A Simple Cipher Governs DNA Recognition by TAL Effectors

Matthew J. Moscou and Adam J. Bogdanove*

TAL (transcription activator–like) effectors of plant pathogenic bacteria in the genus *Xanthomonas* contribute to disease or trigger defense by binding host DNA and activating effector-specific host genes (*1–5*). Specificity depends on a variable number of imperfect, typically 34, amino acid repeats (*6*). Polymorphism is primarily at repeat positions 12 and 13, which we call the repeat-variable diresidue (RVD). We show that the RVDs of TAL effectors correspond directly to the nucleotides in their target sites, one RVD to one nucleotide, with some degeneracy and no apparent context dependence. Our finding explains TAL effector specificity, enables target site prediction, and opens prospects for use of these proteins in research and biotechnology.

Several considerations suggested that RVDs would specify DNA targets. Structural predictions place residues 12 and 13 on a solvent-exposed surface (*6*). Binding of TAL effector AvrBs3 to the *UPA20* promoter is mediated by its repeats (*3*). And, the AvrBs3 target *Bs3* is activated also by TAL effector AvrHah1, which despite overall dissimilarity has a similar sequence of RVDs (*7*). We thus anticipated a one-to-one correspondence between RVDs and contiguous nucleotides in the target site. To test this hypothesis, for each of 10 known TAL effector–target gene promoter pairs, we scanned for RVD-nucleotide alignments with minimal entropy.

Low entropy sites were present in each promoter. However, for AvrBs3, only one mapped to the 54–base pair (bp) *UPA20* promoter fragment that is sufficient and necessary for activation, and it coincided with the *UPA* box common to genes directly activated by 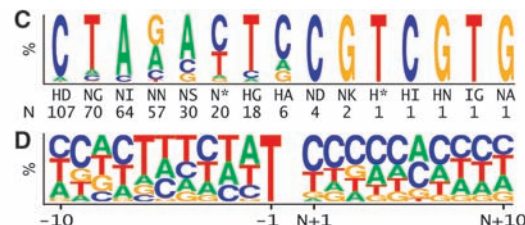AvrBs3 (*3*). For effectors PthXo1 and AvrXa27, only one site each overlapped a polymorphism between the activated and nonactivated alleles of their respective targets, *Os8N3* and *Xa27*. Across the alignments at these three sites, RVD-nucleotide associations were consistent. Remaining alignments were selected on the basis of those associations, resulting in exactly one site per TAL effector–target pair (Fig. 1). A T precedes each site.

To assess the specificity conferred by the RVD-nucleotide associations, we generated a weight matrix based on their frequencies and scanned about 60,000 annotated promoters in rice (cv. Nipponbare) for best matches to the five TAL effectors from the rice pathogen *X. oryzae*. For four, the experimentally identified target gene was the best or nearly best match. Better matches were not preceded by a T, were not represented on the microarray used to identify the target, or lacked introns and expressed sequence tag evidence. Scanning the reverse complement promoter sequences yielded no better scoring alignments than the forward sites for the known targets. The known target of the fifth effector, AvrXa27, is the disease resistance gene *Xa27* (*1*). The poorer rank for this match (5368) may reflect a suboptimal or calibrated host adaptation. Better scoring sites likely comprise genes targeted by AvrXa27 for pathogenesis.

We obtained 10 more alignments by scanning all rice promoters with 40 additional *X. oryzae* TAL effectors. We retained the best alignments for which the downstream gene was activated during infection based on public microarray data (http://PLEXdb.org, accession OS3). Here too a T precedes each site, and no reverse-strand sites scored better. The RVD-nucleotide alignments constitute a strikingly simple cipher (Fig. 1C).

There is some degeneracy in the cipher. Strong associations may represent binding anchors. Weak ones may provide flexibility or result from neighbor effects. Analysis of the latter, however, yielded no signal, suggesting context independence.

Sequences flanking the 20 target sites tend to be C-rich after the site and G-poor throughout (Fig. 1D). The sites usually begin within 60 bp upstream of the annotated transcriptional start. None are closer than 87 bp to the translational start.

Annotation of TAL effector targets, now feasible, will aid identification of host genes important in disease. Adding TAL sites may enhance efficacy and durability of resistance genes like *Xa27*. TAL effectors may also be useful for targeted gene activation as well. Whether TAL effectors function in nonplant cells or are amenable to protein fusion are unknown. Elucidating their interaction with host transcriptional machinery and their structure bound to DNA are important next steps in defining the function and utility of these proteins.

Fig. 1. The TAL effector–DNA recognition cipher. (**A**) A generic TAL effector showing the repeat region (open boxes) and a representative repeat sequence with the RVD underlined. (**B**) Best pattern matches (low-entropy alignments) for several TAL effector RVD and target gene promoter sequences. An asterisk indicates a deletion at residue 13. (**C**) RVD-nucleotide associations in the alignments in (B) and 10 more alignments obtained by scanning all rice promoters with 40 additional *X. oryzae* TAL effectors, retaining for each effector the best alignment for which the downstream gene was activated during infection. (**D**) Flanking nucleotide frequencies for the 20 TAL effector target sites. Positions are relative to the 5′ end of the target site; *N*, length of target site. Logos were generated using WebLogo (http://weblogo.berkeley.edu/).

References and Notes

1. K. Gu *et al.*, *Nature* **435**, 1122 (2005).
2. B. Yang, A. Sugio, F. F. White, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10503 (2006).
3. S. Kay, S. Hahn, E. Marois, G. Hause, U. Bonas, *Science* **318**, 648 (2007).
4. A. Sugio, B. Yang, T. Zhu, F. F. White, *Proc. Natl. Acad. Sci. USA* **104**, 10720 (2007).
5. P. Römer *et al.*, *Science* **318**, 645 (2007).
6. S. Schornack, A. Meyer, P. Römer, T. Jordan, T. Lahaye, *J. Plant Physiol.* **163**, 256 (2006).
7. S. Schornack, G. V. Minsavage, R. E. Stall, J. B. Jones, T. Lahaye, *New Phytol.* **179**, 546 (2008).
8. We thank J. Boch, T. Lahaye, and U. Bonas for useful discussion and K. Dorman, N. Lauter, A. Miller, and S. Whitham for suggestions. This work was funded by the NSF.

Department of Plant Pathology and Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, USA

*To whom correspondence should be addressed. E-mail: ajbog@iastate.edu