TALE: an emerging tool for genome editing and genetic engineering

Basavraj Khanppnavar, Shashi Baruah and Suvendra Kumar Ray

Transcription activator-like effectors (TALEs) are one of the important virulence tools used by a genus of plant pathogen bacteria belonging to *Xanthomonas* spp. Naturally TALEs function as eukaryotic transcription factors that bind to specific host promoter sequence and thus regulate gene expression. In brief, they re-programme the host cells resulting in improved bacterial multiplication inside the host cells. TALEs are injected into the host cell cytoplasm via the type III protein secretory system, otherwise known as Hrp (hypersensitive response and pathogenicity) secretion system in plant pathogenic bacteria[1–3]. TALE-mediated gene expression is also found to be responsible for disease symptoms in plants. These effectors have been predominantly found in *Xanthomonas*, but its less related homologues also exist in *Ralstonia solanacearum*, another plant pathogenic bacterium[4]. Some examples of TALE and its target host are listed in Table 1. It is obvious from Table 1 that such an important virulence factor is almost restricted to *Xanthomonads* (except one in *Rastonia solanacearum*), though type III protein secretory system is found in many plant and animal pathogenic bacteria. So the question regarding evolution of this kind of virulence factor will be of significant interest considering the fact that the mechanism of TALE is yet to be discovered in any eukaryote.

What makes TALE interesting to researchers is its novel mechanism of recognition of specific DNA sequences, unlike other DNA-binding protein motifs such as the zinc finger, the helix–turn–helix and the leucine zipper which recognize target DNA sequence based on consensus sequence. TALE proteins have DNA recognition mechanism mediated by two adjacent amino acids, referred to as the 'repeat variable di-residue' (RVD)[5,6]. The relationship between the preferred binding site of a TALE and its successive RVDs constitutes a simple code, with each repeat independently specifying its targeted base (Figure 1). This DNA recognition code was first cracked computationally by Moscou and Bogdanove[5], by screening for non-random alignments between the variable di-amino acids in the TALEs and DNA sequences of target promoters. Presence of DNA recognition codons has made it easier to find the target DNA sequence that the TALE is going to bind by directly looking at the amino acid sequence of TALE. Recently, many scientists have also validated these codes experimentally.

Members of this effectors family are highly conserved and differ mainly in the amino acid sequence and number of RVD repeats in the central variable domain. The number and order of RVD repeats in a TALE DNA-binding domain determine its specific activity (Figure 1). The number of RVD repeats varies from 1.5 to 33.5 in TALE DNA-binding domain. There are more than 20 different RVD repeats in TALE, but only seven are found to be more common and important. These are HD, NG, NI, NN, NS, HG and 'N*' which specify C, T, A, G/A, A/C/T/G, C/T and T along the DNA strand respectively. HG accounts for nearly 90% of all repeats. 'N*' corresponds to a 33-residue repeat in which RVD appears to be missing its second residue[7]. As there are 20 possible amino acids, the different possibilities of RVDs are $20^2 = 400$ and this combination can specify only four bases. Thus this can be the obvious reason for the ambiguity and degeneracy observed in TALE specificity.

Recently, studies[7,8] have revealed the mechanism of DNA recognition based on nuclear magnetic resonance and X-ray crystallographic studies on dHax and PthXo1. These studies have indicated that the TALE repeats are α-helical, similar to tetratricopeptide (TPR) fold[7]. All repeats of TALE associate to form a right-handed super-helix. This super-helical structure tracks along the DNA duplex and wraps itself along major groove of DNA. The two variable amino acids in the RVD loops are positioned in close proximity to the sense strand and play different biochemical roles in the major groove of DNA. One of these amino acids makes specific contact with a nucleotide in the DNA-sense strand, while the other stabilizes the contact between the DNA and the protein[7,8]. Thus TALE fits well into DNA and it also does not change the structure of DNA when bound to it (Figure 2).

Among more than 20 codes identified in TALE RVDs, only few are found more frequently. The RVD HD $\rightarrow$ C (HD recognizes cytosine); the carboxylate oxygen atom of Asp accepts an H bond from the amine group of cytosine in

Table 1. Some examples of TALE and their corresponding target host and host genes

Bacterial species	TAL effector	Target genes	Host species	Reference
Xanthomonas campestris pv. *vesicatoria*	AvrBs3	*upa20/Bs3*	Pepper	13
	AvrBs4	*MtN3*	Tomato	
X. gardneri	AvrHah1	*W,Bs3*	Pepper	14, 15
X. oryzae pv. *oryzicola*	Tal-C1c	*Os11N3*	Rice	13, 16
X. oryzae pv. *oryzae*	Tal9a, PthXo1, Avrxa5	*Os8N3, EV TFIIAγ1*	Rice	17
X. campestris pv. *armoraciae*	Hax4	*Bs4,ArtX1-10,CN*	Tomato	6, 15
X. campestris pv. *manihotis*	pTHB	*H,Bs3*	Cassava	15, 18
X. campestris pv. *malvacearum*	Avrb6	*W, MtN3* family	Cotton	13, 15, 19
Ralstonia solanacearum	Brg11	*Bs3*	Broad host range	20

CN, Enhance chlorosis and necrosis tissue; W, Increase in water-soaked appearance of tissue; H, Hypertrophy, cell enlargement; EV, Major enhancement of virulence on the basis of increased bacterial populations in plants.

Table 2. Some recent successful advances in TALE-mediated genome modification

Organism	Modification	Reference
Human	Treatment of inherited disease, Friedreichs ataxia by gene expression of frataxin	21
Zebra fish	Heritable *in vivo* gene knockout	12
Rice	Xoo-resistant rice lines by disruption of susceptible genes	22
Rat	Heritable IgM locus disruption	23
Arabidopsis	TALEN directed to ADH1 in *Arabidopsis thaliana* protoplasts	24
Silkworm *Bombyx mori*	Endogenous *BmBLOS2* gene disruption	25
HIV-derived DNA–RNA hybrid	Potential control of DNA replication and retroviral infections by a specific DNA–RNA hybrid recognition employing TALENs	8

Xoo, *Xanthomonas oryzae* pv. *oryzae*.

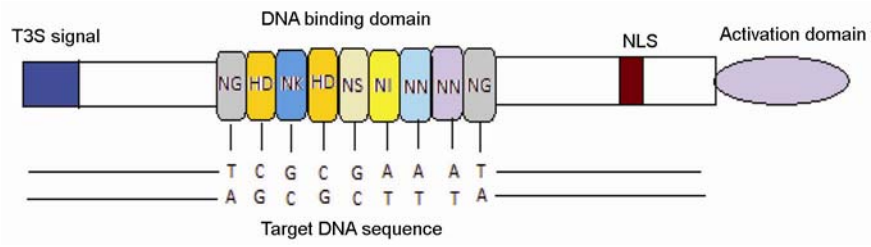

Figure 1. Systematic alignment of TALE showing its DNA-binding domain, type III secretory (T3S) signal sequence, nuclear localization signal sequence (NLS) and transcriptional activation domain. Each box in a DNA-binding domain indicates a single repeat variable di-residue (RVD) repeat specifying its target base, e.g. NG recognizes nucleotide base thymine and while HD recognizes nucleotide base cytosine. Each RVD in α-helical structure covers an axial distance of around 3 Å, which is close to 3.4 Å axial distance of nucleotide base in the DNA helix.

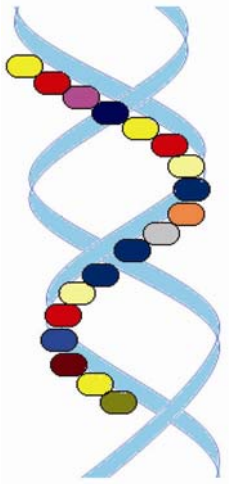

Figure 2. Systematic representation of TALE DNA-binding domain bound to its target DNA strand. Each colour box indicates a different RVD. All the RVD repeats constitute a super-helical structure which wraps around the major groove of DNA and exposes RVDs to bind DNA specifically.

TALE repeats. In the case of NS $\rightarrow$ A, the hydroxyl group of Ser donates an H-bond to the N7 atom of adenine. Compared with HD, NS is nonselective in that it can recognize all four bases. Similar to adenine, guanine also contains a N7 atom, which is likely recognized by Ser in the same manner[7,8]. A more recent structural and biophysical analysis of TALE, dHax3 showed that four additional cryptic repeats are formed immediately upstream of the central repeat region, and that this region provides the initial binding energy required for high-affinity target binding and sequence-specific recognition[9].

Strategies of genome manipulation have great application in both basic and applied research, including therapeutic interventions for genetic diseases, generation of stress-resistant crops and high-yielding plants and animals. But the success of genetic manipulation and genetic engineering was always limited by the difficulty in targeting specific genes and modifying them efficiently. The novel mechanism of DNA recognition by simple code between amino acids in TALEs and DNA bases in their target sites has created great interest among scientists for its application for *in vivo* and *in vitro* site-specific DNA and genome modification (Figure 3). Numerous groups have designed artificial TALEs capable of recognizing new DNA sequences in a variety of experimental systems. Such engineered TALEs have been used to create artificial transcription factors that can be used to target and activate endogenous genes in tomato, *Arabidopsis thaliana* and human cells.

The engineered TALEs can be fused to nuclease, such as catalytic domain of Fok1 to produce TAL effector nuclease or TALEN. These nucleases can be used for creating double strand break (DSB) at specific site. Further DSBs can be repaired by the donor-free and error-prone non-homologous end-joining repair (NHEJ), resulting in mutagenesis with small deletions or insertions that disrupt the gene of interest. Alternatively, DSBs can also be used for insertion or addition by homologous recombination with donor template for functional correction of gene (Figure 4). Recently, TALEN-based technology was used successfully to generate rice variety resistant to pathogen, *Xanthomonas oryzae* by disrupting the disease susceptible(S) gene *Os11N3* in rice.

Earlier, zinc finger domains fused with endonuclease (ZFNs) were used for generating site-directed DSBs, whose mechanism of recognition is based on nucleotide triplets. But TALENs have proved to be more efficient and more versatile than ZFNs[10].

Alternatively, a TALE constructed with activator or repressor as a functional can be used to regulate expression of a particular gene. Simultaneously, the other possible theoretical functional domains can be chromatin-modifying proteins such as cytidine deaminases, histone acetyltransferases or deacetylases or DNA methyltransferases. Recently, engineered TALEs are being successfully used for genome modification and TALEs are also portable to diverse cell lines[11]. Some of the successful advances in TALE-

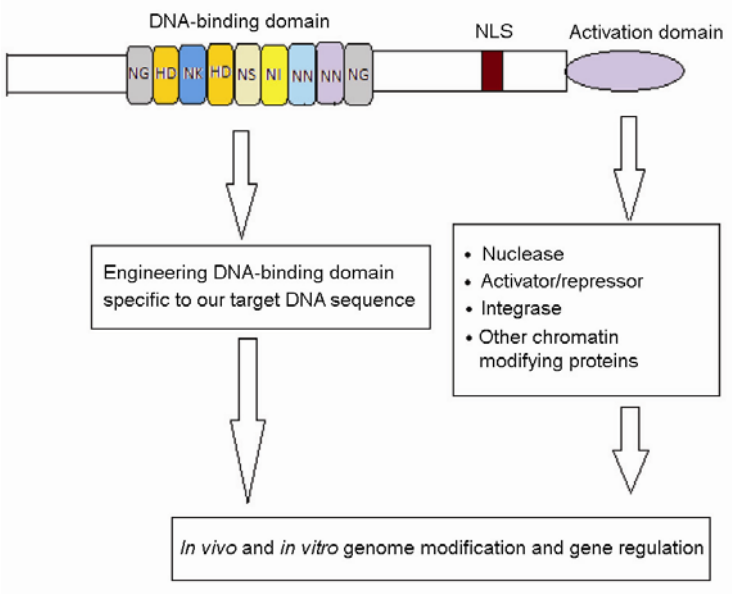

Figure 3. Systematic representation of possible modifications of TALE.

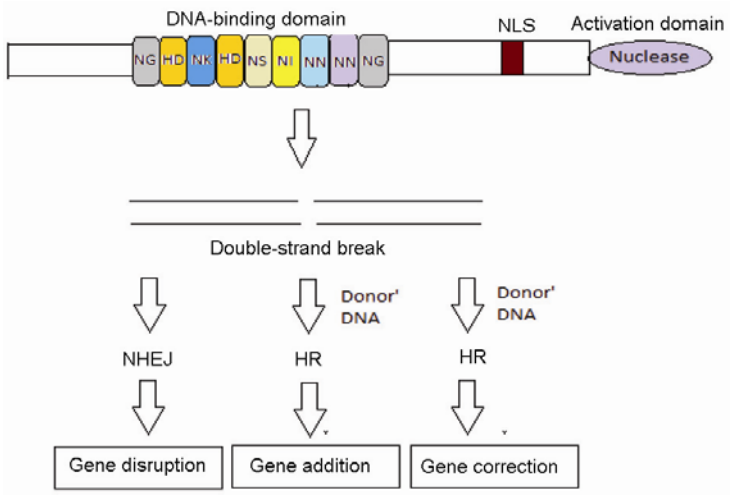

Figure 4. Potential genome editing application based on generation of double-strand break by TALE nuclease for gene disruption, gene addition and functional correction of mutated gene.

mediated genome modification are given in Table 2.

Recent studies carried out on the efficiencies of RVDs showed that DNA recognition by RVD differs in binding efficiency. Based on binding efficiency, RVDs can be classified into strong, intermediate and weak[10,12]. It is found that weak RVDs always compromise TALE activity and its function relies on the presence of strong RVDs. Also, there is difficulty in specific recognition of guanine because the common RVD NN recognizes guanine and adenine, whereas the guanine-specific RVD NK apparently functions poorly when compared to NN[8]. So it important to find new RVDs for potentially improved recognition of guanine. Among naturally occurring TALEs and their corresponding target sites there are no neighbour effects, but still positional and overall biases in nucleotide specificity and RVD composition need to be assessed in engineered TALEs. Also, there is need to understand the molecular dynamics and biophysical basis of RVD efficiency and specificity. Another issue to be resolved is the influence of mismatch tolerance and chromatin effects such as methylation on functional activity of TALEs. These are essential to define rules for TALE design and RVD choices. Further, it will be interesting to find whether TALEs may be exploited for site-directed mutagenesis with single base addition or deletion.

1. Casper-Lindley, C. *et al.*, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 8336–8341.
2. Fujikawa, T. *et al.*, *Mol. Plant–Microbe Interact.*, 2006, **19**, 342–349.
3. White, F. F. and Yang, B., *Plant Physiol.*, 2009, **150**, 1677–1686.
4. Heuer, H. *et al.*, *Appl. Environ. Microbiol.*, 2007, **73**, 4379–4384.
5. Moscou, M. J. and Bogdanove, A. J., *Science*, 2009, **326**, 1501.
6. Boch, J. *et al.*, *Science*, 2009, **326**, 1509–1512.
7. Mak, A. N.-S. *et al.*, *Science*, 2012, **335**, 716–719.
8. Yin, P. *et al.*, *Cell Rep.*, 2012, **2**, 707–713.
9. Gao, H. *et al.*, *Cell Res.*, 2012, **22**, 1716–1720.
10. Mussolino, C. and Cathomen, T., *Curr. Opin. Biotechnol.*, 2012, **23**, 644–650.
11. Miller, J. C. *et al.*, *Nature*, 2010, **29**, 135–136.
12. Moore, F. E. *et al.*, *PLoS ONE*, 2012, **7**, e37877.
13. Schornack, S. *et al.*, *Mol. Plant–Microbe Interact.*, 2005, **18**, 1215–1225.
14. Schornack, S. *et al.*, *New Phytol.*, 2008, **179**, 546–556.
15. White, F. F. *et al.*, *Mol. Plant Pathol.*, 2009, **10**, 749–766.
16. Guo, W. *et al.*, *Microbiology*, 2011, **158**, 505–518.
17. Boch, J. and Bonas, U., *Annu. Rev. Phytopathol.*, 2010, **48**, 419–436.
18. Ogunjobi, A. A., Fagade, O. E., Dixon, A. G. O. and Bandyopadhyay, R., *J. Agric. Biotechnol. Sustain. Dev.*, 2010, **2**, 87.
19. Yang, Y., De Feyter, R. and Gabriel, D. W., *MPMI J.*, 1994, **7**, 345–355.
20. Heuer, H. *et al.*, *Appl. Environ. Microbiol.*, 2007, **73**, 4379–4384.
21. Tremblay, J. P. *et al.*, *Hum. Gene Ther.*, 2012, **23**, 883–890.
22. Li, T. *et al.*, *Nature*, 2012, **30**, 390–392.
23. Tesson, L., *et al.*, *Nature*, 2011, **29**, 695–696.
24. Cermak, T. *et al.*, *Nucleic Acids Res.*, 2011, **39**, e82.
25. Sajwan, S. *et al.*, *Insect Biochem. Mol. Biol.*, 2013, **43**, 17–23.

ACKNOWLEDGEMENTS. B.K. thanks DBT, New Delhi for M Sc Biotechnology scholarship. We thank Anjan Barman for carefully reading the manuscript and for his useful comments.

Basavraj Khanppnavar, Shashi Baruah and Suvendra Kumar Ray are in the Department of Molecular Biology and Biotechnology, Tezpur University, Napaam, Tezpur 784 028, India.*
**e-mail: basuraj1991@gmail.com*