domains bind DNA as monomers whereas Smad1/5/8 form helix-swapped dimers. To examine the relevance of the dimerization phenomenon and to exclude a possible crystallography-induced dimeric state, we studied these MH1 domains in solution. The results show that Smad5/8 domains populate dimers and open monomers in equilibrium, whereas their Smad2/3/4 counterparts adopt monomeric closed conformations. We also found that swapping the loop1 sequence between Smad5 and Smad3 results in the Smad5 chimera-DNA complex crystallizing as a monomer, revealing that the loop1 sequence determines the monomer/dimer propensity of Smad MH1-domains. We propose that distinct MH1-dimerization status of TGFβ and BMP activated Smads influences the interaction with specific *loci* genome-wide by distinct R-Smad and Smad4 complexes.

Significance

TGFβ- and BMP-activated R-Smads were believed to have different preferences with respect to the recognition of DNA motifs and to respond to specific activation inputs. However, recent results indicate that several types of R-Smads can be activated by similar receptors and that all Smads might recognize various DNA motifs. These results pose new questions as to why different types of R-Smads have been conserved for more than 500 million years if they could have a redundant function. They also raise questions as to how different Smad complexes recognize specific clusters of DNA motifs genome-wide.

Here, using structural biology approaches, we elucidated some of the rules that help define the dimers of Smad-DNA complexes and propose how dimers and monomers could influence the composition of Smad complexes as well as the recognition of specific *cis*-regulatory elements genome-wide.

Highlights

R-Smads and Smad4 interact with 5GC and GTCT sites using a conserved binding mode.

Functional differences of TGFβ- and BMP-activated R-Smads are not exclusively related to DNA specificity.

Dimer/monomer propensities are detected in solution and in the absence of DNA.

Author Contributions: L.R., Z.K., and T.G. designed and performed most experiments and coordinated collaborations with other authors. L.R. and E.A. cloned, expressed and purified all proteins, L.R., R.F., C.T., N.M., and J.C performed EMSA and thermal denaturation experiments. L.R., E.A., T.G., T.C., P.M.M., and M.J.M. performed the SAXS and NMR measurements and analyzed the data. Z.K. B.B. and R.P. screened crystallization conditions, collected X-ray data and determined the structures. Z.K., R.P., T.G., J.A.M., and M.J.M. analyzed the structures. 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.

The authors declare no conflict of interest.

Data Deposition: NMR assignments and chemical shifts have been deposited in the Biological Magnetic Resonance Data Bank, BMRB entry 27548, and the Small-angle scattering data and models have been deposited in SASBDB, entries SASDE32 (Smad5) and SASDE42 (Smad8). Densities and coordinates have been deposited in the Protein Data Bank, entries 6FZS (Smad5), 6FZT (Smad8), 6TBZ (Smad5_3 chimera), and 6TCE (Smad5_gly mutant).

This article contains supporting information online.

Introduction

The gene responses activated by the TGFβ cytokine family (a term that includes the transforming growth factor β, bone morphogenetic proteins (BMP), Nodal, Activin and other members) play important roles in embryo development, apoptosis, tissue homeostasis, repair, and immunity [1,2]. These critical roles demand a high level of conservation and fidelity of the TGFβ signaling elements in healthy organisms [1].

The main TGFβ signal transduction mechanism is the Smad pathway, with Smad transcription factors being responsible for the transmission of the signals from the membrane receptor into the nucleus [3]. Receptor-activated Smads (R-Smads) and Smad4 (Co-Smad) are versatile proteins. They all contain a DNA-binding domain (MH1) and a protein-protein interaction region composed of the linker and the MH2 domain [4,5]. The MH1 and MH2 domains are highly conserved across Smad proteins and along evolution, whereas the linker has a higher sequence variability. R-Smad linkers contain PY motifs and phosphorylatable Ser/Thr residues, which are recognized by cofactors containing pairs of WW domains (Supplementary Figure S1A) [6,7]. After being phosphorylated at the MH2 domains by the TGFβ receptor, activated R-Smads interact with Smad4 and define the canonical trimeric functional unit. Once in the nucleus, and upon linker phosphorylation, the trimeric Smad complex is ready to define a new set of interactions with cofactors and with *cis*-regulatory elements, interactions that go on to modulate the outcome of the signaling network [8-10].

R-Smad proteins were believed to have different specificities regarding the recognition of DNA motifs and to respond to specific BMP- and TGFβ-activation inputs [11]. 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 sequence conservation of the MH1 domains (Supplementary Figure S1B) and recent experimental pieces of evidence indicate that the separation between DNA binding preferences of R-Smads is subtler than initially thought. For instance, combined TGFβ and BMP receptors influence Smad1/5-driven responses [12] and the MH1 domains of Smad3 and Smad4 proteins interact —efficiently and specifically— with GC-rich motifs grouped in the 5GC consensus GGC(GC)|(CG) [13]. This 5GC consensus is functionally relevant for TGFβ-activated Smads and for Smad4, and it overlaps with the palindromic and compressed 6-BRE site GGCGCC, previously defined as the GC-rich target sequence of BMP-activated Smads [14]. Crystal structures of Smad2/3 and Smad4 bound to GTCT and 5GC sites, as well as those of Smad1 and Smad5 bound to the

GTCT site, have been determined [13,15-18]. These structures reveal that all R-Smads and Smad4 MH1 domains are able to interact with the SBE, as well as with specific 5GC sites *in vitro,* and perhaps also *in vivo,* as the analysis of numerous ChIP-Seq experiments showed that *cis*-regulatory elements bound by the different Smad proteins are enriched in clusters of SBE and 5GC sites [13]. In all these structures, the Smad proteins interact with the 5GC and SBE sites using a distinctive binding mode. Notably, while keeping the same DNA binding characteristics, these crystal structures showed that Smad3 and Smad4 MH1 domains adopted closed conformations [13,15,18], whereas Smad1 and Smad5 domains adopted a dimer organization, where the α1 helix of one monomer was interchanged with the α1' of a second monomer [16]. To add more complexity to these observations, a structure of Smad5 bound to the palindromic compressed BRE has been determined [17]. This model contains some parts that are not well-defined and shows the fewest specific hydrogen bonds between conserved protein residues and DNA bases of all Smad complexes.

In the search for new clues to clarify how BMP-activated Smad proteins interact with non-compressed GC sites and to decipher the characteristics that define monomers and dimers of MH1 domains, we focused our attention on studying the BMP-activated Smads, which are prone to form dimers in the presence of DNA. We set out to study the interaction of Smad5/8 MH1 domains with non-compressed GC sites using X-ray crystallography, as well as to examine the conformations of these MH1 domains in solution and in the absence of DNA. These structures confirmed that MH1 domains (dimers) of BMP-activated Smads interact efficiently with a non-compressed 5GC site using the canonical protein-DNA binding mode and covering seven base pairs with a single MH1 domain. Moreover, Nuclear Magnetic Resonance, Ion Mobility-Mass Spectrometry and SAXS (Small-angle X-ray scattering) revealed open monomeric and dimeric conformations in equilibrium, thereby indicating that dimers are also present in non-crystallographic conditions. The analysis of these complexes and of the domains in solution also allowed us to clarify the molecular basis for the dimer formation, which is dependent on the loop1 length and sequence. In fact, after swapping the loop1 sequence of Smad5 by that of Smad3, we shifted the monomer/dimer equilibrium towards a predominantly monomeric domain in crystals bound to DNA. Perhaps, the functional differences between Smads are independent from their capacity to interact with specific GC or GTCT motifs. Instead, these differences could have arisen from the dimerization propensity observed in Smad1/5/8 proteins, absent in Smad2/3 and Smad4.

In the context of full-length proteins, R-Smads form heterotrimers with other R-Smads or with Smad4 via contacts between their MH2 domains. Since not all possible R-Smads and Smad4 combinations are found in cellular experiments [19,20], it is very tempting to suggest that other protein parts like the MH1 domains —and their capacity to form monomers or dimers— help define the selection of the Smad components for a given ternary complex.

Results

Smad5 and Smad8 complexes with the 5GC site

We first examined how Smad5/8 recognize 5GC-sites and if the recognition mode is similar to that described for other Smads [13,16] or to the BRE-GC interaction [17]. To test these hypotheses, we determined the structures of the Smad5 and Smad8 MH1 domains bound to the non-compressed 5GC GGCGC motif also using X-ray crystallography. For the constructs, we used the domain boundaries described in the Smad1/GTCT complex [16], which lack the first 10 protein residues since these residues were included in the Smad5/GTCT complex but were disordered [17]. In both cases, the GTCT-bound Smad1/5 proteins were oriented as dimers and the interaction with SBE was unaffected by the presence or absence of these additional residues.

Before setting up the crystallization screenings, we study the protein-DNA interactions by EMSA assays (Supplementary Figure S1C) and observed that the interaction with 5GC motifs is in the same range of concentration as that observed for the SBE sequence. The best diffracting crystals were obtained with a 16bp dsDNA TGCAGGCGCGCCTGCA containing the 5GC sequence (underlined). These crystals diffracted at 2.31 Å and at 2.46 Å resolution for Smad5 and Smad8 MH1 domains, respectively. We solved the Smad5/5GC complex by molecular replacement using a model derived from the Smad1/GTCT complex (PDB: 3KMP) and then used the Smad5 complex to refine that of Smad8 bound to the same DNA. In both complexes, the asymmetric unit (ASU, space group $P2_12_12_1$) contained a dimer of Smad MH1 domains bound to one 5GC site (Smad5 shown in gold and gray Figure 1A and Smad8 in violet and gray Figure 1B), with the α1 helix being swapped between monomers. Crystallization conditions, data collection and statistics are shown in Table 1.

The electron densities for the Smad5 and Smad8 proteins and the bound DNA, are well defined (Supplementary Figures S1D,E,F). Smad5 and Smad8 MH1 structures display all the characteristic features of MH1 domains [5,13]. They are composed of four

helices (arranged as a four-helical bundle) and three anti-parallel pairs of short strands (β1-β5, β2-β3, and β4-β6) and the fold is stabilized by the presence of a Zn^{2+} tightly coordinated by three cysteines and one histidine.

The protein-DNA binding region comprises the loop following the β1 strand, and the β2–β3 hairpin (residues 70–83, highlighted in blue, Figure 1C). 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 network of hydrogen bonds (HBs) with the first four consecutive base pairs of the GGCGCg motif (Figure 1C). The complex is further reinforced by a set of HB interactions between Ser79, Leu72, Gln77, (backbone atoms) and His101 and His102 (side chain) with G8', G10' and C3 bases (Figure 1C, middle and right). There is also a set of nine well-ordered water molecules bound at the interface of the protein-DNA-binding site that contribute to the stability of the complex (Supplementary Table S1). Similar interactions are observed for the Smad8-5GC complex (Figure 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 Smad1 bound to the GTCT site (3KMP, 123 aligned residues, Cα RMSD of 0.30 Å, Supplementary Table S2). The observed contacts are collected as a cartoon in Figure 1E showing that one bound MH1 domain covers the 3-CAGGCGC-9 area.

Overall these results show that Smad5/8 MH1 domain dimers are observed in complexes with 5GC motifs as well as with the SBE site previously characterized and corroborate that dimers seem to be a characteristic of these MH1-DNA complexes. Given the sequence conservation at the MH1 dimer interface, we predict that homo- and hetero-dimers of BMP-activated Smads can also occur in a cellular context.

SBE/5GC sites: One binding mode for all Smads

With the exception of the monomer/dimer arrangement of the MH1 domains, the protein-DNA binding interface of the Smad5 and Smad8 complexes is very similar to those of Smad4 and Smad3 bound to the same GGCGC motif (PDBs: 5MEY, 5OD6) (Figure 2A) [13]. The similarity of 5GC complexes corresponding to all R-Smads and to Smad4 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 topology of the major groove (the principal binding site of all complexes) is conserved between the different bound 5GC DNAs (Supplementary Figure S2A,B), as characterized using Curves [21]. This analysis reveals that the major

groove widths in these 5GC complexes are larger than in the complexes with the GTCT site, with all major groove depths being comparable.

These complexes also revealed that one MH1 domain is efficiently accommodated on one full DNA major groove with the clear distinction of minor and major grooves observed in all SBE and 5GC Smad complexes. This is in contrast to the arrangement observed in the compressed BRE-Smad5 complex [17] that shows a highly-distorted DNA due to the presence of two MH1 domains bound to six bp (Figures 2B and Supplementary Figure S2C,D). In this BRE complex, both minor and major DNA grooves show a similar depth and width due to the distortion introduced by the presence of two adjacent MH1 domains bound to the motif, an arrangement that could have perhaps been stabilized during the crystallization process.

Considering that *cis-regulatory* elements bound by Smad proteins often contain clusters of SBE, 5GC and BRE sites [13], we propose that the most probable binding mode used by Smad proteins to recognize DNA *in vivo* is that of the 5GC and SBE complexes. It seems very unlikely that two MH1 domains would interact with a compressed BRE site —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 (as shown in goldenrod in Figure 2B) using the full protein binding interface and perfect accommodation to the DNA major groove.

MH1 domain stability, a property independent from dimer or monomer propensities

In the literature, the presence of open/dimeric conformations of Smad1 has been related to the lower thermal stability of its MH1 domain with respect to R-Smads that populate closed-monomeric conformations [16]. To study how the structural differences correlate with protein stability, we performed thermal unfolding experiments of the recombinant MH1 domains of Smad3, Smad4, Smad5, and Smad8. Average values obtained from triplicates revealed that melting point values of Smad8 (54.5 ± 0.2ºC) are similar to those previously reported for Smad1, while the values of Smad3 and Smad5 are slightly higher and similar (57.8 ± 0.4ºC), with Smad4 being the most stable (65.1 ± 0.2ºC) (Figure 2C). Upon DNA binding, all MH1 domains are stabilized (2ºC in the case of Smad4 and up to 6ºC for the R-Smads).

The higher T_m of Smad5 (identical to Smad3 despite their distinct structural features) was surprising considering the high level of sequence/structural similarity of Smad5 with Smad1/8. Looking for an explanation of these results, we compared the

sequences of Smad1/5/8 and observed that there are only eight sequence variations in these MH1 domains, most of them conservative (none at the dimer interface, Figure 2D, Supplementary Figures S1B and S2E). The most different residue is Ile109 (loop 5), a position often occupied by hydrophobic residues in all Smad proteins including Smad5, but a Cys in Smad1 and Smad8. Introducing this mutation in Smad5 (Ile109Cys) caused a 3 ºC decrease in the melting temperature, a T_m value close to that of native Smad1 and Smad8 domains. This residue (Ile109Cys) precedes Cys110, which is required for Zn coordination and perhaps the presence of two consecutive Cys residues might compete for metal coordination and facilitate some misfolding processes, as detected by the thermal denaturation experiments (Figure 2C).

All in all, these observations suggest that small differences in the domain sequence (such as Ile109 to Cys) might have a more prominent role in the stability of MH1 domains than dimer/monomer propensities only.

Smad5/8 MH1 domains display a dimer/monomer equilibrium in solution and in the absence of DNA

N-terminus helix-swapped dimers have been considered a crystallization induced state potentiated upon DNA binding. We hypothesized that if the Smad1/5_8 MH1 domains can associate as dimers in crystals, is because there might be an intrinsic flexibility solution of these dimeric domains, absent in the monomeric Smad4 and Smad2 previously characterized [13,22]. This flexibility will allow the presence of an ensemble of conformations, which can facilitate the interchange of structural elements between monomers and dimers. To clarify this hypothesis, we sought to analyze the different species in solution, using Small-Angle X-ray scattering (SAXS) and Nuclear Magnetic Resonance (NMR). We also studied the different species using Ion mobility coupled to Mass spectrometry.

The SAXS data obtained for both Smad5 and Smad8 indicated an interval of radius of gyration of 17.6-19.3 Å and 16.7-18.6 Å respectively, which is between 15.8 Å expected for a compact monomeric form and 23.6 Å for the fully formed dimer (Figure 3A). Therefore, in order to fit the experimental data accurately, we incorporated dimeric models, together with open and closed conformations of monomeric domains (Figure 3B). The analysis indicated that the best fit was consistent with an equilibrium of open monomeric particles and dimers, and that the closed and compact monomeric conformations were not abundant in solution (Figure 3C).

The presence of several conformations in solution was further analyzed by NMR. To this end, we acquired backbone triple resonance experiments for the Smad5 MH1 domain. The analysis of the carbon chemical shift (CCS) values allowed us to identify all elements of secondary structure. However, the CCS values corresponding to the first helix indicate that, in solution and in the absence of DNA, this helix is slightly shorter and more flexible than expected for a compact MH1 domain structure (Figure 3D).

Furthermore, we also measured longitudinal, transverse relaxation times (*T1* and *T2*), and heteronuclear ^{1}H-^{15}N-nuclear Overhauser effect (hetNOE) to characterize the relaxation properties of the domain. Similar *T1* and *T2* experiments were acquired for Smad3 for comparison. Weak signals as well as lower heteronuclear NOE values were also observed from some of the residues located at the end of the first α1 helix and in loop1, supporting the presence of several conformations. We also obtained average correlation times (τ_c) of 12.1 ns and of 13.5 ns at 850 and 600 MHz respectively in contrast to 9.8 ns for both Smad3 and Smad4 MH1 domains [13]. Whereas values of 10 ns correspond to 15-kDa compact structures tumbling as monomers, we interpret the values obtained for Smad5 as an indication of the presence of larger conformations including monomers and dimers in equilibrium.

Finally, Smad5 dimers were also detected in the gas phase by ion mobility mass spectrometry (IM-MS). The analyses of m/z and drift time values revealed that Smad5 can populate monomeric, dimeric, and even larger oligomeric conformations. All these distinct species are resolved at different drift times (Figure 3E, Supplementary Figure S3A-C), with monomer and dimer forms (more compact) traveling faster than oligomeric ones. It is worth noting that, under the same conditions, the most abundant MH1 species of Smad3 (Figure 3F, Supplementary Figure S3D) and Smad4 (Supplementary Figure S3E) were monomeric forms, while signals arising from the dimer conformations were detected as minority species. Attempts to measure Smad8 samples were unsuccessful under these experimental conditions since these samples lost the bound Zn^{2+}after buffer exchange.

Altogether, these results indicate that the different studies of these proteins in non-crystallographic conditions are consistent with a high degree of plasticity of the MH1 domains of Smad5/8 present as dimers in equilibrium with monomers mostly populating open —and not closed— conformations. These results indicate that the monomeric or dimeric properties exhibited in the crystal structures of the different Smad/DNA complexes are also displayed by the domains in solution and in the absence of DNA.

Swapping the loop1 sequence between Smads converts MH1 dimers into monomers

Our findings regarding the presence of dimers in solution challenge the previous hypothesis that crystallization conditions enriched an atypical domain-swap conformation [16]. Our results also point to the first helix and the loop1, which are swapped in the dimeric conformation, as the most flexible areas in the monomers (Figure 3D). When structures of Smad2-5 MH1 domains are superimposed, we observed the variable lengths of loop1 and of the α2 helix (Figure 4A). In fact, Smad proteins are highly conserved in metazoans and in general, the differences are more marked with respect to Smad4 than among the five R-Smads [4] (Supplementary Figure S4A). As highlighted in the figure, BMP Smads have four residues in loop1, whereas the same loop has six residues in Smad3 and in Smad4, and sixteen in Smad2 [22]. Smad2/3 long loops can bridge the distance to accommodate the α1 helix packed to the same monomer, even for the α2 helix, which is one turn longer than in Smad4 (Supplementary Figure S4A). In the case of Smad3 (PDB 5OD6), its loop1 shows internal hydrogen bond contacts that help stabilize the turn, favoring the intramolecular packing of the first two helices (Figure 4B) [16]. These hydrogen bonds are not essential since they are absent in the monomers of Smad2 [22] and in the Smad4 structures (PDB 5MEY), Supplementary Figure S4B).

In Smad1/5/8, the combination of a long α2 helix (as long as in Smad2/3) with a short loop1 (shorter than in Smad2/3/4) explains the difficulties of these BMP activated Smads to maintain a compact and monomeric structure (Figure 4C). Instead, they select dimers (Figure 4D, Supplementary Figure SB) and ensembles of partially unfolded and flexible monomer conformations as we detected by NMR and SAXS data in solution (Figure 3B). Actually, when we extended the loop1 of Smad5 by introducing several Gly residues or by swapping the loop1 sequence with that of Smad3, we detected monomers as the main species in native mass spectrometry (MS) for these two mutant constructs (Supplementary Figure S3F).

Prompted by the MS results, we set to determine whether the crystallization process would turn these mutant constructs to select dimeric properties in crystals yet again. We found that both mutants are now arranged as monomers in the crystals, with the same packing interface for helices 1 and 2 as in the dimer but with both helices belonging to the same monomer as in Smad2/3/4. In the Smad5_3 chimera, the loop1 sequence is well defined in the electron density and OMIT maps (Figure 4D-E, Supplementary Figure S4C-D) whereas in the mutant containing the Gly insertion,

the loop could not be fully traced in the electron density map (Supplementary Figure S4E). With respect to the DNA recognition, the binding mode is conserved, although the Smad5_3 crystal structure contains a single protein bound to the DNA molecule. With this arrangement, only one 5GC site is occupied per DNA, yielding a protein-DNA complex in which five consecutive bp of the 5GC motif are specifically recognized by Smad5 side chains (Figure 4F).

In summary, these structures confirmed that the propensity to form dimers or monomers is encoded by the loop1 sequence, which has been conserved during metazoan evolution, and not artifacts enhanced by the crystallographic conditions.

Discussion

In eukaryotes, the association between transcription factors (TFs) to form homo- and hetero-dimers is a common feature employed by many TF families [23-25]. This capacity of association has implications in the regulation of specific cellular responses, in the stability of the proteins, in the optimal selection of DNA binding sites and in determining overall affinity. Smad proteins also follow this rule and associate with other Smads and with cofactors, ensuring the efficient interaction of these complexes with *cis*-regulatory elements genome-wide [5,13,26].

Comparison of the Smad5/8 complexes determined here to those previously characterized for Smad2/3 and Smad4 indicates that all R-Smads and Smad4 are able to interact specifically with 5GC and SBE sites by means of a conserved binding mode, mostly using the β2-β3 hairpin. Only the long isoform of Smad2 shows additional contacts from residues in the E3 insert, exclusively present in this specific isoform [22]. Indeed, the main difference observed in all MH1-DNA complexes is not in the recognition of DNA, but in the MH1 domain itself (monomer/dimer propensities). Whereas TGFβ-activated Smads and Smad4 interact with DNA as MH1 monomers, BMP-activated Smads form MH1 dimers by swapping the α1 helix between two monomers (Figure 5A).

R-Smad-Smad4 heterotrimers formation has been observed with over-expressed as well as with endogenous full-length proteins [27,28]. It is well accepted that Smads associations are driven *via* direct contacts of the conserved MH2 domains of R-Smads and a single Smad4 protein, as shown in the crystal structures of various complexes of MH2 domains [4,5,29,30]. However, despite the high level of MH2 domain conservation among different R-Smads, not all variants of Smads complexes seem possible or biologically supported, as for instance, a BMP-Smad/TGFβ-Smad/Smad4 complex or a complex made of two TGFβ-Smads and one BMP-Smad have never been

experimentally detected [20,31], suggesting that a second layer of selection might exist to favor the composition of some complexes over others.

One of these selection rules could include holding one dimer of Smad1/5/8 MH1 domains and one monomer of either Smad2/3/4 (whose MH1 domains cannot form dimers) thereby suggesting that these dimers may also occur in full-length Smad complexes, in native conditions. The formation of these complexes would be facilitated by the flexibility provided by the long linkers (80 residues) connecting the MH1 and MH2 domains and they should not compromise either, the formation of MH2 domain trimers as detected in crystals of isolated MH2 domains [5].

The monomer/dimer formation of MH1 domains would also have implications in the selection of several DNA sites recognized simultaneously by the complexes. In the case of Smad4 and BMP-activated Smad heterotrimeric complexes, the dimers of MH1 domains will select DNA sites separated by at least the distance of the two DNA binding sites in the dimer (a distance ≥ 60 Å) and with the Smad4 bound as close as 22 Å (the distance between two consecutive DNA sites that allows the interaction of two MH1 domains without steric hindrance, Figure 5B, 5C). Furthermore, BMP-activated Smad MH1 dimers can only bind to DNA strands with a certain relative orientation imposed by the dimer structure. For instance, we found that the Smad5 MH1 dimer docks well to a linker histone-depleted nucleosome structure (Figure 5B), with the minimal distance between two potential Smad1/5 binding sites being equal to 80-90 bp or larger (hundreds of bp) if the sites are located within nucleosome-depleted DNA or in linkers (Figure 5C). In contrast, Smad4 complexes with TGFβ-activated Smads (all MH1 domains monomers) can bind adjacent motifs, thereby allowing a much more compact interaction with DNA (Schematic representation, Figure 5C). These distinct structural features are in agreement with ChIP-Seq peaks where Smad binding motifs are distributed as clusters. Remarkably, clusters recognized by BMP-activated Smads contain few adjacent motifs whereas the clusters often contain several consecutive sites in TGFβ-activated ones [13].

Overall, our results reveal two new hypotheses for the function of Smad complexes *in vivo*. First, the composition of Smad complexes can be modulated by the association through dimers *via* the BMP-activated Smad MH1 domains, and not only through MH2 domain interactions. We propose that this feature has been among the keys to shaping two classes of R-Smad proteins since the origin of metazoans. The second hypothesis is related to DNA recognition. Although all Smad proteins can interact with the same DNA motifs, finding the optimal DNA sites for a given Smad trimer must fulfill certain specific spatial requirements. This implies that not all theoretically available DNA motifs

can be recognized by a given complex and also, that promoters whose Smad binding sites are separated by variable distances can be recognized by a given Smad complex. This versatility can explain how so many different regions in the genome are bound to Smad complexes, as reported in the literature [32,33].

All findings available till now suggest that the selection of optimal DNA targets in a native context is the result of a collaborative approach between the different Smad complexes and bound cofactors. This selection process seems to be also modulated by the internal association of Smad proteins, where all components fit in order to fine-tune the context-dependent action of BMP and TGFβ signals. Certainly, additional experiments, as well as structures of full-length Smad complexes bound to DNA, will finally illustrate how these different layers of interactions are defined and modulated.

Acknowledgments

We thank Dr. N. Berrow (IRB Barcelona) for some DNA constructs, The Automated Crystallography Platform staff (IRB Barcelona-CSIC) and 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 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 800 MHz. We also thank J. Cordero for some preliminary experiments.

T.G. is a PhD student funded by a Severo Ochoa and by the BBVA. Z.K., R.F, R.P., and B.B are co-funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie COFUND actions of 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 Bio-SAXS 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.

Methods

Detailed methods on protein expression and purification, Differential Scanning Fluorometry, NMR experiments and analysis, crystallography experiments, TWIM-MS experiments, and SAXS data acquisition and analysis can be found in SI Appendix, Methods. The atomic coordinates have been deposited in the Protein Data Bank, Small-angle scattering data and models in the SASBDB database, and NMR assignments and chemical shifts in the Biological Magnetic Resonance Data Bank (BMRB).

Table 1 Data collection and refinement statistics

GGCGC complex	SMAD5 6FZS	SMAD8 6FZT	SMAD5_3 6TBZ	SMAD5_gly 6TCE
Data collection	ESRF-ID30a3	ESRF-ID30a3	ALBA-BL13	ALBA-BL13
Space group	$P2_12_12_1$	$P2_12_12_1$	$P3_2$	$I4_122$
Cell dimensions				
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
α, β, γ(°)	90,90,90	90,90,90	90 90 120	90 90 90
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)
R_{meas}	0.130 (1.413)[a]	0.097 (1.403)[a]	0.087 (0.885)[a]	0.049 (3.208)[a]
$I/\sigma(I)$	11.4 (1.5)[a]	14.3 (1.3)[a]	10.9 (1.9)[a]	21.1 (0.6)[a]
$CC_{1/2}$	99.8 (52.9)[a]	99.8 (57.2)[a]	99.9 (57.8)[a]	100.0 (31.5)[a]
Spherical completeness (%)	83.0 (22.0)[a]	79.7 (22.4)[a]	48.2 (7.3)[a]	90.1 (34.6)[a]
Ellipsoidal completeness (%)	92.4 (44.6)	92.0 (54.3)	88.8 (68.2)[a]	91.9 (40.6)
Redundancy	9.3 (9.9)[a]	6.0 (6.3)[a]	5.0 (3.9)[a]	7.9 (7.9)[a]
Refinement				
Resolution (Å)	29.77-2.31	59.11-2.46	46.00-1.78	53.11-2.92
No. reflections	16665	15793	11400	4823
R_{work} / R_{free}	0.193/0.241	0.193/0.237	0.193/0.226	0.210/0.253
No. atoms	2815	2852	1702	1167
Protein	2010	2057	988	842
DNA	656	656	656	324
Zinc ions	2	2	1	1
Water	143	123	57	0
B factors				
Protein	48.27	58.72	51.29	141.57
DNA	88.88	116.19	85.68	116.17
Zinc ions	40.51	53.18	27.44	156.80
Water	48.06	55.94	35.78	NA
R.m.s. deviations				
Bond lengths (Å)	0.010	0.010	0.008	0.010
Bond angles (°)	1.01	1.04	0.92	0.97

* anisotropic correction, (spherical and ellipsoidal)

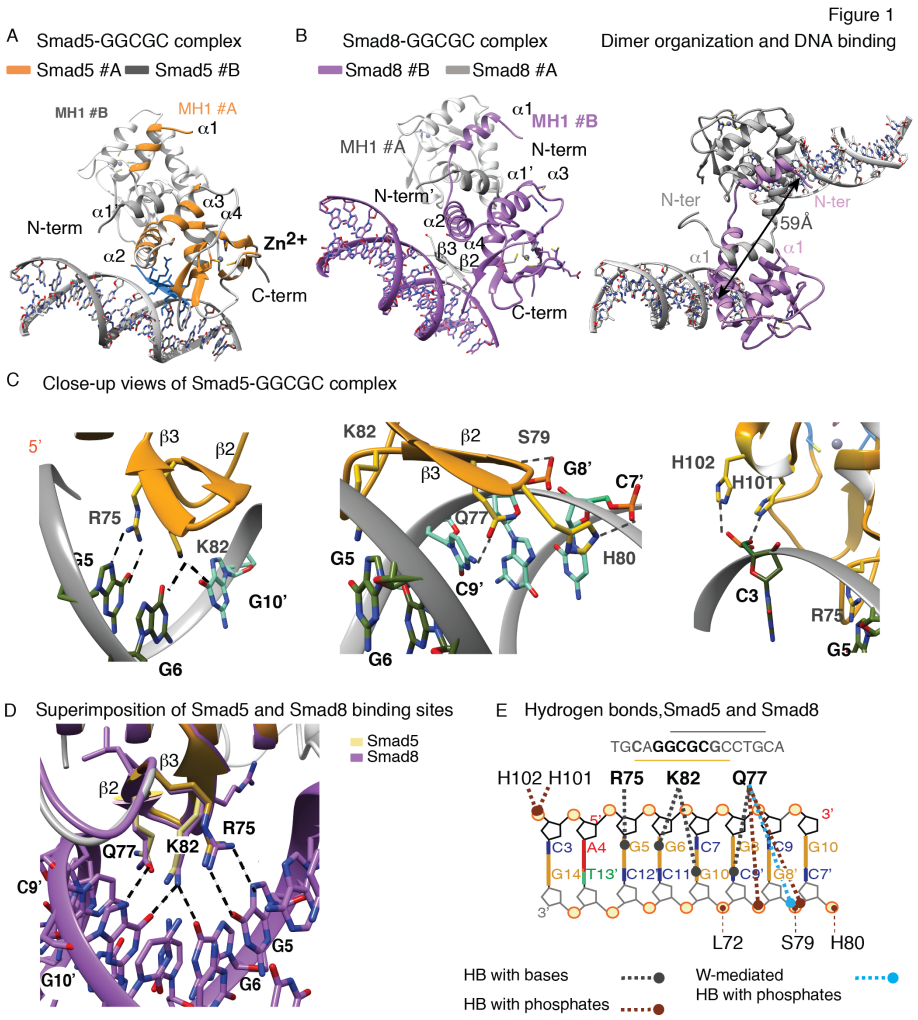
[a]Values in parentheses are for the highest resolution shell.

Ramachandran outliers: 0% (Smad5, Smad8, Smad5_3) 1.7% for Smad5_gly.

Ramachandran preferred/allowed: 98.39/1.61 Smad5 and 96.84/3.16% Smad8, 97.3%/2.7% for Smad5_3ch, and 88.2%/10.1% for Smad5-gly.

Figures and figure legends

Figure 1. Structures of Smad5 and Smad8 bound to the GGCGC site



A. Ribbon diagram of the Smad5 MH1 domain homodimers in complex with the GGCGC DNA motif (ASU). Monomers are shown in orange and gray respectively. The elements of secondary structure and the Zn (gray sphere) binding site are indicated.

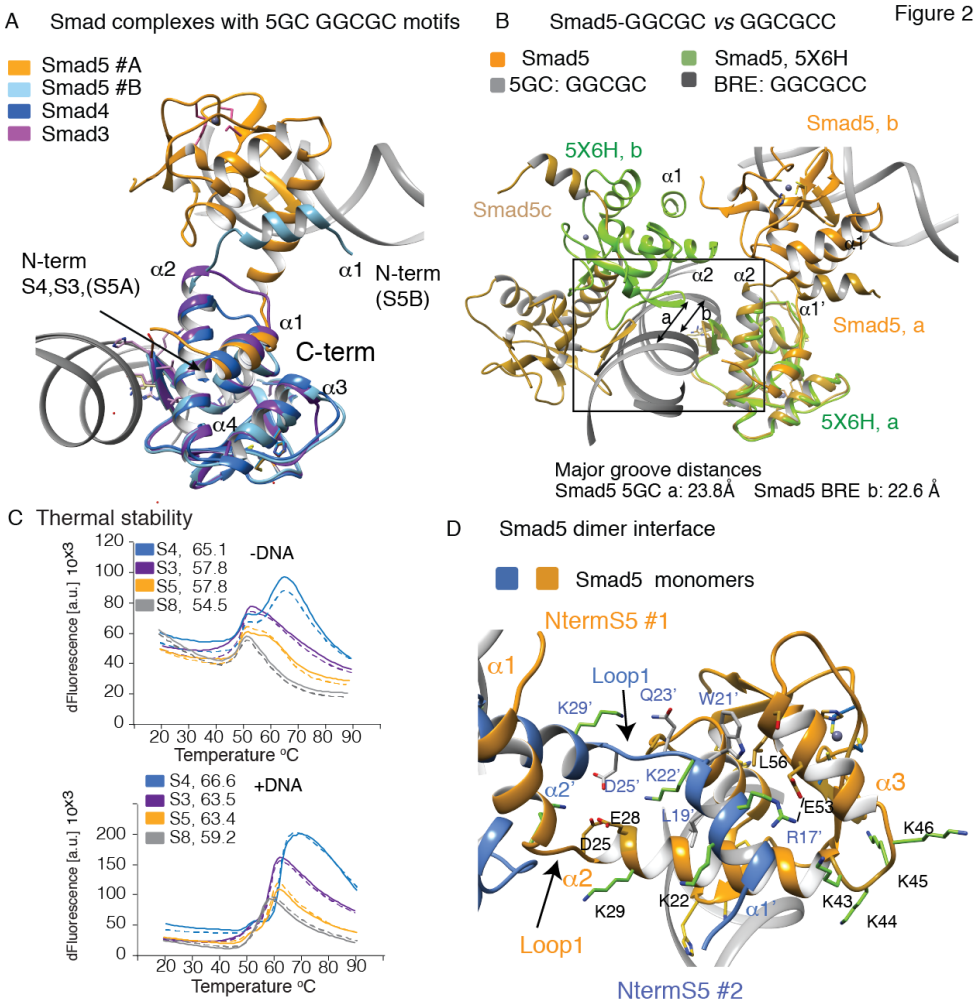
B. Ribbon diagram of the homodimeric Smad8 MH1 in complex with the GGCGC DNA motif (ASU). Monomers are shown in purple and gray, respectively. Each monomer is bound to a different DNA, the distance between DNA binding sites is indicated.

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 bases involved in the interaction are labeled.

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.

Figure 2: Structural comparison of different Smads bound to DNA



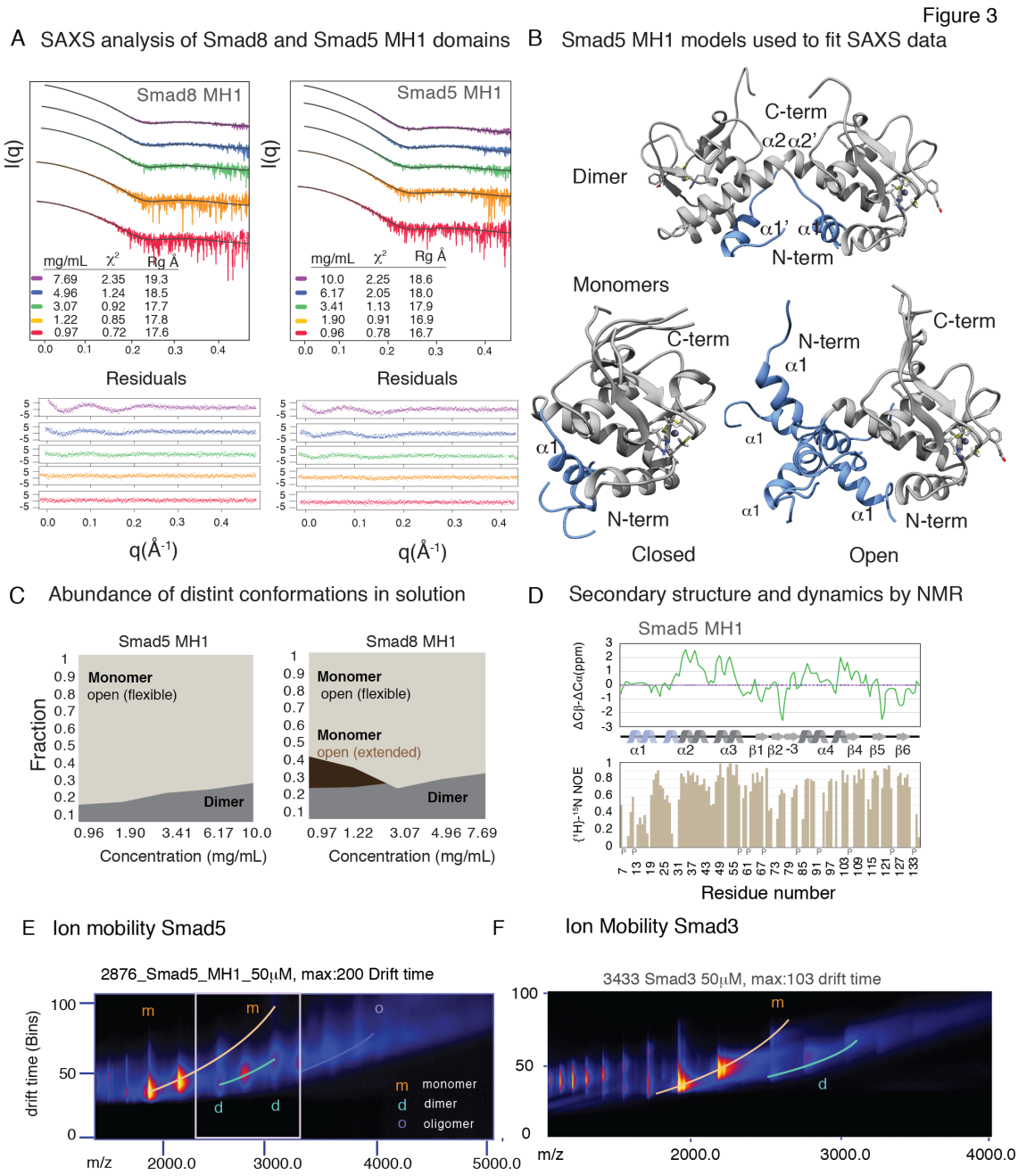
A. Overlay of the complexes of Smad5 MH1 (dimer, gold and light blue) and the monomers corresponding to Smad4 (royal blue) and Smad3 (purple) complexes. All backbones are shown as ribbons. Some secondary structural elements are labeled. N-term of all proteins are indicated.

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 [16] although the electron-density map for this region is not well defined.

C. Thermal denaturation curves (duplicates) of different MH1 domains in the absence and in the presence of DNA (GGCGC motif). Melting temperatures are indicated in °C.

D. Smad5 residues located at the dimer interface are labeled in blue and black. These residues are also conserved in Smad1 and Smad8.

Figure 3: Monomer/dimer equilibrium of Smad5 and Smad8 in solution



A. Small-angle X-ray scattering (SAXS) analysis of the Smad5 and Smad8 MH1 domains at five protein concentrations. Experimental and graphical output of the best fit shown for each concentration.

B. Structural models used for the fitting of the SAXS data. Models were generated from the X-ray structures of Smad5 and Smad8 determined in this work. The closed monomeric structure was generated using the Smad4 structure (PDB: 3QSV) as a template.

C. Relative abundance of the different conformers in solution. Only open monomers and dimers satisfied the experimental data.

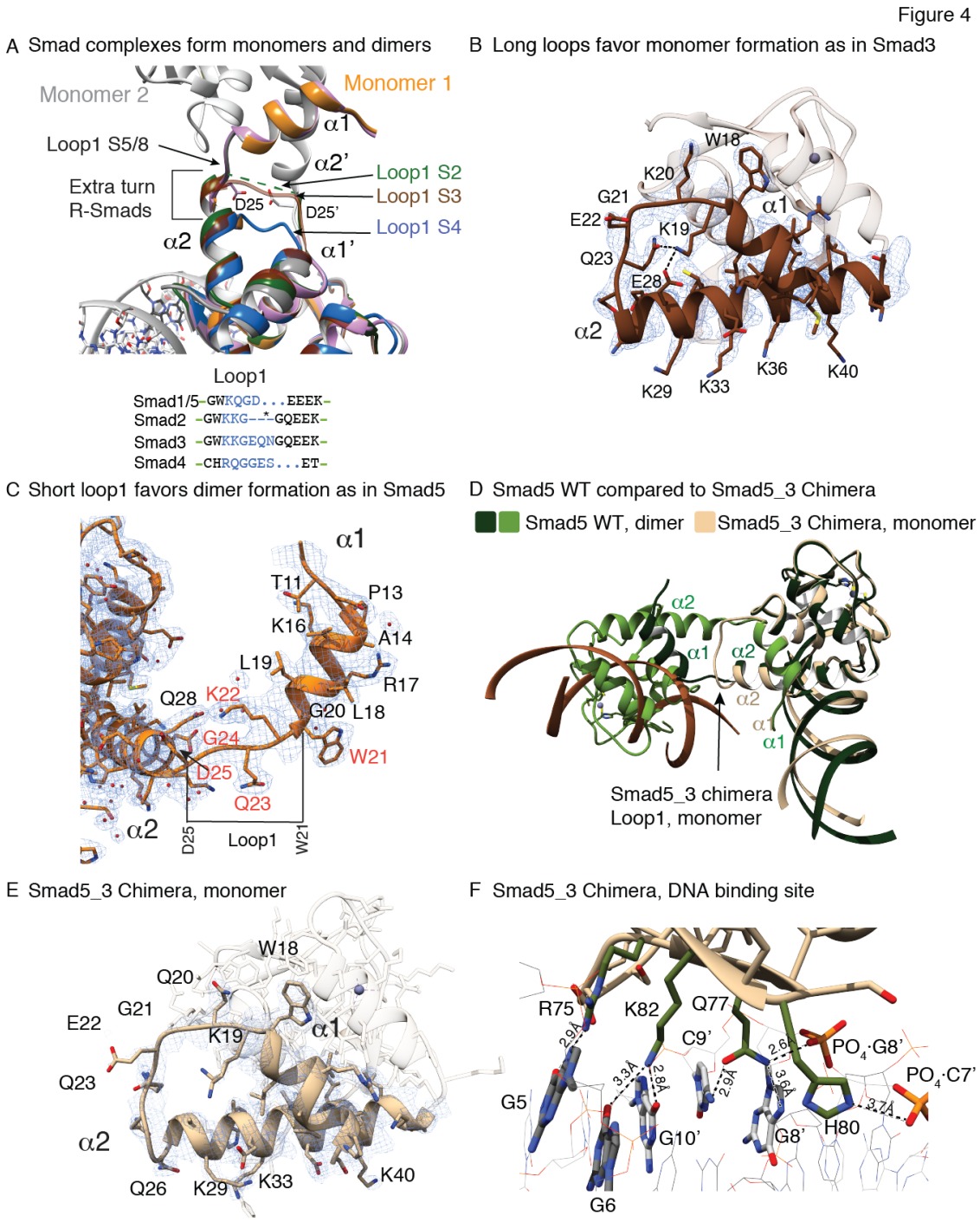
D. NMR analysis of the Smad5 MH1 domain. Secondary structure elements based on experimental ^{13}Cα-^{13}Cβ chemical shift differences. Positive and negative values

indicate α helices and β sheet structures respectively. The differences for the first helix are close to zero, an indication of variable conformations. Below, hetNOE values at 800 MHz ^{1}H Larmor frequency ordered according to the residue number. Proline (P) residues indicated along the sequence.

E. Plot of the mobility drift time versus *m/z* for the Smad5 MH1 domain (aqueous solution, 200 mM ammonium acetate buffer) showing the multidimensional data, (instrumental settings are provided in Supplementary Table S3). The mass spectrum from all ions shown in Supplementary Figure S3A. Peaks corresponding to monomers (m) and dimers (d) present in the gas phase are labeled.

F. Plot of the mobility drift time versus *m/z* for the Smad3 MH1 domain as in E. The mass spectrum from all ions shown in Supplementary Figure S3D.

Figure 4: Structural features defining dimer/monomer propensities



A. Close-up of loop1 in the different structures. Differences in the loop1 and in the second helix, which is longer in R-Smads than in Smad4, are indicated with arrows and as a bracket, respectively. The positions of D25 for both monomers (the last residue of the α1 helix) and their separation in the dimer are indicated in Å, to highlight the difficulty of Smad1/5/8 to bridge this distance with a short loop.

B. A section of the electron-density map (loop1) contoured at 1.0σ corresponding to the Smad3 monomer (pdb: 5OD6). Some side chains are labeled.

C. A section of the electron-density map for the Smad5 MH1 dimer. To indicate the contacts at the protein dimer interface, some elements of the secondary structure and selected residues are labeled surrounding the α1 helix.

D. Superposition of Smad5 dimer and monomer (green and beige). The loop1 in the monomer (chimeric construct) and the dimer interface (Smad5 WT) are indicated with arrows.

E. A section of the electron-density map for the Smad5_3 chimera showing the loop1 orientation and the monomer arrangement. Some elements of the secondary structure and selected residues are labeled surrounding the α1 helix.

Close-up view of the Smad5_3 chimera bound to DNA. Residues located at the protein-DNA binding site are shown in light gray. Hydrogen bonds are indicated as dotted lines.

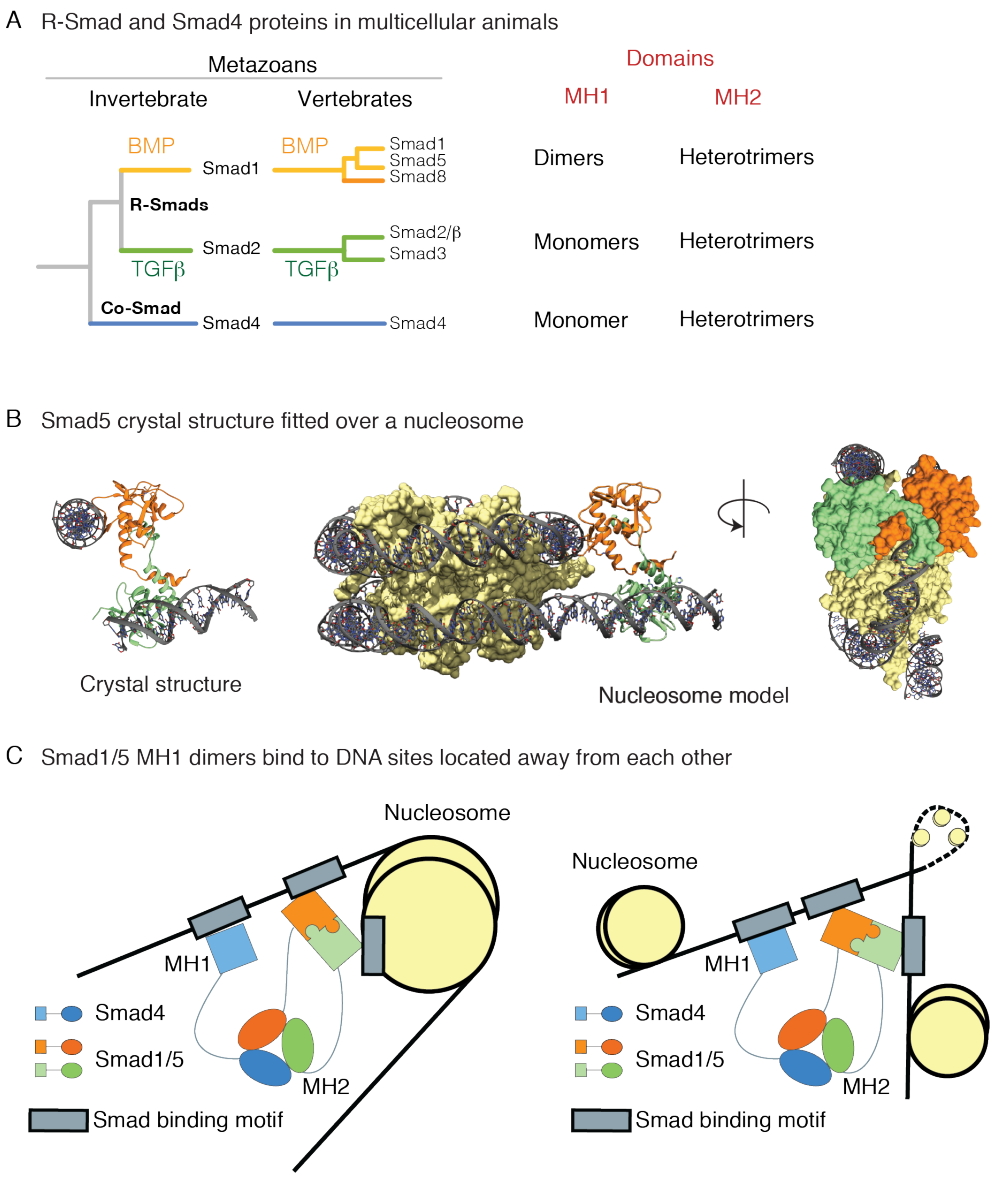


Figure 5: Structure-based models of Smad heterotrimers interacting with DNA clusters

A. Schematic representation of the different Smad proteins in metazoans, including the duplications observed in vertebrates, adapted from [4]. For each R-Smad protein we indicated the propensity of their MH1 domains to fold as monomers and dimers and of their MH2 domains to associate as heterotrimers with Smad4.

B. Model of MH1 Smad5 dimer – nucleosome complex. *Left:* Crystal structure of MH1 Smad5 dimer (orange and green) bound to two distant GGCGC sites (6FZS). *Middle/right:* MH1 Smad5 dimer docked manually on a nucleosome structure described in the literature [34] (6GEJ, in yellow and grey) with two Smad binding sites being approximately 85 bp away.

C. Simplified schematic representation of Smad1/5-Smad4 heterotrimers binding to SBE/5GC sites within chromatin. Monomeric MH1 Smad4 is free to bind any site within the spanning range of the Smad1/5-Smad4 heterotrimer, while MH1 Smad1/5 dimers can bind simultaneously only to DNA motifs that are either approx. 85 bp (*left*) or hundreds of bp away (*right*) from each other. In both scenarios the Smad1/5 DNA binding sites must be spatially separated by 50-60 Å. Conversely, MH1 Smad2/3 monomeric complexes of Smad2/3-Smad4 heterotrimers are not restricted for the location of their binding sites. Smad cofactors and other chromatin-interacting proteins are omitted for simplicity.

Supplemental Information

Methods

Protein production and cloning:

The Smad5 (Uniprot: Q99717-1, Ser9-Arg143), the Smad5_gly, Smad5_3 mutant construct and Smad8 (O15198-1, Thr14-Pro144) domains were cloned using synthesized DNA templates with optimized codons for Bacterial expression (Thermo Fisher Scientific) and confirmed by DNA sequencing (GATC Biotech). Proteins were expressed fused to an N-terminal His-tag followed by a TEV or 3C protease cleavage site in *E. coli* BL21(DE3) or C41(DE3) pLysS and purified following standard procedures [13]. Unlabeled and labeled samples were prepared using Luria Broth (LB) and minimal media (M9) cultures, respectively (Melford). D_2O (99.95%, Silantes), $^{15}NH_4Cl$ and/or D-[^{13}C] glucose (Cambridge Isotope Laboratories, Inc) were used to prepare the labeled samples [35]. Cells were cultured at 37°C to reach an OD_{600} range of 0.6-0.8. After induction with IPTG (final concentration of 0.4 mM) and overnight expression at 20°C, bacterial cultures were centrifuged and cells were lysed (EmulsiFlex-C5, Avestin) in the presence of Lysozyme and DNase I and in PBS buffer at pH 7.5. The soluble supernatants were purified by nickel affinity chromatography (HiTrap Chelating HP column, GE Healthcare Life Science) using an NGC™ Quest 10 Plus Chromatography System (Bio-Rad). Eluted proteins were digested with TEV or 3C proteases (at 4°C or room temperature respectively) and further purified by ion-exchange chromatography using a HiTrap SP HP and size-exclusion chromatography on a HiLoadTM Superdex 75 16/60 prep grade columns (GE Healthcare) equilibrated in 20 mM Tris-HCl buffer (pH 7.2), 80 mM NaCl and 2 mM TCEP.

Purified proteins were verified by Liquid chromatography-Mass Spectrometry (LC-MS) using an ACQUITY UPLC Binary Sol MGR LC system (Waters) equipped with a BioSuite Phenyl 1000Å column (Waters, 10 µm RPC 2.0x75 mm) at a flow rate of 100 µL/min. The column outlet was directly connected to the mass spectrometer, which acquired full MS scans (400-4000 m/z) working in positive polarity mode. Samples were eluted using a linear gradient from 2% to 5% B in 5 min and from 5% to 80% B in 60 min (A= 0.1% Formic Acid, FA, in water, B= 0.1% FA in CH3CN) and analyzed using the MassLynx™ Software, (V4.1.SCN704, Waters). The purity of the recombinant proteins was over 95%, as shown by the Mass Spec analysis.

Duplex DNAs

Duplex DNAs were annealed using complementary single-strand HPLC-purified DNAs. DNAs were mixed at equimolar concentrations (1 mM), heated at 90°C for 3 min and allowed to cool to room temperature for 2 h. DNAs (with and without fluorophores) were purchased at Biomers and/or at Metabion, Germany.

Differential scanning fluorometry

Experiments were performed in a StepOnePlus Real-time PCR System (Applied Biosystems) using 96-well plates (MicroAmp Fast 96-Well Reaction Plate, Applied Biosystems) and a 25 μL total volume (for each reaction). Melting curves were acquired in triplicates to determine the average melting temperature (Tm), and lysozyme and Smad4 MH1 domains were used as positive controls. All samples (0.5 mg/mL) were prepared in 50 mM Tris-HCl (pH 7.2), 150 mM NaCl and 2 mM TCEP. SYPRO Orange dye (Sigma-Aldrich) was used at 10X starting from a 5000X dilution. The plate was sealed with optical quality sealing film (Platemax, Axygen) and centrifuged at 2000*g* for 30 s. Samples were equilibrated for 60 s and analyzed using a linear gradient from 25ºC to 95ºC in increments of 1ºC/min, recording the SYPRO orange fluorescence throughout the gradient. Tm values were determined from the maximum value obtained from the first derivative of the sigmoid curve, using the Applied Biosystems® Protein Thermal Shift™ Software.

NMR chemical shift assignment and perturbation experiments

NMR data were recorded on a Bruker Avance III 600-MHz spectrometer equipped with a quadruple (^{1}H, ^{13}C, ^{15}N, ^{31}P) resonance cryogenic probe head and a z-pulse field gradient unit at 298 K. Backbone ^{1}H, ^{13}C and ^{15}N resonance assignments were obtained by analyzing the 3D HNCACB and HN(CO)CACB experiment pair (33). Experiments were acquired as Band-Selective Excitation Short-Transient-type experiments (BEST) with TROSY and Non-Uniform Sampling (NUS) [36,37]. This strategy allowed us to unambiguously assign 110 of the 121 possible amides (131 residues, 10 of them prolines). Comparison of the Smad5 MH1 (Cα and Cβ) chemical shifts to reference values, as well as the ^{15}N edited-NOESY data, corroborated the presence of bound Zn^{2+} and of four helices and six strands, characteristic of the MH1 fold. The strands are ordered as three anti-parallel pairs: $\beta_1\beta_5$, $\beta_2\beta_3$, and $\beta_4\beta_6$. The presence of many long-range interactions confirmed that, in the absence of DNA, the structure of the MH1 domain is well defined in solution. Chemical shifts have been deposited in the BMRB (entry 27548). For the titration experiments, HSQCs were recorded using a

Non-Uniform Sampling (NUS) acquisition strategy to reduce experimental time and increase resolution.

T1 and T2 relaxation measurements were acquired using standard pulse sequences [36]. The rotational correlation time of the Smad5 MH1 domain (τc) was calculated assuming slow molecular motion, τc larger than 0.5 ns, and only J(0) and J(ωN) spectral density terms contributed to the overall value.

$\tau_c \approx \frac{1}{4\pi\nu_N}\sqrt{(6\frac{T1}{T2}-7)}$, where ν_N is the ^{15}N resonance frequency.

Protein samples (400 μM for backbone experiments) were equilibrated in a 20-mM TRIS buffer containing 100 mM NaCl at pH7.2 supplemented with 10% D_2O. Spectra were processed with NMRPipe [38] and MddNMR [37], and analysis was performed with CARA [39].

X-ray

High-throughput crystallization screening and optimization experiments were performed at the HTX facility of the EMBL Grenoble Outstation [40]. Human Smad5 was concentrated to 5 mg/mL prior to the addition of the annealed DNAs (Metabion) dissolved in 20 mM Tris-HCl pH 7, 10 mM NaCl and 2 mM TCEP. The final molar protein DNA ratio was 1:1. Screenings and optimizations were prepared by mixing 100 nL of the complex solution and 100 nL reservoir solution in 96-well plates. Crystals of the complexes were grown by sitting-drop vapor diffusion at 4°C. Crystals were obtained with several DNAs of different lengths and in different conditions, and in all cases they were reproducible. Several datasets were acquired for the best diffracting crystals and were then analyzed. Final conditions for the three best diffracting complexes were optimized as follows:

Smad5: 0.2 M NaF, 0.1 M bis-tris propane pH 7.5, 20% PEG3350

Smad8: 0.2M NaF, 0.1 M bis-tris propane pH 8.5, 20% PEG3350

Smad5_3 chimera: 0.05 M sodium citrate pH 5.5, 22% PEG3350

Smad5_gly: 0.05 M Hepes pH 7.0, 21% PEG Smear Medium (PEG 2000, 3350, 4000, and 5000 MME).

All crystals were cryoprotected in mother liquid supplemented with glycerol and frozen in liquid nitrogen. Diffraction data for Smad5 and Smad8 complexes were recorded at the ESRF in Grenoble (France) (beamline ID30a3) and Smad5_3 chimera and Smad5_gly data at the ALBA Synchrotron Light Facility (BL13-XALOC beamline),

Barcelona, Spain. The data were processed, scaled and merged with autoPROC [41]. Initial phases were obtained by molecular replacement using PHASER [42,43] from the CCP4 and PHENIX suites [44,45] (search model PDB code: 3KMP) with anisotropic correction. REFMAC [46] phenix.refine [44] and BUSTER [47] were employed for the refinement, and COOT [48] for the manual improvement of the models. For Smad5 mutants, the PDB-REDO server was used for the selection of data resolution cutoff (paired-refinement) and for the structure model optimization [49]. Water molecules bound at the DNA-protein interface were selected when they participated in at least three hydrogen bonds (cutoff distance of 3.5 Å). Figures describing the structures were generated with UCSF Chimera [50].

TWIM-MS experiments

Experiments were performed using a Synapt G1 HDMS mass spectrometer (Waters UK Ltd., Manchester, UK). 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 [51] 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).

SAXS data

Data were collected on samples of Smad5 and Smad8 MH1 domains at protein concentrations ranging from 0.96 to 10 mg/mL for Smad5 and from 0.97 to 7.69 mg/mL for Smad8_9. All samples were concentrated in 20 mM Tris buffer, 150 mM NaCl, and 2 mM TCEP, pH 7.2. Data were acquired at Beamline 29 (BM29) at the European Synchrotron Radiation Facility (ESRF; Grenoble, France). Protein samples were centrifuged for 10 min at 10,000*g* prior to data acquisition. Experiments on BM29 were collected at an energy of 12.5 keV and data were recorded on a Pilatus 1M detector at 10ºC. For each sample and buffer, 10 exposure frames of 1 s were collected, and the

exposure set was combined during data reduction to produce each SAXS curve. Buffer subtraction was performed after data reduction. Image conversion to the 1D profile, data reduction, scaling and buffer subtraction were done by the software pipeline available at the BM29 beamline. Further processing was done with the ATSAS software suite and Scatter [52]. Guinier plot calculation (for the estimation of the radius of gyration Rg) was performed with PRIMUS, included in the ATSAS suite, using low q regions (qmax × Rg<1.3). Small-angle scattering data and models were deposited in the SASBDB database under entry codes SASDE32 and SASDE42 respectively. Regarding the possible conformations in solution, the following four states were considered plausible: the crystallographic dimer; the closed monomer present in other crystal structures of MH1 domains; and two open monomers where the extended N-term helix was considered either as in the crystal structure or flexible. Molecular modeling for the open and free structure was performed with the Rosetta modeling software suite, using the FloppyTail application [53] and starting from the MH1 crystal structures of Smad5 and Smad8_9 determined in this work (PDB: 6FZS and 6FZT respectively). For closed monomer models, we used Modeller [54] and the Smad4 MH1 structure (PDB: 3QSV), whereas the dimers were directly taken from the complexes of Smad5 and Smad8 (PDB: 6FZS and 6FZT). In all cases, DNA and water molecules were removed and secondary structure elements were restrained, except for the flexible N and C-terminal tails and the flexible N-term helix. For each state, one thousand conformers were simulated, in order to generate sufficient conformational sampling. Theoretical SAXS curves were calculated using CRYSOL [55] and compared to the experimental data, selecting the linear combination of the four states with the lowest chi-squared, as implemented in ATSAS.

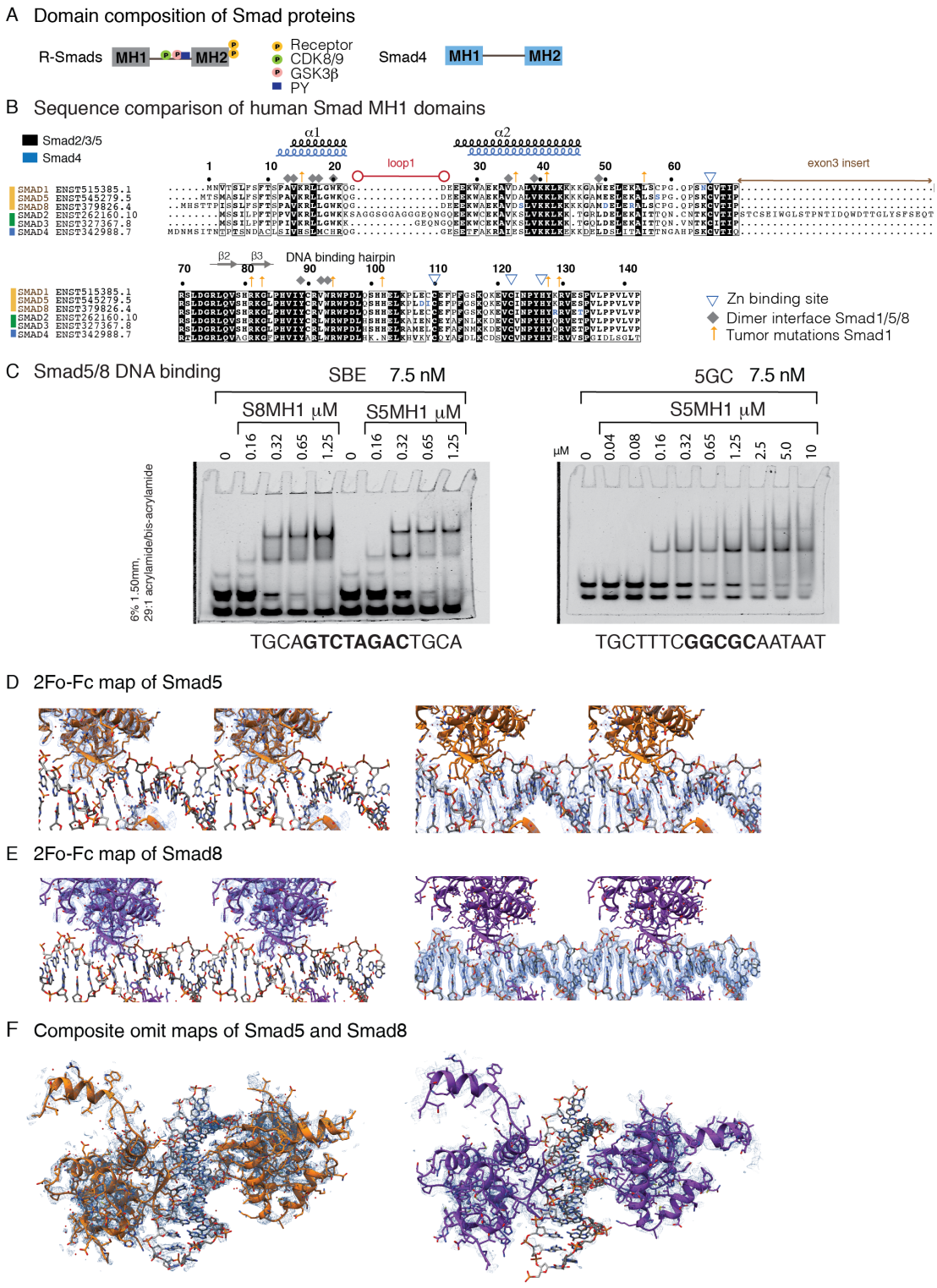
The chi-squared metric for N data points was calculated using the equation:

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\left[I^{calc}(q_i) - I^{obs}(q_i)\right]^2}{\sigma_i}$$

A Key Resources Table is included describing material and resources.

Supplementary Figures

Supplementary Figure S1



A. Domain composition of R-Smads and Smad4. Different phosphorylation sites of R-Smads are indicated as yellow (receptor), green (CDK8/9) and pink (GSK3) circles.

B. Sequence alignment of the R-Smads and the Co-Smad Smad4 MH1 domain (human proteins), where strictly conserved residues are highlighted. The four residues that coordinate Zn are indicated with a triangle. The residues that define the first two helices (as observed in the structures of R-Smads and in Smad4) and the DNA binding hairpin are represented on the top of the alignment. The variable length of the loop1 is indicated with a red line. Residues participating in the dimerization interface of BMP-activated Smads are indicated as gray diamonds. Tumor mutations described for Smad1 are indicated by arrows. The insert corresponding to the Smad2 long isoform is labeled. Alignments were generated with MAFFT [56] and BoxShade server.

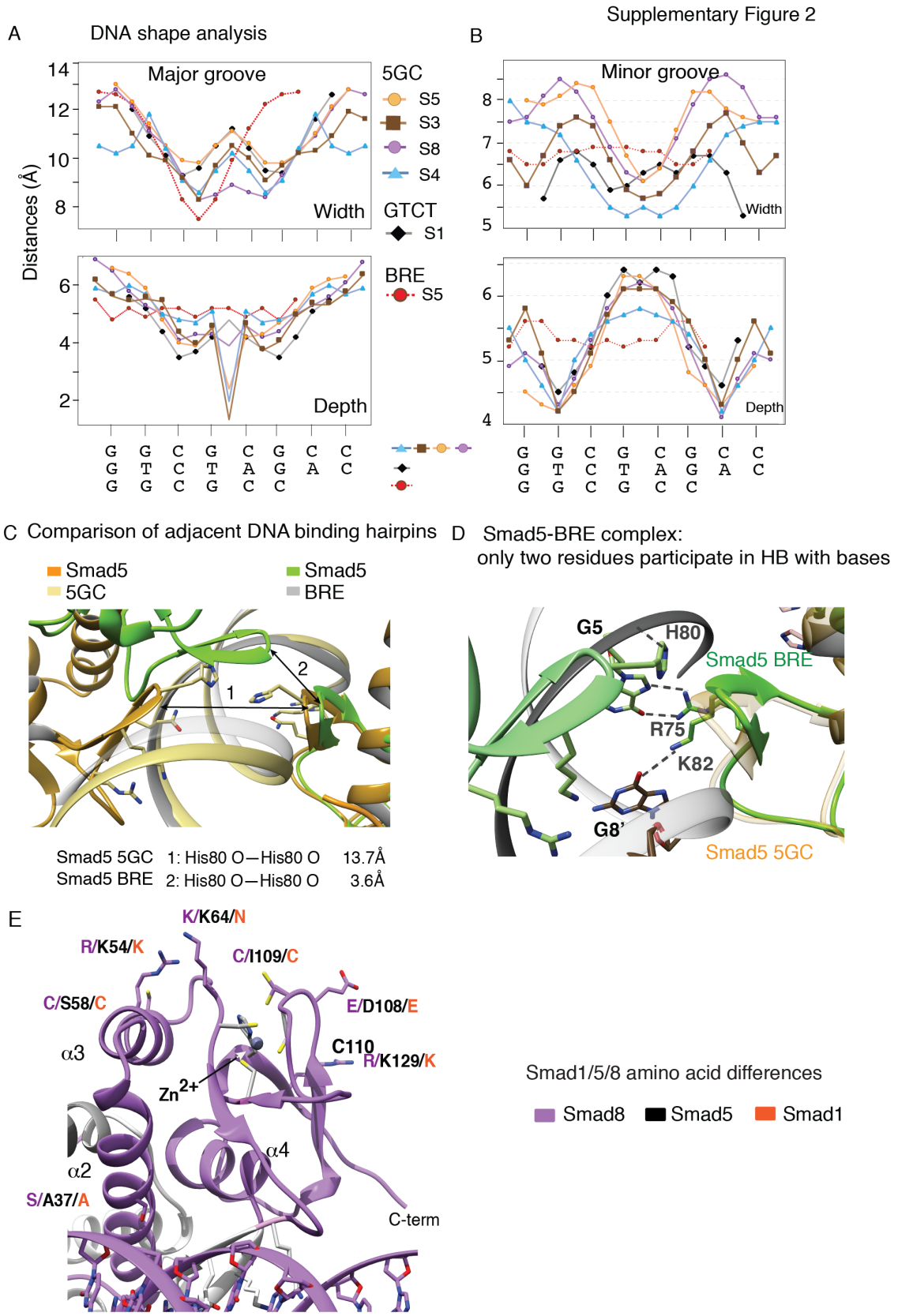
C. EMSA assays showing the interactions of Smad5 and Smad8 with the SBE sequence and the Smad5 interaction with the 5GC sequence. The binding affinity is similar for both sequences. DNA sequences are shown at the bottom, DNAs were at a 7.5nM fixed concentration and MH1 domain concentration for each lane is shown on top.

D. Section of the electron-density map (stereo-view) contoured at 1.0σ for the Smad5 MH1 complex. The map was calculated with coefficients 2Fo – Fc, and the refined model (ribbon) is superimposed. The density map is shown for the protein or for the DNA. Some residues and elements of secondary structure are indicated.

E. Section of the electron-density map (stereo-view) representing the protein-DNA complex of Smad8 as in panel A.

F. Composite OMIT maps for the Smad5 MH1 domain (left) and the Smad8 MH1 domain (right).

Supplementary Figure S2



A. Major groove width and depth analyses of the DNA bound to Smad5 (orange), Smad8 (lavender), Smad3 (brown), Smad4 (blue), Smad1 (black) and Smad5-BRE complex (red). The values were calculated using Curves+ [21].

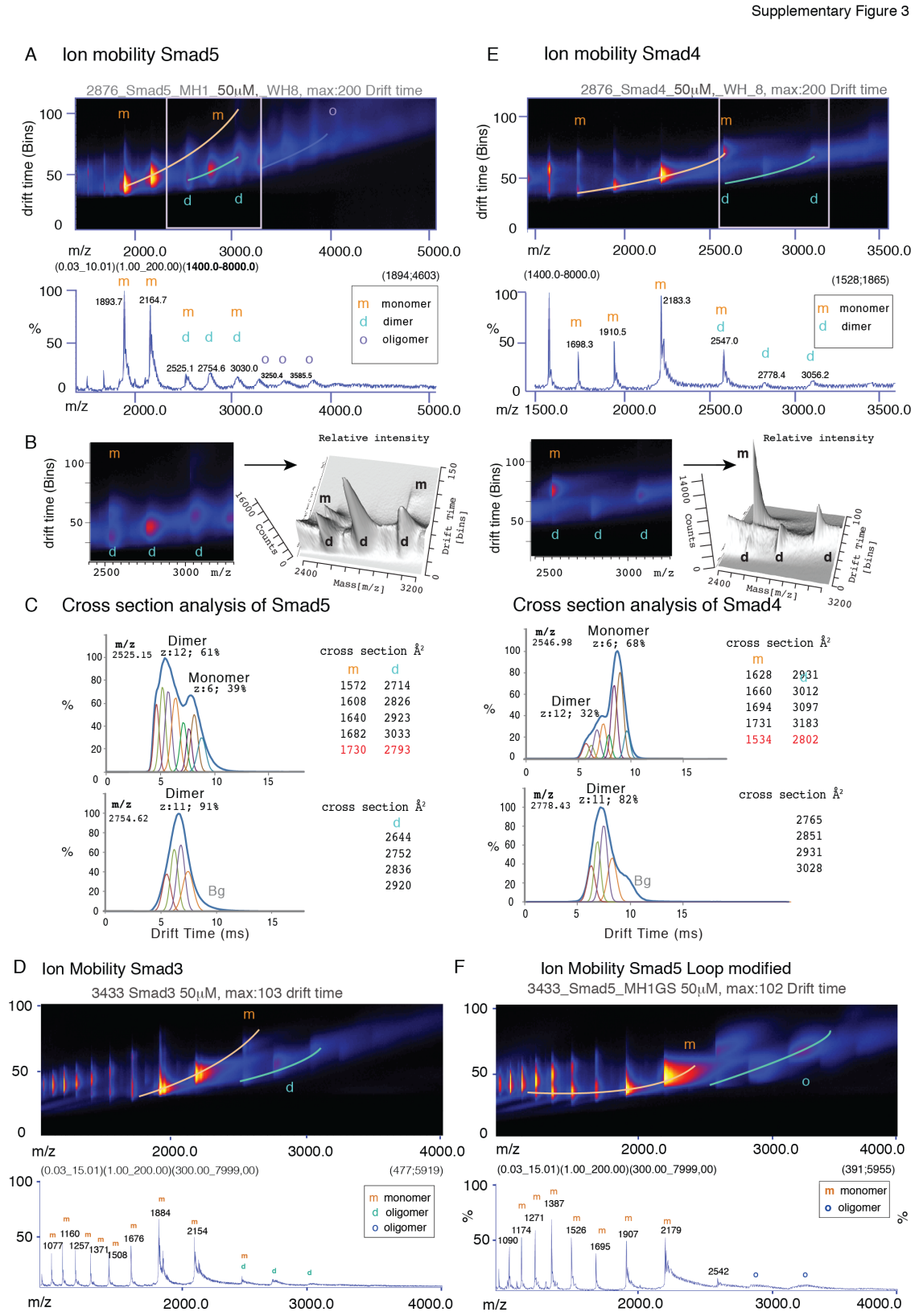
B. Minor groove width and depth analyses of the DNA bound to Smad5 (orange), Smad8 (lavender), Smad3 (brown), Smad4 (blue), Smad1 (black) and Smad5-BRE complex (red). The values were calculated using Curves+ [21].

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 the interactions of R75 and K82 with DNA are comparable in both structures.

E. Sequence differences between Smad1/5/8 on the Smad8 MH1 structure. The differences cluster around the Zn^{2+} binding region.

Supplementary Figure S3. Multidimensional IM-MS data of Smad5, Smad 3 and Smad4 MH1 domains



A. Plot of the mobility drift time versus *m/z* for the Smad5 MH1 domain (aqueous solution, 200 mM ammonium acetate buffer) showing the multidimensional data, (instrumental settings are provided in table S3). The mass spectrum from all ions shown below. Peaks corresponding to monomers (m) and dimers (d) present in the gas phase are labeled.

B. Close-up representation (left: 2D contour plot and right, 3D plot) of the conformations corresponding to monomers and dimers that are present at a given m/z value but that are separated by ion mobility.

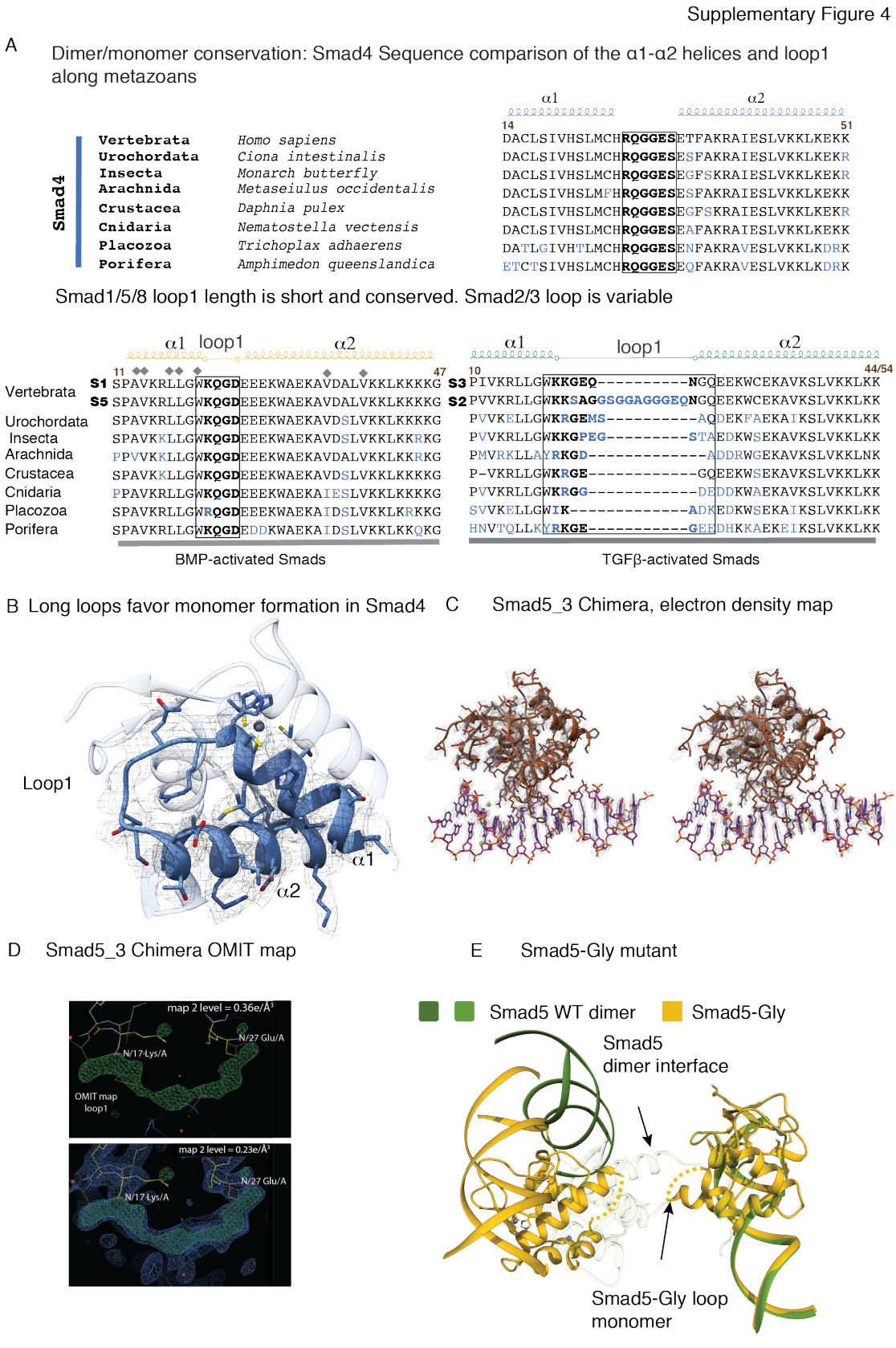
C. Ion mobility time distributions for three selected charge states. Collision cross-sections and the conformations used for the fitting. Calibration curves using standard proteins were used prior to the cross-section analysis.

D. Equivalent results to panel A are shown for Smad3

E. Equivalent results to panels A, B and C are shown for Smad4

F. Equivalent results to panel A are shown for Smad5 with the modified loop 1.

Supplementary Figure S4: Sequence conservation and dimer interface



A. Comparison of the α1-α2 regions along different metazoans. The specific organisms used in the comparison are indicated in the Smad4 alignment.

B. As in Figure 4C, for the Smad4 MH1 monomer (pdb: 5MEY). Smad4 does not show side-chain contacts stabilizing the loop1 orientation, in opposition to Smad3, while still maintaining the monomeric arrangement. Some elements of the secondary structure are labeled.

C. Section of the electron-density map (stereo-view) contoured at 1.0σ for the Smad5_3 chimera complex. The map was calculated with coefficients 2Fo – Fc, and the refined model (ribbon) is superimposed.

D. OMIT and electron-density maps corresponding to the loop1 (Smad5_3 chimera). The refined model is superimposed and some residues are labeled.

E. Superposition of the Smad5 dimer and the Smad5_gly monomer with Loops 1 labeled. The loop in the Smad5_gly monomer is indicated as a dotted line since its corresponding electron-density is not well defined.

Supplementary Tables

Supplementary Table S1. Analysis of Water Molecules (WM) that mediate interactions between MH1 domains and the GGCGC DNA

Color code:

Underlined: WM between Monomer B and DNA

Black: WM between Monomer A and DNA

Gray: Crystallographic WM between Monomer A and DNA

Protein complex	Water number	Atoms involved, chain identifier, BB: backbone distances			
Smad5	**W23**	PO_4 of G8 (C) 2.63 Å	Gln77 (A) 2.78 Å	BB His80, (A) 3.1 Å	BB Ser79 (A) 3.5 Å
	W24	PO_4, of G8 (C) 2.7, 3.4 Å	BB Gln77 (A) 2.7 Å		
	W36	PO_4 of G10 (C) 2.7 Å	BB Asp73 (A) 3.0 Å	BB Leu72 (A) 3.5 Å	
	W53	PO_4, G8 (D) 2.8, 3.6 Å	BB Gln77 (B) 2.8Å		
	W58	PO4 of G8 (D) 2.7 Å	Gln77 (B) 2.9 Å	BB His80 (B) 3.4 Å	BB Ser79 (B) 3.3 Å
	W100	T1' Carbonyl 3.5 Å	Glu51 (A) 3.0Å	BB Gly47 (A) 3.2 Å	
	W118	PO_4 of G8 (D) 2.76,3.40 Å	BB Ser79 (B) 3.5Å		
	W122	PO_4 of C9 (D) 2.7Å	BB Arg75 (B) 3.0 Å		
	W130	N7 of G8 (D) 2.8 Å	His80 (A) 3.6/3.3 Å	W85 3.1 Å	W59 3.1 Å
	W131	N7 of G6 (C) 3.3Å	N4 of C7 (C) 2.9Å	Gln77 (B) 3.1Å	W132, 3.3Å
Smad8	**W49**	PO_4 of G10 (C) 3.2Å	Asp73 (A) 2.8 Å	BB Asp73 (A) 3.6 Å	Ser71 (A) 3.5 Å
	W90	PO_4 of C9 (D) 3.0 Å	Lys41 (B) 2.8 Å	W69 3.2Å	
	W91	G8 PO_4 (D) 2.7 3.33Å	Ser37 (B) 3.3Å	W69 3.1 Å	
	W95	PO_4 of C9 (D) 3.1Å	BB Pro69 (A) 3.6Å		
	W96	PO_4 of G10 (D) 2.8Å	BB Leu72, (B) 3.5Å	BB and Asp73 (B) 3.1Å, 3.6Å	W108, 2.8Å
	W97	PO_4 of C9 (D) 2.8 Å	BB Arg75 (B) 3.1 Å		

Supplementary Table S2. RMSD values obtained from the comparison of several MH1 domain

RMSD - Cα	Smad8 (5-135)	Smad1:3KMP (9-132)	Smad4:3QSV (8-135)
Smad5 (10-132)	0.25 Å	0.30 Å	0.60 Å

RMSD values for the Smad5 MH1 domain superimposed to Smad8, Smad1 and Smad4 MH1 domains. Compared residues used for the fit are indicated in parenthesis. The PDB IDs are indicated for Smad1 and Smad4.

Supplementary Table S3. Experimental parameters used for ion mobility

NanoESI + V Resolution mode		DC potentials (V)		Traveling wave parameters	
m/z range	500-800	Trap Collision Energy	6	Source Wave Velocity (m/s)	300
Spray voltage (kV)	1.75	Transfer Collision Energy	4	Source Wave Height (V)	0.2
Gas pressure (psi)	0.5	Trap DC Entrance	5	Trap Wave Velocity (m/s)	100
Source Temperature (°C)	40	Trap DC Bias	15	Trap Wave Height (V)	0.2
Sampling Cone (V)	40	Trap DC Exit	5	IMS Wave Velocity (m/s)	300
Extraction Cone (V)	3	IMS DC Entrance	5	IMS Wave Height (V)	8
Desolvation temperature (°C)	250	IMS DC Exit	2	Transfer Wave Velocity (m/s)	200
Cone Gas Flow (L/hr)	10	Transfer DC Entrance	2	Transfer Wave Height (V)	8
Desolvation Gas Flow (L/hr)	300	Transfer DC Exit	2	IMS T-Wave pressure (mbar of N_2)	0.48
Trap Gas Flow (mL/min)	8				
IMS Gas Flow (mL/min)	25				
Backing region (mbar)	5.77				
Trap and Transfer T-Wave region pressure (mbar of Ar)	$5.95 \cdot 10^{-2}$				

Experiments acquired on a SYNAPT G1-HDMS (Waters, Manchester, UK)

References

1 Massagué, J. How cells read TGF-beta signals. *Nat Rev Mol Cell Biol* **1**, 169-178, doi:10.1038/35043051 (2000).
2 Massagué, J. TGFbeta signalling in context. *Nat Rev Mol Cell Biol* **13**, 616-630, doi:10.1038/nrm3434 (2012).
3 Massagué, J. TGF-beta signal transduction. *Annu Rev Biochem* **67**, 753-791, doi:10.1146/annurev.biochem.67.1.753 (1998).
4 Macias, M. J., Martin-Malpartida, P. & Massagué, J. Structural determinants of Smad function in TGF-beta signaling. *Trends Biochem Sci* **40**, 296-308, doi:10.1016/j.tibs.2015.03.012 (2015).
5 Shi, Y. & Massagué, J. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell* **113**, 685-700 (2003).
6 Macias, M. J., Wiesner, S. & Sudol, M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Lett* **513**, 30-37, doi:10.1016/s0014-5793(01)03290-2 (2002).
7 Morales, B. *et al.* NMR structural studies of the ItchWW3 domain reveal that phosphorylation at T30 inhibits the interaction with PPxY-containing ligands. *Structure* **15**, 473-483, doi:10.1016/j.str.2007.03.005 (2007).
8 Alarcon, C. *et al.* Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. *Cell* **139**, 757-769, doi:10.1016/j.cell.2009.09.035 (2009).
9 Aragon, E. *et al.* A Smad action turnover switch operated by WW domain readers of a phosphoserine code. *Genes Dev* **25**, 1275-1288, doi:10.1101/gad.2060811 (2011).
10 Gao, S. *et al.* Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signaling. *Mol Cell* **36**, 457-468, doi:10.1016/j.molcel.2009.09.043 (2009).
11 Wang, R. N. *et al.* Bone Morphogenetic Protein (BMP) signaling in development and human diseases. *Genes Dis* **1**, 87-105, doi:10.1016/j.gendis.2014.07.005 (2014).
12 Ramachandran, A. *et al.* TGF-beta uses a novel mode of receptor activation to phosphorylate SMAD1/5 and induce epithelial-to-mesenchymal transition. *Elife* **7**, doi:10.7554/eLife.31756 (2018).
13 Martin-Malpartida, P. *et al.* Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors. *Nat Commun* **8**, 2070, doi:10.1038/s41467-017-02054-6 (2017).
14 Katagiri, T. *et al.* Identification of a BMP-responsive element in Id1, the gene for inhibition of myogenesis. *Genes Cells* **7**, 949-960 (2002).
15 Baburajendran, N., Jauch, R., Tan, C. Y., Narasimhan, K. & Kolatkar, P. R. Structural basis for the cooperative DNA recognition by Smad4 MH1 dimers. *Nucleic Acids Res* **39**, 8213-8222, doi:10.1093/nar/gkr500 (2011).
16 BabuRajendran, N. *et al.* Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-beta effectors. *Nucleic Acids Res* **38**, 3477-3488, doi:10.1093/nar/gkq046 (2010).
17 Chai, N. *et al.* Structural basis for the Smad5 MH1 domain to recognize different DNA sequences. *Nucleic Acids Res* **45**, 6255-6257, doi:10.1093/nar/gkx226 (2017).
18 Shi, Y. *et al.* Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. *Cell* **94**, 585-594 (1998).
19 Flanders, K. C. *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* **62**, 846-863, doi:10.1369/0022155414550163 (2014).
20 Gronroos, E. *et al.* Transforming growth factor beta inhibits bone morphogenetic protein-induced transcription through novel phosphorylated Smad1/5-Smad3 complexes. *Mol Cell Biol* **32**, 2904-2916, doi:10.1128/MCB.00231-12 (2012).
21 Blanchet, C., Pasi, M., Zakrzewska, K. & Lavery, R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* **39**, W68-73, doi:10.1093/nar/gkr316 (2011).
22 Aragon, E. *et al.* Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-beta signaling. *Genes Dev* **33**, 1506-1524, doi:10.1101/gad.330837.119 (2019).
23 Amoutzias, G. D., Robertson, D. L., Van de Peer, Y. & Oliver, S. G. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci* **33**, 220-229, doi:10.1016/j.tibs.2008.02.002 (2008).
24 Hou, C. & Tsodikov, O. V. Structural Basis for Dimerization and DNA Binding of Transcription Factor FLI1. *Biochemistry* **54**, 7365-7374, doi:10.1021/acs.biochem.5b01121 (2015).
25 Huang, Y. H., Jankowski, A., Cheah, K. S., 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* **5**, 10398, doi:10.1038/srep10398 (2015).
26 Aragon, E. *et al.* Structural basis for the versatile interactions of Smad7 with regulator WW domains in TGF-beta Pathways. *Structure* **20**, 1726-1736, doi:10.1016/j.str.2012.07.014 (2012).
27 Jayaraman, L. & Massagué, J. Distinct oligomeric states of SMAD proteins in the transforming growth factor-beta pathway. *J Biol Chem* **275**, 40710-40717, doi:10.1074/jbc.M005799200 (2000).
28 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* **17**, 4056-4065, doi:10.1093/emboj/17.14.4056 (1998).
29 Chacko, B. M. *et al.* Structural basis of heteromeric smad protein assembly in TGF-beta signaling. *Mol Cell* **15**, 813-823, doi:10.1016/j.molcel.2004.07.016 (2004).
30 Miyazono, K. I. *et al.* Hydrophobic patches on SMAD2 and SMAD3 determine selective binding to cofactors. *Sci Signal* **11**, doi:10.1126/scisignal.aao7227 (2018).
31 Derynck, R. & Budi, E. H. Specificity, versatility, and control of TGF-beta family signaling. *Sci Signal* **12**, doi:10.1126/scisignal.aav5183 (2019).
32 Lopez-Rovira, T., Chalaux, E., Massague, J., Rosa, J. L. & 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* **277**, 3176-3185, doi:10.1074/jbc.M106826200 (2002).
33 Morikawa, M., Koinuma, D., Miyazono, K. & Heldin, C. H. Genome-wide mechanisms of Smad binding. *Oncogene* **32**, 1609-1615, doi:10.1038/onc.2012.191 (2013).
34 Willhoft, O. *et al.* Structure and dynamics of the yeast SWR1-nucleosome complex. *Science* **362**, doi:10.1126/science.aat7716 (2018).
35 Marley, J., Lu, M. & Bracken, C. A method for efficient isotopic labeling of recombinant proteins. *J Biomol NMR* **20**, 71-75 (2001).
36 Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. Backbone dynamics of calmodulin studied by 15N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. *Biochemistry* **31**, 5269-5278 (1992).

37 Orekhov, V. Y. & Jaravine, V. A. Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. *Prog Nucl Magn Reson Spectrosc* **59**, 271-292, doi:10.1016/j.pnmrs.2011.02.002 (2011).
38 Delaglio, F. *et al.* NMRPipe : a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-293, doi:10.1007/ BF00197809 (1995).
39 Bartels, C., Xia, T. H., Billeter, M., Guntert, P. & Wuthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biomol NMR* **6**, 1-10, doi:10.1007/BF00417486 (1995).
40 Zander, U. *et al.* Automated harvesting and processing of protein crystals through laser photoablation. *Acta Crystallogr D Struct Biol* **72**, 454-466, doi:10.1107/S2059798316000954 (2016).
41 Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr D Biol Crystallogr* **67**, 293-302, doi:10.1107/S0907444911007773 (2011).
42 McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr D Biol Crystallogr* **63**, 32-41, doi:10.1107/S0907444906045975 (2007).
43 McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674, doi:10.1107/S0021889807021206 (2007).
44 Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877, doi:10.1107/S2059798319011471 (2019).
45 Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235-242, doi:10.1107/S0907444910045749 (2011).
46 Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**, 240-255, doi:10.1107/S0907444996012255 (1997).
47 Smart, O. S. *et al.* Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr D Biol Crystallogr* **68**, 368-380, doi:10.1107/S0907444911056058 (2012).
48 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and Development of Coot. *Acta Crystallographica Section D - Biological Crystallography* **66**, 486 (2010).
49 Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213-220, doi:10.1107/S2052252514009324 (2014).
50 Pettersen *et al.* UCSF Chimera - a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).
51 Bush, M. F. *et al.* Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. *Anal Chem* **82**, 9557-9565, doi:10.1021/ac1022953 (2010).
52 Franke, D. *et al.* ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr* **50**, 1212-1225, doi:10.1107/S1600576717007786 (2017).
53 Kleiger, G., Saha, A., Lewis, S., Kuhlman, B. & Deshaies, R. J. Rapid E2-E3 assembly and disassembly enable processive ubiquitylation of cullin-RING ubiquitin ligase substrates. *Cell* **139**, 957-968, doi:10.1016/j.cell.2009.10.030 (2009).
54 Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815, doi:10.1006/jmbi.1993.1626 (1993).
55 Svergun, D., Barberato, C. & Koch, M. H. CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Cryst.* **28**, 768-773, doi:10.1107/S0021889895007047 (1995).

56 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066, doi:10.1093/nar/gkf436 (2002).