HHS Public Access

Author manuscript

Nat Struct Mol Biol. Author manuscript; available in PMC 2010 September 06.

Published in final edited form as:

Nat Struct Mol Biol. 2010 January ; 17(1): 117–123. doi:10.1038/nsmb.1742.

THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves

Alex Sabogal[1,*], **Artem Y. Lyubimov**[1,2,*], **Jacob E. Corn**[1], **James M. Berger**[1,2,†], and **Donald C. Rio**[1,2,3,†]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[2]California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, CA 94720, USA

[3]Center for Integrative Genomics, University of California, Berkeley, Berkeley, CA 94720, USA

Abstract

THAP-family C_2CH zinc-coordinating DNA-binding proteins function in diverse eukaryotic cellular processes, such as transposition, transcriptional repression, stem-cell pluripotency, angiogenesis and neurological function. To determine the molecular basis for sequence-specific DNA recognition by THAP proteins, we solved the crystal structure of the *Drosophila melanogaster* P element transposase THAP domain (DmTHAP) complexed with a natural 10-base pair site. In contrast to C_2H_2 zinc fingers, DmTHAP docks a conserved β-sheet into the major groove and a basic C-terminal loop into the adjacent minor groove. We confirmed specific protein-DNA interactions by mutagenesis and DNA binding assays. Sequence analysis of natural and *in-vitro*-selected binding sites suggests several THAPs (DmTHAP, human THAP1 and THAP9) recognize a bipartite TxxGGGx(A/T) consensus motif; homology suggests THAP proteins bind DNA through a bipartite interaction. These findings reveal the conserved mechanisms by which THAP-family proteins engage specific chromosomal target elements.

Introduction

Recent genome sequencing efforts have identified the THAP domain, originally characterized as the N-terminal site-specific DNA-binding domain of the P element transposase of *Drosophila melanogaster*1,2, in over 300 proteins from animal genomes and

Contact: Donald C. Rio, don_rio@berkeley.edu. †Corresponding authors .
*These authors contributed equally to this work.

Author Contributions A. S. and D. C. R. conceived the experiments. D. C. R. synthesized brominated DNA oligonucleotides. A. S. purified the proteins, nucleic acids, and crystallized the complex. J. E. C., A. Y. L., and J. M. B. provided guidance in crystallography trials. A. Y. L. and A. S. collected and analyzed the structural data; J. M. B. assisted with model building and refinement; A. Y. L. solved the structure and made all structural models. A. S. performed the DmTHAP mutagenesis and biochemistry, the SELEX experiment for human THAP9, and the THAP DNA binding site sequence analysis. A. S., A.Y. L., J. M. B., and D. C. R. wrote the paper.

Accession codes: 3KDE

parasitic mobile elements3-6. Approximately 80 amino acids long, THAP domains are characterized by a Cys-X_{2-4}-Cys-X_{35-50}-Cys-X_2-His zinc-coordinating motif, and other signature elements, including a C-terminal AVPTIF sequence4,7. Mutations of these conserved sequence elements disrupt folding and DNA binding *in vitro* 1,8,9, and have been implicated by human genetics in neurological diseases when mutated or truncated9. THAP domains are the second-most common zinc-coordinating DNA-binding domain after the C_2H_2 class of zinc-fingers4,10,11. Typical of large DNA-binding protein families, primary sequence conservation among THAP homologues is low11, although secondary and tertiary structures, particularly the characteristic βαβ fold, are strongly conserved7,12.

The phylogenetic distribution of THAP proteins (which includes evidence of a recently active P element transposase-related THAP9 gene in zebrafish13), combined with the absence of THAPs in non-animal species, suggests a recent incorporation of the domain into eukaryotic genomes by domestication of an ancestral mobile element5,13. More generally, THAP proteins are thought to share a common ancestral DNA-binding fold with the P element transposase4,5. Other features often shared between the THAP family of transcription factors and P element transposases include: 1) the stereotypical location of the THAP domains at the N-termini of their resident open reading frames, 2) a basic nuclear localization signal (NLS; amino acids 64-67 in DmTHAP) embedded within or near the THAP domain, and 3) a C-terminal leucine-zipper or coiled-coil dimerization domain (amino acids 100-150 in P element transposase). These features allow THAP family transcription factors to enter the nucleus, bind to DNA with high affinity, and form higher-order oligomeric complexes with regulatory components, thereby linking DNA targeting functions with the regulation of chromatin remodeling and transcriptional repression14,15. Signature THAP sequence elements, including the C_2CH zinc-coordinating motif, are found in 12 human proteins, several of which have been functionally characterized as nuclear DNA-binding proteins (THAP016, THAP117, THAP518, THAP714, and THAP1119).At present, the mechanism by which THAP proteins recognize specific DNA sequences is unknown. Molecular insights into recognition are key to understanding how THAP family transcription factors are targeted to chromosomal sites to modulate key cellular processes. Indeed, many of the cellular THAP proteins studied to date act as transcription factors that control the expression of diverse sets of genes implicated in angiogenesis, apoptosis, cell cycle regulation, stem cell pluripotency, and epigenetic gene silencing8,14,15,17,19-21. THAP family members also have been implicated in a variety of human disease pathways from angiogenesis20 and heart disease18, to neurological defects9 and multiple types of cancer20-22.

To better understand THAP-DNA interactions, we purified a minimal 77-amino acid THAP domain (DmTHAP) from the *Drosophila melanogaster* P element transposase, which is necessary and sufficient for high-affinity DNA binding1,2,7, and determined its crystal structure in complex with a naturally-occurring 10 bp DNA site. Our results show that DmTHAP specifically recognizes sequence elements in a bipartite manner using both the major and minor grooves of its target DNA site. Minor groove recognition is achieved by a combination of direct base contacts and indirect sequence readout of DNA deformation through a variable, basic loop. By contrast, the adjacent major groove is recognized

sequence-specifically by the central β-sheet of the domain. Due to their common ancestry, the sequence-specific DNA binding events of other THAP proteins can be postulated at a molecular level. In particular, the binding sites of two human THAPs (hTHAP1 and hTHAP9) appear to share common features with loci recognized by DmTHAP, including the sequence identity and spacing to create a TxxGGGx(A/T) consensus target motif. Contrary to proposed helix-groove models for THAP-DNA interactions7, THAP domains instead engage appropriate target sites in complex genomes by a conserved bipartite β-sheet and loop-dependent readout mechanism.

Results

Overall fold and secondary structure elements

To visualize how THAP proteins interact with specific DNA sequences, we determined the crystal structure of DmTHAP in complex with a naturally occurring 10-base pair DNA site at 1.74Å resolution by single wavelength anomalous dispersion (SAD) methods. The quality of the resultant electron density maps (Table 1) allowed unambiguous mapping of both direct and water-mediated DNA-protein contacts. The final model includes the entire 10-base pair DNA substrate and residues 1 – 76 of the transposase, excluding two disordered amino acids in loop 4 (Pro57 and Ala58) (Figs. 1a, 1b).

As expected, DmTHAP adopts a βαβ fold characteristic of THAP domains seen previously in apo-NMR structures of human THAP1 and THAP2, and the *C. elegans* C-terminal binding protein (CtBP)7,12. Structurally, the core fold of DmTHAP aligns well with other members of the THAP family (1.39, 0.71 and 1.46 Å rmsd for hTHAP1, hTHAP2 and *C. elegans* CtBP, respectively, Fig. 1c and Supplementary Fig. 1). The rest of the molecule is composed of loops, of which loop 4 is the most variable in length, sequence and structure (Fig. 1d and Supplementary Fig. 1). DmTHAP binds DNA as a monomer, making a total of 17 direct and water-mediated base-specific contacts with two non-overlapping regions that span the entire binding site (Fig. 1e). This interaction buries ~2380 $Å^2$ of total surface area at the nucleoprotein interface.

Major Groove Protein-DNA Interactions

The main-chain atoms of the N-terminal methionine (Met1) recognize the 3′ GA sequence from the major groove at positions 9 and 10 (Figs. 1e, 1f, 2a). The β-sheet further interacts with the central GTGG sequence of the major groove, corresponding to positions 6-9 (Figs. 1e, 1f, 2b). His18 and Gln42 from the two β-strands, along with the N-terminus, make a total of six direct contacts with six bases and engage both strands of the DNA duplex in the major groove (Figs. 1e, 1f, 2b). The main-chain atoms of Tyr3, Leu16 and Asn40, along with the side-chain of Gln42, further interact with five additional bases in the major groove via bridging water molecules (Fig. 1e). Given the variability of the amino acid composition in the THAP domain β-sheet (Fig. 1d, Supplementary Fig.2), and the ability of water to accommodate different hydrogen bond donors and acceptors23, the structure indicates that some THAP paralogs will be able to accommodate major groove sequences that differ from DmTHAP.

Minor Groove Protein-DNA Interactions

Loop 4 (Arg65 and Arg67) interacts with the AT-rich sequence in the minor groove (positions 2-4, Figs. 1e, 1f, 2c, 2d). Loop 4 is the most variable portion of THAP domains4, yet at least one basic amino acid is found in this region (Fig. 1d and Supplementary Figs. 1 and 3). In DmTHAP, Arg65 contacts T7 directly and A4 through a bridging water molecule, while Arg67 makes water-mediated contacts with T3, A18 and A19 (Figs. 1e, 2c, 2d). By contrast, Arg66 projects away from the DNA and occupies two conformations, both of which are engaged in π-stacking interactions with Trp53 (Fig. 2e). This residue structurally restricts one end of loop 4, directing the main chain to allow Arg65 and Arg67 to project into the minor groove. Arg66 also interacts with Asp45, Cys44 and His47, thus anchoring loop 4 to the zinc-coordinating core of DmTHAP. Together, the base of loop 4 and the central β-sheet create two ridges that project into adjacent DNA minor and major grooves, respectively (Fig. 1f).

In addition to direct contacts with bases in both grooves, indirect readout of deformable DNA sequences plays a role in specific site recognition by DmTHAP. The main chain atoms of Lys64, Arg65, and Arg66 all interact with the backbone phosphates of A19 and G6, resulting in a noticeable narrowing of the minor groove, which is localized to the region contacted by loop 4 (Supplementary Figs. 4 and 5) .Distortions of local base-pair geometry appear to be most pronounced at positions 2, 3 and 4, corresponding to minor groove binding by Arg65 and Arg67, as analyzed using the programs 3DNA and CURVES+ (Supplementary Figs. 4 and 5). However, it is unknown at this time if the DNA distortion is a result of DNA binding, or is intrinsic to the DmTHAP binding sequence.

Validation of Specific Protein-DNA interactions by EMSA

We utilized electrophoretic mobility shift assays (EMSA) to determine the contribution of key residues in each groove towards the overall affinity. To examine the role of Met1, we deleted Tyr2 and Lys3, expecting that the truncated construct (ΔY2,K3) would perturb the position of the starting amino acid relative to the 3′ GA sequence. ΔY2,K3 displayed a partially reduced affinity (~3-fold) compared to wild type DmTHAP (Fig. 3a, c), suggesting that the N-terminus makes a modest contribution to the overall DNA-binding affinity. By contrast, the H18A and Q42A mutations substantially impaired DNA binding (~12-fold, ~15-fold reduction, respectively), with the double H18A Q42A mutant protein exhibiting an even greater (~20-fold) reduction in affinity (Fig. 3a, c). The mutations R65A and R67A led to a similar loss of DNA binding (~21-fold and ~17-fold respectively), with an even greater (~42-fold) loss of binding for the R65A R67A double mutant (Fig. 3b, c). The R66A mutation resulted in a complete loss of binding (Fig. 3b), which may be attributable to a possible destabilization of the core DmTHAP structure. Taken together, the biochemical analysis of base-specific contacts in both the major and minor grooves validates the DNA-protein interactions observed in the co-crystal structure.

Bipartite DNA targeting by Other THAP Proteins

Despite poor sequence conservation, the known tertiary structures of THAP proteins are highly similar, suggesting that the DNA recognition strategies used by DmTHAP are preserved among different THAP homologs. In support of this proposal, superposition of

three previously-reported DNA-free structures of hTHAP1, hTHAP2 and *C. elegans* CtBP 7,12 with the DNA-bound DmTHAP seen here results in plausible binding orientations for all proteins (Supplementary Fig. 3). In particular, each of these related THAP domains seems capable of interacting with DNA in a manner analogous to DmTHAP, with the conserved β-sheets of all three proteins docking into the major groove without steric hindrance (Supplementary Fig. 3). Homology-based structural models of all twelve human THAP proteins (hTHAP0 - hTHAP11) further indicate that the DNA-binding β-sheet is likely conserved across the THAP family (Supplementary Fig. 2). Although specific interactions with DNA cannot be inferred from these models, the apparent diversity of putative major groove-binding elements suggests that paralogous THAP domains likely recognize a variety of distinct target site sequences in the major groove, most of which are unknown. Similarly, we note that the orientation of loop 4 with respect to the minor groove may also be variable, although in all cases some degree of engagement between this element and DNA can be modeled (Supplementary Fig. 3). Together, the structural models indicate that most THAP family members rely on a bipartite model for engaging DNA, and that the diversity of binding elements in the β-sheet likely correlates with a diversity of recognition sequences in the major groove.

THAP binding site analysis

To determine whether THAP binding sites contain any signature sequence elements, we performed an alignment of experimentally-verified natural target sites for DmTHAP2,24 and hTHAP120 with target sites determined by SELEX for human THAP18 and THAP9. These alignments allowed us to subdivide known THAP-binding regions on the DNA into major and minor-groove-interacting sub-sites (Fig. 4). The natural sites for the P-element transposase and human THAP1, as well as the SELEX motifs for human THAP1 and THAP9, are all 9-11 base pairs in length. This metric appears to correspond to a single THAP domain binding site, and is consistent with the ~10 base pair DNA duplex used in our co-crystallization experiments.

Position 3 in the DmTHAP minor groove sub-site contains a conserved A-T base pair, which both interacts with the basic loop 4 and is a region of local distortion (Figs. 1e, 1f, 2d and Supplementary Figs. 4 and 5). Interestingly, an A-T base pair is found at the same position in the hTHAP1 and hTHAP9 binding sites reported to date, suggesting it is a critical recognition determinant for these proteins, as it is for DmTHAP (Fig. 4). Both hTHAP1 and hTHAP9 also contain at least one basic side chain in the loop 4 region (Fig. 1d), which could mediate binding the conserved A-T base pair in a manner analogous to DmTHAP. Moreover, in the SELEX motifs, the spacing between the conserved T at position 3 is ~5 base pairs, or one DNA half-turn, away from the next conserved sequence block (GGG or GGGCA), which comprises the major groove sub-site (Fig. 4); the spacing between the major and minor groove sub-sites further is restricted to two base pairs in all available THAP target sites. The DmTHAP structure reveals that this spacing is necessary for the protein to arch over the DNA backbone and bind both grooves on the same face of the duplex (Fig. 1f). Taken together, these results suggest that a common core set of DNA sequence motifs may be conserved between DmTHAP and the THAP1 and THAP9 subfamilies.

Discussion

DmTHAP Utilizes a Novel DNA-targeting Mechanism

The ability of DmTHAP to use a β-sheet for recognizing the DNA major groove differs dramatically from the binding mode employed by canonical C_2H_2 zif268 zinc fingers, to which it has been compared previously (Figs. 5a, 5b). The typical ~30-amino acid C_2H_2 zinc-finger motif presents an α-helix into the major groove of DNA25. Classical C_2H_2 zinc-finger proteins also are highly modular, recognizing extended DNA sequences through the use of several tandem copies of the domain10,11,25. By contrast, most THAP protein family members have only a single N-terminal THAP domain4, possibly due to a need for the N-terminal amino group to contact DNA.

β-sheet/major groove interactions have been observed in other structures, such as the Arc and MetJ repressors26, the N-terminal domain of the λ integrase27 and the Tn916 transposase DNA-binding domain28. However, notable differences between these structures and DmTHAP also are present (Fig. 5). For example, the β-sheets of Arc and MetJ are composed of strands donated by individual subunits of a homodimer, whereas DmTHAP is monomeric. The λ integrase N-terminal domain is similar to DmTHAP in combining a major groove-binding β-sheet with a minor groove-binding element, but uses a 3_{10} helix, rather than a loop. The DNA-binding domain of Tn916 transposase uses a β-sheet to bind the major groove and a loop to engage the minor groove. However, the minor groove contacts of the Tn916 DNA-binding domain are predominantly with the phosphate backbone rather than with the bases, and therefore do not appear to be sequence-specific. Overall, the RRR sequence of DmTHAP loop 4 is perhaps most reminiscent of the “AT-hook” motif found in HMG proteins29, in which two arginine residues, separated by a single amino acid, insert into the minor groove to contact specific bases. Taken together, these comparisons indicate that THAP domains utilize a unique combination of DNA-recognition strategies to engage their target sites, allowing for the possibility of engineering of novel DNA binding specificities.

Direct Sequence Readout by β-sheet Side Chains

The N-terminus of THAP proteins, up to the first zinc-coordinating cysteine, is typically 2 - 4 residues long4. Therefore, it appears likely that an interaction between the N terminal-most methionine and DNA is often preserved across the THAP family. By contrast, the β-sheet residues used by DmTHAP to bind DNA show remarkably little sequence conservation (Fig. 1d)4. It seems likely that variation at these β-sheet positions, along with variation in the precise length and composition of the N-terminus, alters the DNA sequence(s) recognized by the THAP proteins through the major groove. In agreement with this premise, a previous study of a natural C-terminal deletion mutant repressor form of P element transposase assessed the effects of the H18A mutation by DNase I footprinting on its natural DNA binding site1 and found that, in the context of the truncated 207 amino acid KP repressor protein, the H18A mutant exhibited non-specific DNA binding behavior while retaining high affinity for DNA duplexes. Interestingly, the most highly conserved THAP domain residues appear to play structural roles in forming and stabilizing the hydrophobic core of the protein 7,12.

Loop 4 Sequence Affects DNA Binding

Of all of the THAP proteins analyzed here, *C. elegans* CtBP has one of the shortest loop 4 regions. Although CtBP retains the consensus C-terminal AVPTIF motif, the internal truncation of loop 4 suggests that the protein may interact with the minor groove in a manner distinct from DmTHAP. Nonetheless, our modeling studies suggest that CtBP loop 4 does retain a pair of lysines that appear to be within interacting distance of the phosphate backbone or perhaps capable projecting into the minor groove (Fig. 1d and Supplementary Fig. 3d). Human THAP11 (Ronin) has a loop 4 similar to CtBP, and may bind DNA in an analogous fashion. By contrast, truncating the C-terminus of DmTHAP at position 73 disrupts the AVPTIF motif (AVPSKV in DmTHAP), resulting in the destabilization of loop 4 and loss of DNA binding1. Thus, the molecular definition of a minimal THAP domain must include the AVPTIF motif to complete the fold and optimally position minor groove binding residues.

A narrowing of the minor groove is observed at the positions bound by the basic loop 4 in DmTHAP, where it likely contributes to DNA site selection by indirect readout (Supplementary Figs. 4 and 5). This phenomenon may be present in other THAPs. For example, the SELEX-derived motifs of several monomeric THAP binding sites indicate that the information content in the minor groove position 3 is higher than background (≥1) for both hTHAP1 and hTHAP9 (Fig. 4), consistent with high minor groove conservation signatures and distorted DNA observed in several replication proteins30.

Bipartite DNA-binding Model Applied to Human THAP1

The bipartite binding model presented here can be used to explain several biochemical and biophysical observations of hTHAP1, as well as the molecular basis for generalized human dystonia (DYT6) in adults9. For example, EMSA studies of human THAP1 using an *in vitro*-derived 11-base pair target sequence (known as THABS, AGTAAGGGCAA) showed binding defects when the core TxxGGCA recognition motif was mutated8. Our model suggests these defects are likely caused by the disruption of key major and minor groove interactions. In the same system, NMR experiments showed measurable changes in chemical shifts occurring upon DNA addition that could be associated with residues identified here as important for DNA binding7. Although not a direct indicator of DNA binding, these data revealed large chemical shifts for several amino acids located in loop 4, which is disordered in the absence of DNA7, presumably due to the docking of loop 4 to the minor groove. These observations, coupled with the hTHAP1 SELEX analysis and structural modeling described above, are consistent with a bipartite targeting mechanism for hTHAP1.

The DmTHAP-DNA structure similarly can explain the defects in genetically-identified hTHAP1 mutants that cause DYT69, a disease that results in abnormal or repetitive movements of the limbs, as well as speech defects31. In one reported deletion mutant, hTHAP1 loop 4 is truncated upstream of the AVPTIF motif that is needed to complete the THAP fold and help position basic residues to bind the minor groove sub-site. This deletion, as well as a single point mutant, Phe81Leu (affecting the phenylalanine position in the AVPTIF motif), both have been shown to dramatically reduce DNA binding9. Phe81 sits far from the DNA-binding interface, but within the AVPTIF motif, and thus may also affect

DNA binding by destabilizing the structure of loop 4. Alternatively, the Phe81Leu substitution could affect other aspects of DNA binding in the context of dimeric, full-length hTHAP1.

The downstream consequences of DNA-binding defects of hTHAP1 are believed to include a reduced repression of hTHAP1 target genes, resulting in aberrant transcriptional programs for genes involved in cell-cycle control and cell-cycle growth17,20. Thus, structural information from the DmTHAP-DNA complex can link substitution or deletion of specific amino acid residues to disruption of neurological function through the role of these residues in DNA binding and structural stability. Furthermore, putative hTHAP1 binding sites can now be better identified with the understanding of how they are recognized by THAP domains. Knowledge of the molecular mechanism of specific DNA site recognition by THAP domains should facilitate the further study of the downstream effects of DNA binding.

THAP Domain Oligomerization and Regulation

While single THAP domains bind to DNA as monomers, many family members are predicted to form dimers (or possibly higher order oligomers) by a common C-terminal leucine-zipper/coiled-coil motif2. Dimerization allows for multi-site DNA binding in THAP proteins, exemplified by the *D. melanogaster* P element transposase2, and postulated for human THAP11 (Ronin), which has a 20 base pair binding site and a predicted leucine-zipper domain19,22. Though uncommon, multi-THAP domain-containing proteins do exist4,12; an extreme example is the open reading frame CG10631 from *D. melanogaster* with 27 tandem THAP domains and with no known function4. Furthermore, human THAP7 and THAP11 are found together in a transcriptional repression complex19, which may utilize several THAP domains for complex multi-site/multi-sequence binding events. Regulation of DNA binding by THAP proteins also is postulated to occur for certain THAP homologs. For example, the THAP1 and THAP5 mRNAs are predicted to be alternatively spliced whereby one isoform lacks a complete THAP domain, while the *Drosophila* transcriptional co-repressor, CtBP, lacks the DNA binding THAP element found in its *C.elegans* counterpart6.

Conclusions

In summary, our structure provides the first general model for DNA recognition by the abundant THAP domain protein family. THAP domains comprise a unique class of C_2CH zinc-coordinating, DNA-binding folds which, in contrast to canonical C_2H_2 zinc fingers such as zif268, as well as the nuclear receptor superfamily and GATA-1 factors3,11, use a β-sheet to bind DNA in the major groove and make additional specific minor groove protein-DNA contacts using a C-terminal basic loop. Based on structural, biochemical, and bioinformatic results, we propose that THAP domains target DNA through a novel bipartite mechanism, with some family members targeting a consensus sequence of TxxGGGx(A/T) that bears readily identifiable major and minor groove sub-sites. Local variations in target DNA sequence can be accommodated by amino acid substitutions in the β-sheet, loop 4, and (to a lesser extent) N-terminal length and sequence. The structural insights presented here

significantly advance our knowledge of THAP domain function and the mechanism of sequence-specific protein-DNA recognition. This analysis should aid in the understanding of yet unstudied biological processes in humans and diverse animals that depend on THAP domain-containing DNA binding proteins.

Methods

Protein Purification

We amplified amino acids 1-77 of the *Drosophila* P-element transposase using primers GCATGAAATCATATGAAGTACTGCAAGTTCTGC and GCGTACTTACCATGGTTACACCTTGGAGGGCACGGCGTC, then subcloned the product into pRSETA (Invitrogen), using growth and expression as described for PN881. We sonicated frozen cell pellets with 10 mL lysis buffer (25 mM Hepes-KOH, pH 7.6, 1 M NaCl, 10% (v/v) Glycerol, 1 mM PMSF, 0.5 mM TCEP, 0.5 μg ml^{-1} of leupeptin, pepstatin, aprotinin, antipain, chymostatin) per gram of frozen bacterial paste. We removed nucleic acids from clarified lysates by addition of 30 ml of Q-Sepharose Fast-Flow resin (Pharmacia) at 4°C for 1 hr. We diluted the flow-through five-fold with buffer A (25 mM Hepes-KOH, pH 7.6, 10% (v/v) glycerol), then filtered, and loaded the material onto a 30 ml SP- Fast-Flow column (Pharmacia) pre-equilibrated with 80% (v/v) buffer A and 20% (v/v) buffer B (25 mM Hepes-KOH, pH 7.6, 1 M NaCl, 10% (v/v) glycerol). We added $ZnSO_4$ and TCEP to final concentrations of 10 μM and 0.5 mM, respectively to elutions. Following dialysis against 10% (v/v) buffer B plus 10 μM $ZnSO_4$, we loaded the solution onto an 8 ml heparin-agarose column, and DmTHAP was eluted with a linear gradient of 10%-55% (v/v) buffer B. Again, $ZnSO_4$ and TCEP were added. Using a 120ml Superdex 75 gel filtration column (GE Healthcare), DmTHAP eluted as a monomer at ~10 kDa, and we concentrated it to ~20 mg ml^{-1}, and froze aliquots in liquid nitrogen in gel filtration buffer (10 mM Hepes-KOH, pH 8.0, 50 mM NaCl, and 0.5 mM TCEP).

We made proteins for EMSA assays by adding a C-terminal $histidine_6$ tag to the DmTHAP construct. We made point mutants by overlapping-primer PCR and expressed them as described above for DmTHAP. We purified proteins using a 1 ml HiTRAP FF column as described by the manufacturer (Pharmacia), then spiked $ZnSO_4$ and TCEP to final concentrations of 10 μM and 0.5 mM, respectively, and loaded the material onto a 24 ml Superdex 75 column (Pharmacia). We froze aliquots in gel filtration buffer as described above. We cloned hTHAP9 amino acids 1-99 into pRSETA (Invitrogen), added a C-terminal $histidine_6$ tag, and purified it similarly.

Preparation of oligonucleotides

We synthesized brominated DNA oligonucleotides on ABI model 392 DNA synthesizer at 1 μmol scale, with an overnight manual elution using 1.5 ml NH_4OH at room temperature. We purified the oligonucleotides using 19% (w/v) polyacrylamide/8.3 M urea denaturing gels, visualized by UV shadowing, and extracted gel slices in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) at 37°C, then desalted the buffer with two rounds of ethanol precipitation. We resuspended purified single-strand oligonucleotides in 10 mM Hepes-KOH, pH 8.0, and 50 mM NaCl, then heated equimolar amounts of each to 65°C and slowly cooled them.

EMSA assays

We performed EMSA assays with the oligo 5′ GAGGTTAAGTGGATGT 3′ and 5′ TACATCCACTTAAC 3′, purified as described above. We 5′ end-labeled the 15mer duplex with T4 PNK (USB), P32 gamma ATP (GE Healthcare) and a P-6 column (Bio-Rad). We measured apparent K_d using 1nM DNA with increasing protein in a 20 μl reaction volume (10 mM Hepes-KOH, pH 8.0, 50 mM NaCl, and 10% (v/v) Glycerol), for 30 min. at room temperature and loaded the reaction onto a native 5% (w/v) polyacrylamide gel. We ran the gel for 1 hr at 150 V at 4°C, in 0.5X TBE buffer (0.089 M Tris-base, 0.089 M boric acid, 2 mM EDTA, pH 8.35), then dried and visualized the results using the Typhoon Phosphoimager system, then performed binding analysis using Prism5 (Graphpad Software).

Co-crystallization of DmTHAP with DNA

We used vapor diffusion methods with a Mosquito crystallization system (TTP LabTech) with 200 nL dropsize to produce diffraction-quality crystals in 24% (w/v) polyethylene glycol (PEG), MW 8000 (Fluka), 5 mM NaCl, 0.05 M CAPSO, pH 9.0 (Hampton Research), 10 mM TCEP at 25°C in ~3-5 days. For cryoprotecting, we incubated the drop with 26% (w/v) PEG, MW 8000, 25 mM NaCl, 0.05 M CAPSO, pH 9.0, 10 mM TCEP, 20% (v/v) xylitol. We collected diffraction data at the Advanced Light Source (ALS) beamline 8.3.1 from a single crystal at wavelength 0.92 Å over a 360° wedge using 1° oscillations. We integrated and scaled reflections in HKL200035 with separate scaling of anomalous pairs. We determined phases by single-wavelength anomalous dispersion (SAD) using Phaser HYSS36. We improved electron density maps by solvent flattening (RESOLVE)37,38 in PHENIX AutoSol Wizard36,39. We manually modeled DNA and protein using Coot40. Automated refinement (Refmac5)41 and manual modeling produced R_{work} and R_{free} values of 17.7% and 21.5%, respectively. We validated the structure using SFCHECK42, PROCHECK43 and Coot. In the final model, 100% of Ramachandran plot values fell into favored regions. We deposited atomic coordinates and structure factors to the Protein Data Bank under the code 3KDE.

We made figures and alignments of DmTHAP with NMR structures using PyMOL44. We performed homology modeling of human THAP proteins using PHYRE45. We calculated DNA distortion using 3DNA46 and CURVES+47. We made structure-based multiple sequence alignments with 3DCoffee48,49 and JalView50.

hTHAP9 SELEX

We performed SELEX experiments using the method of Roulet and Bucher51, modified by Ogowa and Biggin (personal communication). Briefly, we incubated ~0.4 mg of recombinant hTHAP9 with 50 μl of TALON superflow (Clontech). We diluted saturated beads 1:5 with unbound resin and used ~10 μl of this slurry in binding experiments. We prepared random target dsDNA by PCR extension with oligos GGATTTGCTGGTGCAGTACAGTGGATCC-[N_{16}]-GGATCCCTTAGGAGCTTGAAATCGAGCAG and CTGCTCGATTTCAAGCTCCT. We incubated 10 μl of protein slurry with random DNA (~1-2 ng), in a 20 μl reaction in 1X SELEX buffer (10 mM Tris_HCl, 7.6, 50 mM NaCl, 5% (v/v) glycerol, 0.1% (v/v) NP40, 10 μM $ZnCl_2$, 5 mM $MgCl_2$), supplemented with 1 μg BSA (NEB) and 1 μg poly dI:dC. The

binding proceeded for 20 min. at room temperature, then we washed twice with wash buffer (SELEX buffer with 5 mM NaCl), then eluted in 100 μl elution buffer (SELEX buffer with 500 mM NaCl). We PCR amplified this fragment using 20 cycles with primers GGATTTGCTGGTGCAGTACA and CTGCTCGATTTCAAGCTCCT. After 4 rounds of selection, >10% of the starting material was retained, amplified, and subsequently sequenced using concatemerization51. We made sequence logos using the Delila program52 with 76 independent sites.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Jodi Gureasko and John Kuriyan (U. C. Berkeley) for use of equipment and technical expertise; Eric Abbate and Mike Botchan (U. C. Berkeley) for crystallography supplies and experimental design; Nat Echols and Tom Alber (U. C. Berkeley) for use of equipment and assistance with data collection; James Holton (A. L. S.) for assistance with data collection; Andy May (Fluidigm) for crystallography supplies and assistance with data collection; David King (U. C. Berkeley, H. H. M. I. Mass-Spec Facility) for mass-spectrometry analysis; Nobuo Ogowa and Mark Biggin (U. C. Berkeley) for reagents and expertise for the SELEX protocol; Ryan Schultzaberger and Mike Eisen (L. B. N. L.) for assistance in SELEX data analysis and for creating the sequence logos; Kathy Collins (U. C. Berkeley) for data analysis; David Wemmer, Mike Levine and Mike Botchan for critical reading of the manuscript. A. Y. L. is supported by an ACS postdoctoral fellowship, J. M. B. by the NCI (CA077307), and D. R. by the NIGMS (GM61987).

References

1. Lee CC, Beall EL, Rio DC. DNA binding by the KP repressor protein inhibits P-element transposase activity in vitro. EMBO J. 1998; 17:4166–74. [PubMed: 9670031]
2. Lee CC, Mul YM, Rio DC. The Drosophila P-element KP repressor protein dimerizes and interacts with multiple sites on P-element DNA. Mol Cell Biol. 1996; 16:5616–22. [PubMed: 8816474]
3. Finn RD, et al. The Pfam protein families database. Nucleic Acids Res. 2008; 36:D281–8. [PubMed: 18039703]
4. Roussigne M, et al. The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. Trends Biochem Sci. 2003; 28:66–9. [PubMed: 12575992]
5. Quesneville H, Nouaud D, Anxolabehere D. Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. Mol Biol Evol. 2005; 22:741–6. [PubMed: 15574804]
6. Nicholas HR, Lowry JA, Wu T, Crossley M. The Caenorhabditis elegans protein CTBP-1 defines a new group of THAP domain-containing CtBP corepressors. J Mol Biol. 2008; 375:1–11. [PubMed: 18005989]
7. Bessiere D, et al. Structure-function analysis of the THAP zinc finger of THAP1, a large C2CH DNA-binding module linked to Rb/E2F pathways. J Biol Chem. 2008; 283:4352–63. [PubMed: 18073205]
8. Clouaire T, et al. The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. Proc Natl Acad Sci U S A. 2005; 102:6907–12. [PubMed: 15863623]
9. Fuchs T, et al. Mutations in the THAP1 gene are responsible for DYT6 primary torsion dystonia. Nat Genet. 2009
10. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct. 2000; 29:183–212. [PubMed: 10940247]
11. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biol. 2000; 1 REVIEWS001.

12. Liew CK, Crossley M, Mackay JP, Nicholas HR. Solution structure of the THAP domain from Caenorhabditis elegans C-terminal binding protein (CtBP). J Mol Biol. 2007; 366:382–90. [PubMed: 17174978]
13. Hammer SE, Strehl S, Hagemann S. Homologs of Drosophila P transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human. Mol Biol Evol. 2005; 22:833–44. [PubMed: 15616143]
14. Macfarlan T, et al. Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor. J Biol Chem. 2005; 280:7346–58. [PubMed: 15561719]
15. Macfarlan T, Parker JB, Nagata K, Chakravarti D. Thanatos-associated protein 7 associates with template activating factor-Ibeta and inhibits histone acetylation to repress transcription. Mol Endocrinol. 2006; 20:335–47. [PubMed: 16195249]
16. Lin Y, Khokhlatchev A, Figeys D, Avruch J. Death-associated protein 4 binds MST1 and augments MST1-induced apoptosis. J Biol Chem. 2002; 277:47991–8001. [PubMed: 12384512]
17. Roussigne M, Cayrol C, Clouaire T, Amalric F, Girard JP. THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. Oncogene. 2003; 22:2432–42. [PubMed: 12717420]
18. Balakrishnan MP, et al. THAP5 is a human cardiac-specific inhibitor of cell cycle that is cleaved by the proapoptotic Omi/HtrA2 protease during cell death. Am J Physiol Heart Circ Physiol. 2009; 297:H643–53. [PubMed: 19502560]
19. Dejosez M, et al. Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells. Cell. 2008; 133:1162–74. [PubMed: 18585351]
20. Cayrol C, et al. The THAP-zinc finger protein THAP1 regulates endothelial cell proliferation through modulation of pRB/E2F cell-cycle target genes. Blood. 2007; 109:584–94. [PubMed: 17003378]
21. De Souza Santos E, De Bessa SA, Netto MM, Nagai MA. Silencing of LRRC49 and THAP10 genes by bidirectional promoter hypermethylation is a frequent event in breast cancer. Int J Oncol. 2008; 33:25–31. [PubMed: 18575747]
22. Zhu CY, et al. Cell growth suppression by thanatos-associated protein 11(THAP11) is mediated by transcriptional downregulation of c-Myc. Cell Death Differ. 2009; 16:395–405. [PubMed: 19008924]
23. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. 2001; 29:2860–74. [PubMed: 11433033]
24. Kaufman PD, Doll RF, Rio DC. Drosophila P element transposase recognizes internal P element DNA sequences. Cell. 1989; 59:359–71. [PubMed: 2553268]
25. Pavletich NP, Pabo CO. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science. 1991; 252:809–17. [PubMed: 2028256]
26. Suzuki M. DNA recognition by a beta-sheet. Protein Eng. 1995; 8:1–4. [PubMed: 7770446]
27. Fadeev EA, Sam MD, Clubb RT. NMR structure of the amino-terminal domain of the lambda integrase protein in complex with DNA: immobilization of a flexible tail facilitates beta-sheet recognition of the major groove. J Mol Biol. 2009; 388:682–90. [PubMed: 19324050]
28. Wojciak JM, Connolly KM, Clubb RT. NMR structure of the Tn916 integrase-DNA complex. Nat Struct Biol. 1999; 6:366–73. [PubMed: 10201406]
29. Geierstanger BH, Volkman BF, Kremer W, Wemmer DE. Short peptide fragments derived from HMG-I/Y proteins bind specifically to the minor groove of DNA. Biochemistry. 1994; 33:5347–55. [PubMed: 8172908]
30. Schneider TD. Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. Nucleic Acids Res. 2001; 29:4881–91. [PubMed: 11726698]
31. Muller U. The monogenic primary dystonias. Brain. 2009; 132:2005–25. [PubMed: 19578124]
32. Elrod-Erickson M, Rould MA, Nekludova L, Pabo CO. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. Structure. 1996; 4:1171–80. [PubMed: 8939742]

33. Schildbach JF, Karzai AW, Raumann BE, Sauer RT. Origins of DNA-binding specificity: role of protein contacts with the DNA backbone. Proc Natl Acad Sci U S A. 1999; 96:811–7. [PubMed: 9927650]
34. Somers WS, Phillips SE. Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. Nature. 1992; 359:387–93. [PubMed: 1406951]
35. Otwinowski, Z.; Minor, W. Processing of X-Ray Diffraction Data Collected in Oscillation Mode. In: Carter, CWJ.; Sweet, RM., editors. Methods in Enzymology. Vol. Vol. 276. Academic Press; Boston: 1997. p. 307-325.
36. Zwart PH, et al. Automated structure solution with the PHENIX suite. Methods Mol Biol. 2008; 426:419–35. [PubMed: 18542881]
37. Terwilliger TC. Maximum-likelihood density modification. Acta Crystallogr D Biol Crystallogr. 2000; 56:965–72. [PubMed: 10944333]
38. Terwilliger TC. Maximum-likelihood density modification using pattern recognition of structural motifs. Acta Crystallogr D Biol Crystallogr. 2001; 57:1755–62. [PubMed: 11717487]
39. Adams PD, et al. PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr. 2002; 58:1948–54. [PubMed: 12393927]
40. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr. 2004; 60:2126–32. [PubMed: 15572765]
41. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica Section D-Biological Crystallography. 1997; 53:240–255.
42. Vaguine AA, Richelle J, Wodak SJ. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr D Biol Crystallogr. 1999; 55(Pt 1):191–205. [PubMed: 10089410]
43. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: A program to check the stereochemical quality of protein structures. J. App. Cryst. 1993; 26:283–291.
44. DeLano, WL. The PyMOL Molecular Graphics System. DeLano Scientific; San Carlos, CA, USA: 2002.
45. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc. 2009; 4:363–71. [PubMed: 19247286]
46. Lu XJ, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc. 2008; 3:1213–27. [PubMed: 18600227]
47. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K. Conformational analysis of nucleic acids revisited: Curves+ Nucleic Acids Res. 2009; 37:5917–29. [PubMed: 19625494]
48. Poirot O, Suhre K, Abergel C, O'Toole E, Notredame C. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Res. 2004; 32:W37–40. [PubMed: 15215345]
49. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol. 2004; 340:385–95. [PubMed: 15201059]
50. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009
51. Roulet E, et al. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol. 2002; 20:831–5. [PubMed: 12101405]
52. Schneider TD, Stormo GD, Yarus MA, Gold L. Delila system tools. Nucleic Acids Res. 1984; 12:129–40. [PubMed: 6694897]

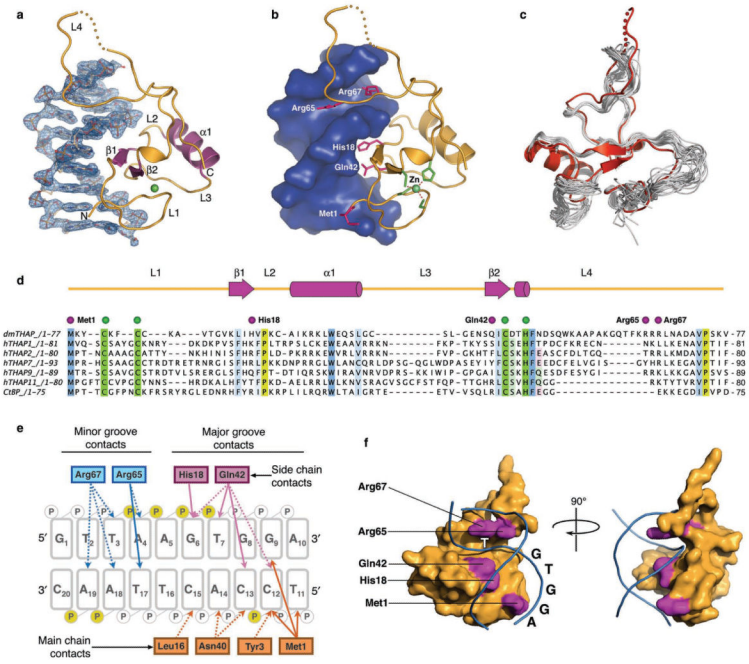


Figure 1.

Structure of DmTHAP-DNA complex and specific interactions with DNA. a) The protein-DNA interface. Experimental electron density map of the DNA (blue mesh) is contoured at 1.5σ. DmTHAP is shown as a ribbon diagram and labeled by secondary structure, with the βαβ motif highlighted in magenta. Zinc is shown as a green sphere. b) Base-specific interactions in the major and minor groove. Interacting amino acids are shown as magenta sticks; DNA is shown in blue surface representation; zinc-coordinating residues are shown as green sticks. c) Structural alignment of DmTHAP (red) and the solution structure of human THAP2 (grey, PDB ID: 2D8R). d) Structure-based multiple sequence alignment of DmTHAP, human THAP1, 2, 7, 9 and 11, and *C. elegans* CtBP. Conserved residues are highlighted; zinc-coordinating C_2CH motif is highlighted in green and indicated by green circles; base-specific DNA-binding residues of DmTHAP are indicated by magenta circles and are labeled. The secondary structure diagram is shown for DmTHAP and labeled as in (a). e) Schematic representation of all base-specific contacts in the major and minor groove. Direct contacts are shown as solid lines, base-specific water-mediated contacts are shown as dashed lines, interacting phosphates are highlighted yellow. f) Surface representation of DmTHAP. Sequence specific DNA-binding residues are highlighted in magenta. DNA backbone is shown as lines with sub-site positions labeled.

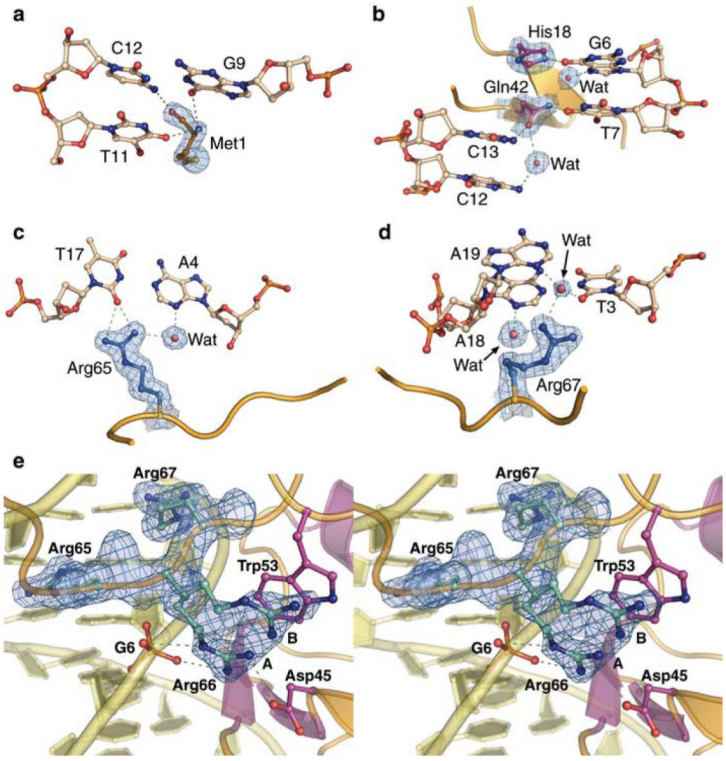
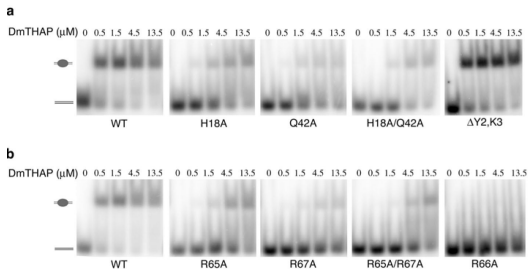


Figure 2.

Base-specific DmTHAP-DNA contacts. Interactions of a) Met1, b) His18 and Gln42, c) Arg65 and d) Arg67 with corresponding bases. Final electron density (calculated using $2F_o$-F_c coefficients and contoured at 1.5σ) is shown for interacting amino acids and bridging water molecules only. A cartoon representation of the β-sheet is shown in b). e) Stereographic representation of the RRR motif. Electron density for Arg65, Arg67 and the alternate conformations of Arg66 are contoured at 1.0σ. Side-chain and main-chain atoms of the RRR motif, as well as the side-chain atoms of Trp53 and Asp45 are shown in ball-and-stick representation.



c

DmTHAP construct	K_D (μM)	$K_{D,mut}$ / $K_{D,WT}$
WT	0.16	n/a
H18A	2.02	12
Q42A	2.48	15
H18A/Q42A	3.24	20
R65A	3.36	21
R67A	2.82	17
R65A/R67A	6.79	42
R66A	n/a	n/a
ΔY2,K3	0.47	3

Figure 3.

DmTHAP specificity mutant affinity determination by EMSA. a) Reduction in affinity observed in the major groove binding mutants H18A, Q42A, H18A Q42A, (ΔY2,K3). b) Reduction in affinity seen with the minor groove binding mutants R65A, R67A, R65A R67A, R66A. In each well, 1nM of a radioactive 15mer duplex DNA containing the specific transposase binding site was incubated with wild-type or mutant DmTHAP protein, allowed to equilibrate, and run on native 5% polyacrylamide gels. c) Table of apparent K_d values and fold reduction compared to wild-type DmTHAP.

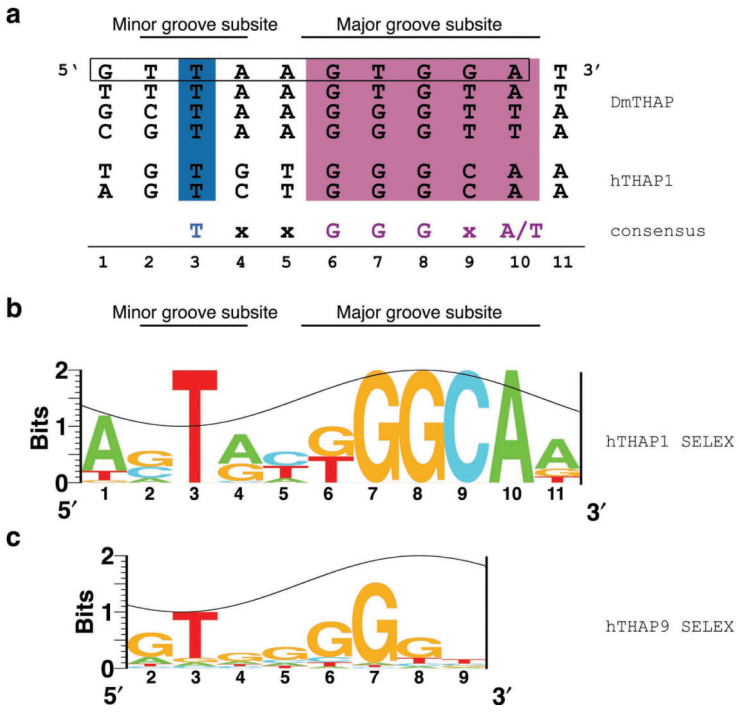


Figure 4.

Bipartite sequence readout by THAP proteins. a) Experimentally-verified naturally-occurring binding sites for the P-element transposase and human THAP1. The consensus major and minor groove sub-sites are highlighted in magenta and blue, respectively. The sequence used for co-crystallization with DmTHAP is boxed. b) Sequence logos made from position-specific scoring matrixes from SELEX experiments of human THAP1 (ref. 8) and c) human THAP9. DNA helical phasing is represented as an 11 base pair SIN wave and positioned based on DmTHAP structure.

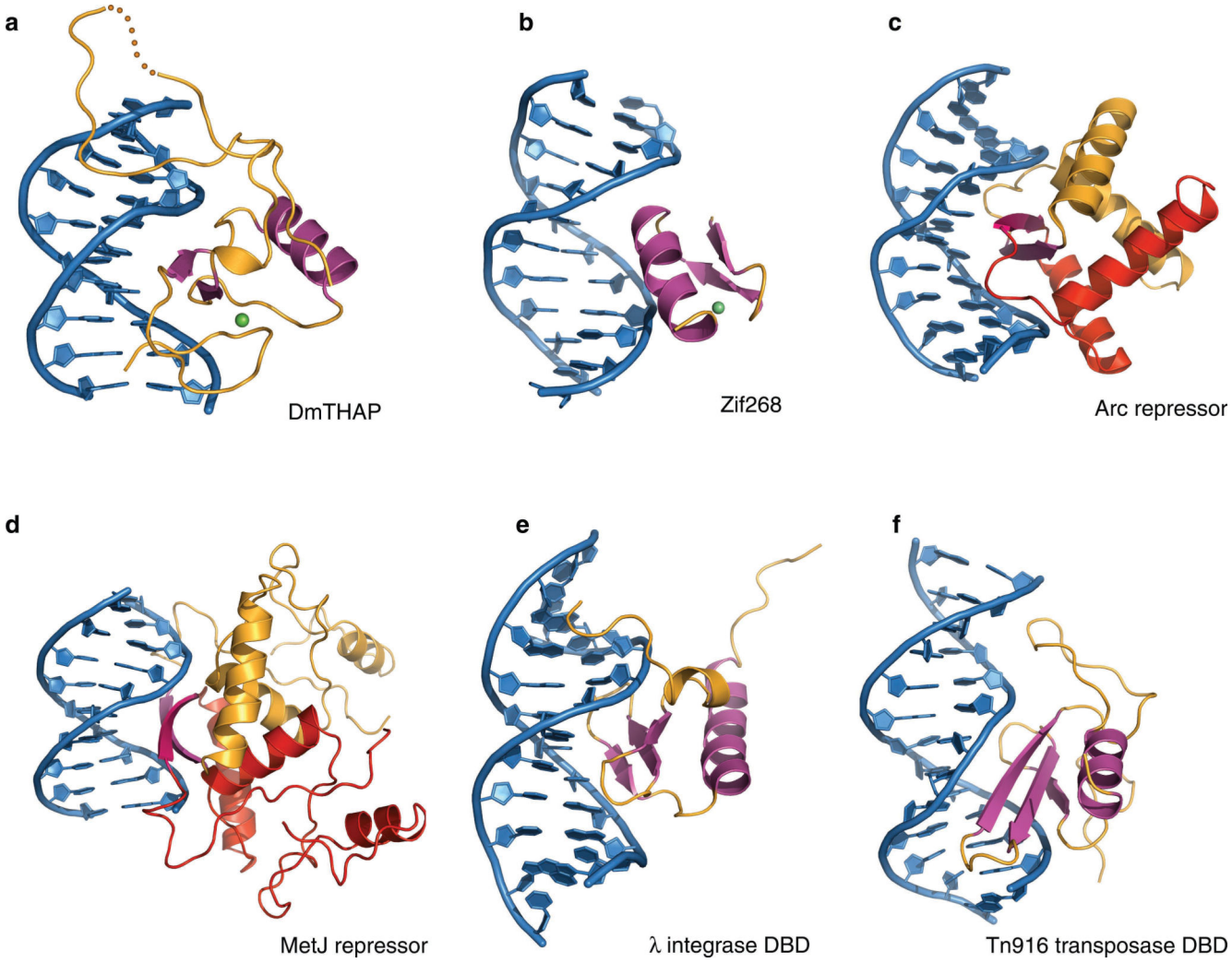


Figure 5.

DmTHAP binds DNA in a manner distinct from the canonical zinc fingers. Cartoon representation of a) DmTHAP and b) Zif268 (PDB ID: 1AAY32) in association with double-stranded DNA. Only a single Zif268 domain is shown. Also distinct from the THAP DNA-recognition interface are the homo-dimeric proteins c) Arc repressor (PDB ID: 1BDT33) and d) MetJ repressor (PDB ID: 1CMA34); colored with each polypeptide in red and yellow, respectively. e) λ-integrase (PDB ID: 2WCC27) and f) Tn916 integrase (PDB ID: 1B6928) DNA-binding domains may be the most similar to THAP domains. Secondary structure color schemes are the same as in Figure 1A.

Table 1

Data collection and refinement statistics

	DmTHAP + 10 bp dsDNA
Data collection	
Space group	$P2_1$
Cell dimensions	
a, *b*, *c* (Å)	28.7, 69.3, 35.1
α, β, γ (°)	90.0, 92.5, 90.0
Wavelength (Å)	0.92
Resolution (Å)	50.0 – 1.74 (1.81 – 1.74)[†]
R_{sym}	0.049 (0.29)
I / σ_I	21.6 (2.6)
Completeness (%)	95.0 (66.8)
Redundancy	3.5 (2.3)
Refinement	
Resolution (Å)	35.1 – 1.74
No. unique reflections	26095 (1841)
R_{work} / R_{free}	17.7 / 21.5
No. atoms	
Protein / DNA	1001
Ligand / ion	1
Water	107
B-factors	
Protein / DNA	30.2
Ligand/ion	25.1
Water	36.6
R.m.s. deviations	
Bond lengths (Å)	0.011
Bond angles (°)	1.03

[*] All data were collected from a single crystal.

[†] Values in parentheses are for highest-resolution shell.