Published in final edited form as:
Science. 2012 February 10; 335(6069): 716–719. doi:10.1126/science.1216211.

The crystal structure of TAL effector PthXo1 bound to its DNA target

Amanda Nga-Sze Mak[1,#], **Philip Bradley**[2,#], **Raul A. Cernadas**[3], **Adam J. Bogdanove**[3], and **Barry L. Stoddard**[1,*]

[1]Division of Basic Sciences, Fred Hutchinson Cancer Research, Center 1100 Fairview Ave. N. A3-025 Seattle WA 98019

[2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. M1-B514 Seattle WA 98109

[3]Department of Plant Pathology and Microbiology, Iowa State University, 351 Bessey Hall, Ames, IA 50011

Abstract

DNA recognition by TAL effectors is mediated by tandem repeats, each 33 to 35 residues in length, that specify nucleotides via unique repeat variable diresidues (RVDs). The crystal structure of PthXo1 bound to its DNA target was determined using high-throughput computational structure prediction and validated by heavy-atom derivatization. Each repeat forms a left-handed, two-helix bundle that presents an RVD-containing loop to the DNA. The repeats self-associate to form a right-handed superhelix wrapped around the DNA major groove. The first RVD residue forms a stabilizing contact with the protein backbone, while the second makes a base-specific contact to the DNA sense strand. Two degenerate N-terminal repeats also interact with the DNA. Containing several RVDs and noncanonical associations, the structure illustrates the basis of TAL effector-DNA recognition.

TAL effectors are proteins that are injected into plant cells by pathogens in the bacterial genus *Xanthomonas.* There they enter the nucleus, bind to effector-specific promoter sequences, and activate the expression of individual plant genes, which can either benefit the bacterium or trigger host defenses (1, 2). In each TAL effector a variable number of tandem amino acid repeats (which are usually 34 residues in length), terminated by a truncated “half repeat,” mediates DNA recognition. Each of the repeats preferentially associates with one of the four nucleotides in the target site (3, 4). The repeats are located centrally in the protein between N-terminal sequences required for bacterial type III secretion and C-terminal sequences required for nuclear localization and activation of transcription (Figure 1a).

The nucleotide specificity of individual TAL effector repeats is encoded by two adjacent residues (located at positions 12 and 13) called the repeat-variable diresidue (RVD) (Figure

*Corresponding author: bstoddar@fhcrc.org 1-206-667-4031, (ph) 1-206-667-3331 (fax).
#These two investigators contributed equally to this study and are co-first authors.

Data deposition
The refined coordinates and corresponding X-ray intensities for the PthXo1-DNA structure have been deposited in the RCSB Protein Data Base (accession code 3UGM).

Supporting online material
Materials and Methods
Figs. S1 to S5
Table S1
References (17-29)

1b and 1c) (4). More than 20 unique RVD sequences have been observed in TAL effectors, but just seven – HD, NG, NI, NN, NS, "N*" (which corresponds to a 33 residue repeat in which the RVD appears to be missing its second residue), and HG – account for nearly 90% of all repeats (5) and respectively specify C, T, A, G/A, A/C/T/G, C/T, and T (3, 4). These relationships enable prediction of targets for existing TAL effectors, and engineering of artificial TAL effectors that bind DNA sequences of choice. Consequently, TAL effectors have received much attention as DNA targeting tools (6).

Nearly all TAL effector binding sites observed in nature are preceded by a T (3, 4). Notably, the protein sequence immediately preceding the canonical TAL effector repeats bears some similarity to the repeat consensus. It has therefore been suggested that this region of the protein may participate in DNA binding by forming a cryptic repeat structure that specifies the T (7).

A recent NMR structural study of 1.5 repeats of TAL effector PthA, and an accompanying SAXS study of the entire protein, indicated that an isolated TAL effector repeat is largely α-helical, similar to a tetratricopeptide (TPR) fold, and that the full-length protein compacts upon DNA binding (8). However, in that study it was unclear to what extent the structure of repeats in the context of the entire protein might differ from an isolated repeat, and the manner in which individual repeats associate with contiguous DNA base pairs was not resolved.

A protein construct corresponding to residues 127 to 1149 of the 23.5 repeat TAL effector PthXo1 from the rice pathogen *Xanthomonas oryzae* (Figure 1 and Supplementary Figure S1) was crystallized bound to a 36 base pair DNA duplex (Supplementary Table S1) containing the target sequence found in the rice genome along with flanking sequences ending in short 3′ overhangs. The structure was determined using a high-throughput computational approach in which structural models built with the Rosetta software package (9) were iteratively refined and selected, guided by molecular replacement searches (Supplementary Figure S2). The best model was subsequently validated using a variety of model-free features of electron density, including anomalous difference peaks calculated from a selenomethionyl derivative (Supplementary Figure S3). The final structure was refined to 3.0 Å resolution to values for R_{work}/R_{free} of 0.264/0.294 and excellent geometry (Table 1).

The structure consists of a relatively unperturbed B-form DNA duplex, with 23 consecutive bases of the target site intimately engaged in the major groove by a superhelical arrangement of TAL effector repeats (Figure 2). The overall dimensions of the protein-DNA complex are approximately 60 Å x 60 Å x 90 Å. The quality of the electron density is excellent from repeat 1 through the middle of repeat 22, and then becomes less well defined.

All of the repeats in the DNA-bound PthXo1 structure form highly similar two-helix bundles (Figure 1c). The helices span positions 3 to 11 and 14 to 33, locating the RVD in a loop between them. A proline located at position 27, creates a kink in the second helix that appears to be critical for the sequential packing and association of tandem repeats with the DNA double helix. The packing of consecutive helices within and between individual repeats is left-handed, in contrast to the right-handed packing of helices found in TPR proteins (10). The modular architecture of the TAL effector repeats is reminiscent of the mitochondrial transcription terminator mTERF (11) and the RNA-binding attenuation protein TRAP (12); however, interactions of those proteins with their nucleic acid targets are structurally distinct from those of TAL effectors with DNA and lack modular correspondence to single nucleotides.

Sequence-specific contacts of PthXo1 to the DNA are made exclusively by the second residue in each RVD to the corresponding base on the sense strand. In contrast, the side chain at the first position of each RVD contacts the backbone carbonyl oxygen of position 8 in each repeat, constraining the RVD-containing loop (Figure 3). Additional, nonspecific contacts to the DNA are made by a lysine and glutamine found at positions 16 and 17. The average root-mean-square-deviation between backbone atoms in any two repeats in the PthXo1 structure is approximately 0.8 A for all atoms; it is slightly greater for the 33-residue "N*" repeats, which are missing one residue in the RVD loops (Supplementary Figure S4). The positions within the core of individual repeats are occupied entirely by small aliphatic residues, while several positions in the interface between repeats correspond to polar residue pairs.

The PthXo1-DNA structure displays five HD-containing repeats (all aligned to cytosines), four 'NG' repeats and one 'HG' repeat (aligned to thymines), one additional 'NG' repeat aligned to cytosine, seven 'NI' repeats (aligned with four adenosines and three cytosines), two 'NN' repeats (both opposite a guanosine), and two 'N*' repeats paired to cytosines (Figure 1a). The observed contacts by individual repeats (Figure 3) correlate well with their specificity and fidelity (or lack thereof) that have been described via bioinformatic and genetic analyses. The sole NS in PthXo1 and one additional N* are located in the last full repeat and the half repeat respectively, which are disordered in the structure.

In the 'HD' RVDs, the aspartate residue makes van der Waals contacts with the edge of the corresponding cytosine base and a hydrogen bond to the cytosine N4 atom. Contacts between cytosine bases in protein-DNA complexes and charged acidic side chains, which exclude alternative base identities via physical and electrostatic clash, have been observed in a wide variety of solved sequence-specific protein-DNA complexes (13).

Both the 'NG' and 'HG' repeats make a contact in which the backbone alpha carbon of the glycine residue forms a nonpolar van der Waals interaction with the methyl group of the opposing thymine base (average distance ~ 3.3 Å). At the one position where an NG is aligned opposite a cytosine base, the backbone carbonyl and alpha-carbon of the same glycine residue displays a less favorable, far more distant contact (~ 6 Å).

The second asparagine residue in the 'NN' RVDs is positioned to make a hydrogen bond with the N7 nitrogen of an opposing guanine base. This RVD associates with either guanosine or adenine with roughly equal frequency (3, 4, 14); the availability of an N7 nitrogen in either purine ring appears to explain that observation (13).

PthXo1 contains two 33 residue 'N*' repeats (7 and 22). Since RVDs are followed immediately by two conserved glycine residues, this repeat is equivalent to an 'NG' repeat in which one of those glycine residues is missing. The crystal structure indicates that the deletion results in a truncated RVD loop that extends less deeply into the DNA major groove, with the glycine at position 13 located a considerable distance (over 6 Å) from the corresponding sense strand base. Consistent with this observation, the observed specificity of the 'N*' repeat is relatively lax (4).

Finally, NI, which is the second most common RVD overall, accounting for roughly 20% of all TAL effector repeats, occurs seven times in PthXo1, and displays an unusual contact pattern to adenosine or cytosine bases. The aliphatic side chain of the isoleucine residue is observed to make non-polar van der Waals contacts to C8 (and N7) of the adenine purine ring, or to C5 of the cytosine pyrimidine ring. These contacts would appear to necessitate desolvation of at least one polar atom in the adenosine ring, without the formation of a compensating hydrogen bond, and might therefore reasonably be expected to represent a reduced affinity interaction.

N-terminal to the canonical repeats, the PthXo1 structure reveals two degenerate repeat folds that appear to cooperate to specify the conserved thymine that precedes the RVD-specified sequence (Figure 4). We have designated these as the 0^{th} and -1^{st} repeats. Residues 221 to 239 and residues 256 to 273 each form a helix and an adjoining loop that resembles helix 1 and the RVD loop in the canonical repeats; the remaining residues in each region are poorly ordered. Those two N-terminal regions converge near the 5′ thymine base, with the indole ring of tryptophan 232 (in the -1^{st} repeat) making a van der Waals contact with the methyl group of that base. Mutation of the thymine reduces TAL effector activity at the target (3, 15). Tryptophan 232, as well as the surrounding residues, is highly conserved across available, intact TAL effector sequences. Some TAL effectors efficiently target sequences preceded by a cytosine rather than a thymine (14, 16). Though less favorable, the packing of tryptophan 232 would be expected to accommodate this substitution.

In addition to revealing folding and interactions of the N-terminal cryptic repeats with the 5′ end of the DNA target site and illustrating the functions of the six most common repeat types in TAL-effector-DNA recognition, the structure provides a basis for prediction of structures that are not represented. For example, an alignment of the 35 residue repeat type found in some TAL effectors with the more common 34 residue repeat type found in PthXo1 (Supplementary Figure S5) indicates that the additional residue (a proline) at position 33 would be located within the relatively disordered turn region that connects the helices of one repeat to the next. The 35 residue repeat therefore can be predicted to be functionally indistinguishable from the 34. Likewise, although the sole ‘NS’ repeat in PthXo1 is in an apparently disordered part of the protein-DNA complex, the overall homogeneity of the repeat structures and the consistent role of the first RVD residue in stabilizing the RVD loop to facilitate specific contacts of the second residue with the DNA should make it possible to computationally model the potential nucleotide interactions of NS, as well as those of rare or artificial RVDs.

The protein-DNA complex studied leaves some questions unanswered, such as the structure of the N and C-terminal portions of TAL effectors that are respectively required for translocation and interaction with host transcriptional machinery. As well, because of the observed disorder at either end, it does not yet precisely define the minimal TAL effector DNA binding domain. However, by demonstrating the essential features that accomplish interaction specificity, the structure provides a foundation for more accurately predicting and efficiently exploiting TAL effector-DNA targeting. More fundamentally, it reveals the hitherto enigmatic structural nature of a simple solution that an important group of pathogens has evolved to manipulate host gene expression in a specific yet highly adaptable manner.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was funded by NIH (grants RL1 CA833133 to BLS, R01GM098861 to BLS and AJB, and R01 GM088277 to PHB), NSF grant 0820831 to AJB, and a Searles Scholars Fellowship to PHB. A.N-S.M. was supported by a training grant from the Northwest Genome Engineering Consortium. The authors thank the staff of ALS beamline 5.0.2 and Lindsey Doyle, Betty Shen, Ryo Takeuchi, Jill Bolduc, and Clarice Schmidt for technical assistance and advice, Tom Edwards and Matt Clifton for collecting SeMet data, and Carl Pabo for helpful discussion.

REFERENCES AND NOTES

1. Kay S, Hahn S, Marois E, Hause G, Bonas U. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. Science. 2007; 318:648. [PubMed: 17962565]
2. Romer P, et al. Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. Science. 2007; 318:645. [PubMed: 17962564]
3. Boch J, et al. Breaking the code of DNA binding specificity of TAL-type III effectors. Science. 2009; 326:1509. [PubMed: 19933107]
4. Moscou MJ, Bogdanove AJ. A simple cipher governs DNA recognition by TAL effectors. Science. 2009; 326:1501. [PubMed: 19933106]
5. Boch J, Bonas U. *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. Annual Review of Phytopathology. 2010; 48:419.
6. Bogdanove AJ, Voytas DF. TAL effectors: customizable proteins for DNA targeting. Science. 2011; 333:1843. [PubMed: 21960622]
7. Bogdanove AJ, Schornack S, Lahaye T. TAL effectors: finding plant genes for disease and defense. Current Opinion in Plant Biology. 2010; 13:394. [PubMed: 20570209]
8. Murakami MT, et al. The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction. Proteins: Structure, Function, and Bioinformatics. 2010; 78:3386.
9. Leaver-Fay A, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011; 487:545. [PubMed: 21187238]
10. D'Andrea LD, Regan L. TPR proteins: the versatile helix. Trends Biochem Sci. 2003; 28:655. [PubMed: 14659697]
11. Jimenez-Menendez N, et al. Human mitochondrial mTERF wraps around DNA through a left-handed superhelical tandem repeat. Nature Struct Mol Biol. 2010; 17:891. [PubMed: 20543826]
12. Hopcroft NH, et al. The interaction of rna with trap: the role of triplet repeats and separating spacer nucleotides. J Mol Biol. 2004; 338:43. [PubMed: 15050822]
13. Rohs R, et al. Origins of specificity in protein-DNA recognition. Annu Rev Biochem. 2010; 79:233. [PubMed: 20334529]
14. Miller JC, et al. A TALE nuclease architecture for efficient genome editing. Nature Biotech. 2011; 29:143.
15. Romer P, et al. Promoter elements of rice susceptibility genes are bound and activated by specific TAL effectors from the bacterial blight pathogen, Xanthomonas oryzae pv. oryzae. New Phytol. 2010; 187:1048. [PubMed: 20345643]
16. Yu Y, et al. Colonization of rice leaf blades by an African strain of Xanthomonas oryzae pv. oryzae depends on a new TAL effector that induces the rice nodulin-3 Os11N3 gene. Molec Plant-Microbe Interactions. 2011; 24:1102.
17. Yang B, Sugio A, White FF. Os8N3 is a host disease-susceptibility gene for bacterial blight of rice. Proc Natl Acad Sci U S A. 2006; 103:10503. [PubMed: 16798873]
18. Bogdanove AJ, et al. Two new complete genome sequences offer insight into host and tissue specificity of plant pathogenic Xanthomonas spp. J Bacteriol. 2011; 193:5450. [PubMed: 21784931]
19. Mak AN, Lambert AR, Stoddard BL. Folding, DNA recognition, and function of GIY-YIG endonucleases: crystal structures of R.Eco29kI. Structure. 2010; 18:1321. [PubMed: 20800503]
20. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods in Enzymology. 1997; 276:307.
21. McCoy AJ, et al. Phaser crystallographic software. J Appl Crystallogr. 2007; 40:658. [PubMed: 19461840]
22. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010; 66:213. [PubMed: 20124702]
23. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr. 2004; 60:2126. [PubMed: 15572765]

24. Winn MD, Murshudov GN, Papiz MZ. Macromolecular TLS refinement in REFMAC at moderate resolutions. Methods Enzymol. 2003; 374:300. [PubMed: 14696379]

25. Laskowski RJ, Macarthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystall. 1993; 26:283.

26. Vriend G. WHATIF: a molecular modeling and drug design program. J Mol Graph. 1990; 8:52. [PubMed: 2268628]

27. Rose PW, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. 2011; 39:D392. [PubMed: 21036868]

28. Maiti R, Domselaar GHV, Zhang H, Wishart DS. SuperPose: a simple server for sophisticated structural superposition. Nucleic Acids Res. 2004; 32:W590. [PubMed: 15215457]

29. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystall. 1976; 32A: 922.

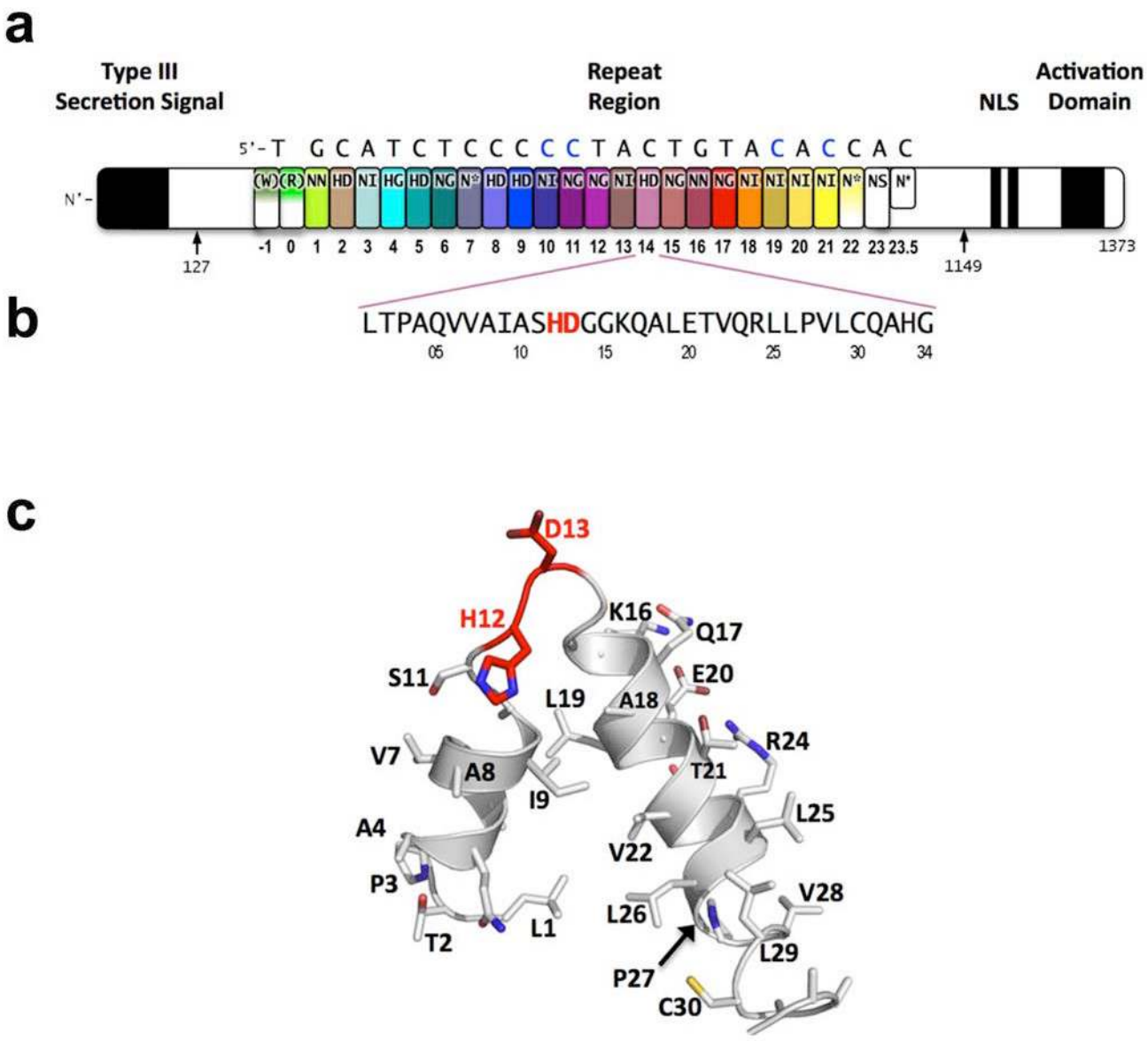

Figure 1. Domain organization of PthXo1 and structure of a single TAL effector repeat

TAL effectors contain N-terminal signals for bacterial type III secretion, tandem repeats that specify the target nucleotide sequence, nuclear localization signals, and a C-terminal region that is required for transcriptional activation. PthXo1 contains 23.5 canonical repeats (color coded to match Figure 2) that contact the DNA target found in the promoter of the rice *Os8N3* gene (17). Blue bases correspond to positions in the target where the match between protein and DNA differs from the optimal match specified by the recognition code (3,4). Arrows indicate the start and end of the crystallized protein construct. In the structure, repeats 22 to 23.5 are poorly ordered, as are the C-termini of the two N-terminal cryptic repeats. The sequence and structure of a representative repeat (#14) is shown; RVD residues (HD) that recognize cytosine are red.

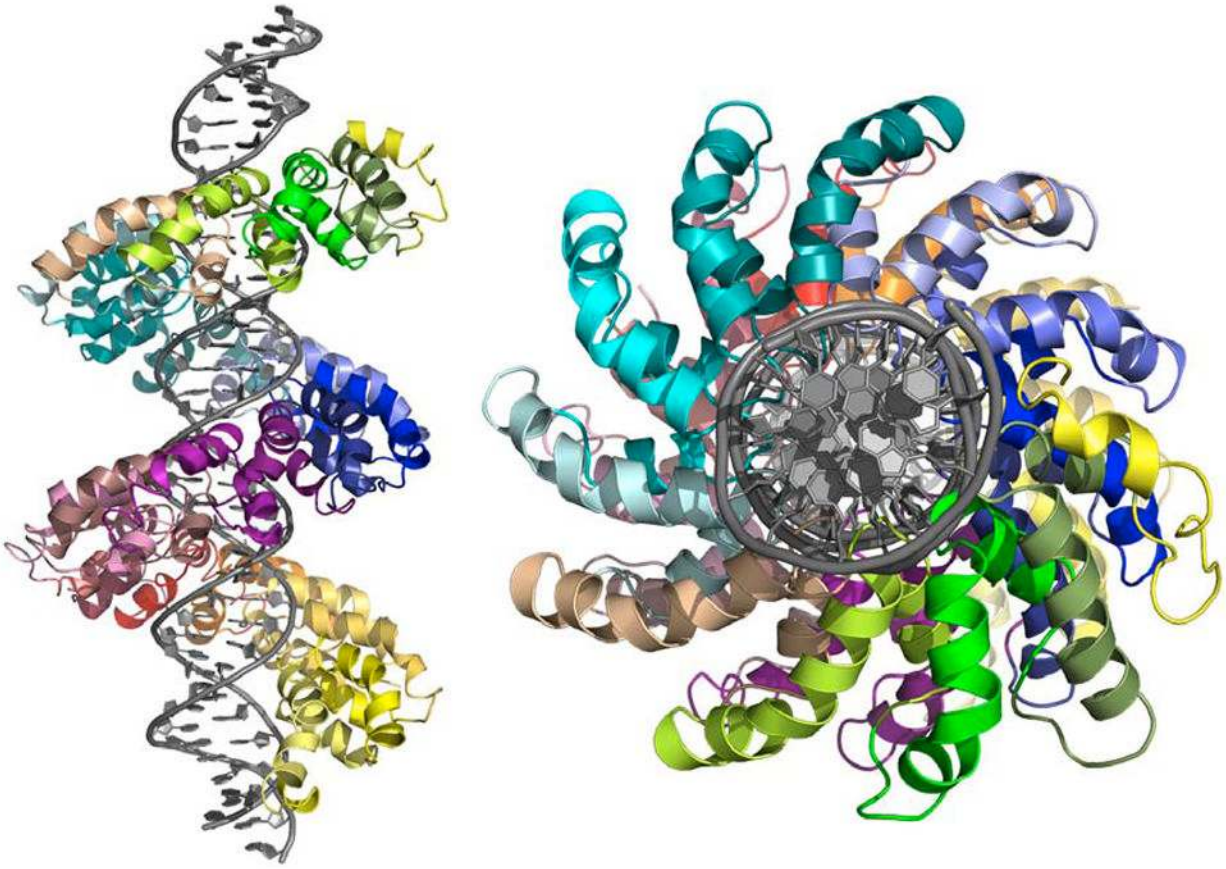

Figure 2.
Structure of the PthXo1 DNA binding region in complex with its target site. The coloring of individual repeats matches the schematic in Figure 1.

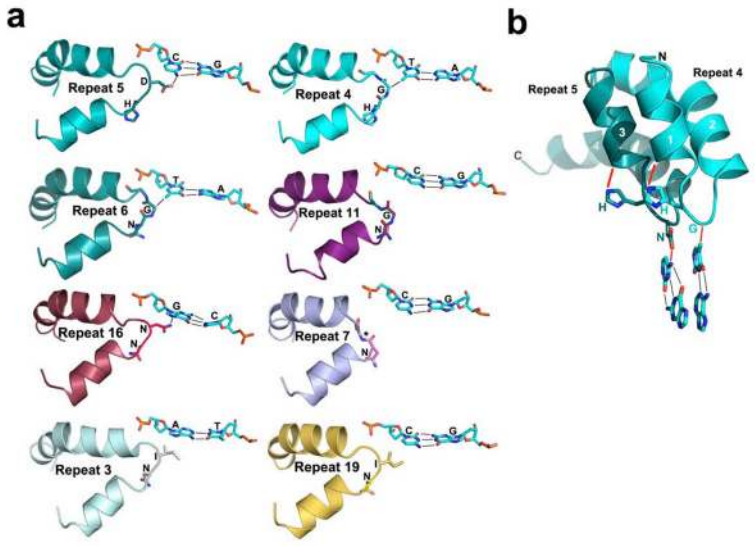

Figure 3. Topology and contacts between TAL effector repeats and DNA bases

Panel a: 8 distinct combinations of RVDs and DNA bases are observed in the structure. HD forms a steric and electrostatic contact with cytosine; HG and NG both form nonpolar interactions between the glycine α-carbon and the thymine methyl group. A "mismatch' between NG and a cytosine results in a longer distance from the RVD to the base. NN associates with either guanine (repeat 16) or with adenine (which would interact with the same N7 nitrogen of the purine base). NI forms a desolvating interface with either adenine (repeat 3) or cytosine (repeat19). The reduction in loop length by one residue in the 'N*' RVD (repeat 7) results in an increased distance to the base. **Panel b:** Two adjacent repeats form a tightly packed left-handed bundle of helices that position the second amino acid of each RVD in proximity to corresponding consecutive bases in an unperturbed B-form DNA duplex. The first residue of each RVD (position 12, either His or Asn) forms H-bonds to the backbone carbonyl oxygen of amino acid position 8 of the same repeat.

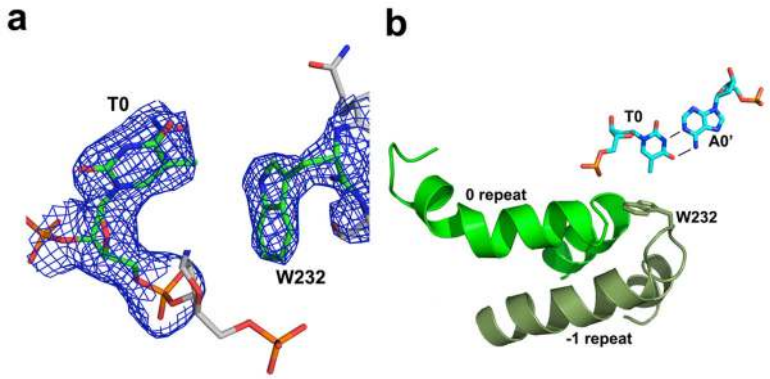

Figure 4. N-terminal cryptic repeats and contacts with 5′ thymine

a: 2Fo-Fc electron density maps contoured around thymine at position '0' and tryptophan 232 in the '−1' repeat. **b:** Residues 221 to 239 and residues 256 to 273 each form a helix and an adjoining loop that resembles helix 1 and the RVD loop in the canonical repeats; the remaining residues in each region are poorly ordered. W232 forms a non polar van der Waals contact with the methyl carbon of the thymine base at position 0.

Table 1

Crystallographic data and refinement statistics

DATA STATISTICS		
Dataset	**WT**	**SeMet**
X-ray source	ALS 5.0.2	APS 21-ID-F
Wavelength(Å)	1.177	1.378
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit Cell (Å)	a = 95.6	a = 100.7
	b = 248.5	b = 247.8
	c = 54.6	c = 54.2
Resolution (Å)[a]	50-3.0 (3.11-3.0)	50-4.0 (4.14 - 4.0)
R_{merge} (%)	0.121 (0.431)	0.087 (0.139)
I/σ (I)	9.3 (3.5)	10.3 (3.7)
Redundancy	5.6 (4.9)	4.9 (5.0)
Completeness (%)	96.6 (90.4)	95.8 (97.7)
Mosaicity (°)	0.8	0.8
Unique Reflections	25841	11591
REFINEMENT		
R_{work}	0.264	
R_{free}	0.294	
Protein Atoms	6086	
DNA Atoms	1552	
Heteroatoms (waters)	216	
Rmsd bond lengths (Å)	0.021	
Rmsd bond angles (°)	2.4	
Average B factor ($Å^2$)	85.1	
Ramachandran (% core, allowed, generous, disallowed)	73.6%, 26.4%, 0%, 0%	