Progress in Biophysics and Molecular Biology 114 (2014) 25–32

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

Review

# TALE: A tale of genome editing

Mingjie Zhang [a,c,1], Feng Wang [b,1], Shifei Li [a], Yan Wang [a], Yun Bai [a,**], Xueqing Xu [a,*]

[a] Department of Medical Genetics, College of Basic Medical Science, Third Military Medical University, Chongqing 400038, People's Republic of China
[b] Department of Clinical Laboratory, Institute of Surgery Research, Daping Hospital, Third Military Medical University, Chongqing 400042, People's Republic of China
[c] Battalion 16 of Cadet Brigade, Third Military Medical University, Chongqing 400038, People's Republic of China

ARTICLE INFO

*Article history:*
Available online 27 November 2013

*Keywords:*
Transcription activator-like effectors
TALE transcription factors
TALE nucleases
Genome editing
Gene targeting

ABSTRACT

Transcription activator-like effectors (TALEs), first identified in *Xanthomonas* bacteria, are naturally occurring or artificially designed proteins that modulate gene transcription. These proteins recognize and bind DNA sequences based on a variable numbers of tandem repeats. Each repeat is comprised of a set of ~34 conserved amino acids; within this conserved domain, there are usually two amino acids that distinguish one TALE from another. Interestingly, TALEs have revealed a simple cipher for the one-to-one recognition of proteins for DNA bases. Synthetic TALEs have been used to successfully target genes in a variety of species, including humans. Depending on the type of functional domain that is fused to the TALE of interest, these proteins can have diverse biological effects. For example, after binding DNA, TALEs fused to transcriptional activation domains can function as robust transcription factors (TALE-TFs), while fused to restriction endonucleases (TALENs) can cut DNA. Targeted genome editing, in theory, is capable of modifying any endogenous gene sequence of interest; this can be performed in cells or organisms, and may be applied to clinical gene-based therapies in the future. With current technologies, highly accurate, specific, and reliable gene editing cannot be achieved. Thus, recognition and binding mechanisms governing TALE biology are currently hot research areas. In this review, we summarize the major advances in TALE technology over the past several years with a focus on the interaction between TALEs and DNA, TALE design and construction, potential applications for this technology, and unique characteristics that make TALEs superior to zinc finger endonucleases.



**Contents**



## 1. Introduction

Genome editing can be used to modify any gene sequence of interest in either cells or whole organisms. Currently, this technique is useful for functional genomic studies. In the future, it may

* Corresponding author. Tel.: +86 23 68753925; fax: +86 23 68752224.
** Corresponding author. Tel.: +86 23 68752258; fax: +86 23 68752224.
*E-mail addresses:* zhangmingjie1989@hotmail.com (M. Zhang), Wangfnew@gmail.com (F. Wang), baiyungene@gmail.com (Y. Bai), buyi.chu@gmail.com (X. Xu).
[1] These authors contributed equally to this work.

0079-6107/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.pbiomolbio.2013.11.006

also be useful in the clinic for the treatment of heritable diseases. Ideally, genome editing technology should be specific for the desired target, efficient, and affordable. Target specificity is typically achieved by gene targeting, in which a specific genomic sequence is recognized, bound, and modified via homologous recombination. While this methodology is popular in the laboratory, its lack of efficiency and its high cost have limited its widespread use. Zinc finger endonucleases (ZFN) are engineered proteins comprised of two distinct domains: a zinc finger domain to recognize and bind to DNA and the endonuclease FokI, which cuts DNA at a specific sequence. ZFN dramatically increase the efficiency of genome editing. Interestingly, drugs based on this methodology are currently being tested in phase 2 clinical trials for the treatment of HIV. Despite its promise, ZFN technology is limited by design complexity and cost; moreover, several companies hold patents for this type of technology.

Over the past several years, a second-generation artificial transcription activator-like effector nuclease (TALEN) has rapidly emerged as an alternative to ZFNs for genome editing. TALENs are similar to ZFNs in that they consist of a FokI nuclease domain fused to a customizable DNA-binding domain. The primary difference between the two is that the DNA-binding domain of TALEs is not a zinc finger domain. Instead, it is a multiple 33–35-amino-acid repeat domain that recognizes a single base pair. This technology implements a simple 'protein–DNA code' in which the modular DNA-binding TALE repeat domains are complementary to individual bases in a unique target-binding site. This feature makes the design and production of TALEs much easier and more efficient than the production of ZFNs. Over the course of a relatively short time period, TALES have been successfully used to perform genome editing in plants, zebrafish, frogs, rats and pigs. The technology has also been used in human somatic and pluripotent stem cells. Importantly, this technology has proven successful in several cell lines and model organisms that have been described as extremely difficult or even impossible to genetically manipulate. In this review, we summarize the major advances in TALE technology over the years. We focus on the interaction between TALEs and DNA, the design and construction of these proteins, how they are superior to ZFNs, and their potential applications.

## 2. The discovery of TALEs

TALEs were first discovered in the bacteria *Xanthomonas oryzae pv. oryzae* (*Xoo*) and *Xanthomonas oryzae pv. oryzicola* (*Xooc*). These two pathovars are responsible for decreasing the world's rice production by causing bacterial blight (BB) and bacterial leaf streak (BLS) in the rice species *Oryza sativa* (Bennetzen and Ma, 2003; Ronald and Leung, 2002). The effector proteins in these bacteria are encoded by a gene family called *avrBs3/PthA* (short for avirulence/pathogenicity A). AvrBs3 is the first member of the avrBs3/PthA family and encodes a 122kD protein with 17.5 repeats. Each repeat, except for the non-full-length repeat at the end (i.e. the 0.5 repeat), consists of 34 tandem amino acids (aa) (Bonas et al., 1989). AvrBs3-like genes with repeat domain diversity are found in 309 strains of *Ralstonia solanacearum biovars* (Heuer et al., 2007).

These natural effector proteins, termed Type III effectors, are secreted by *Xanthomonas* bacteria via their type III secretion system (T3SS) and delivered into host cells where they regulate the transcription of genes involved in disease pathogenesis (Szurek et al., 2002). The T3SS is encoded by the *hrp* (hypersensitive response and pathogenicity) gene cluster. Expression of genes from this cluster facilitates transfer of proteins from bacteria to eukaryotic cells (Cornelis and Van Gijsegem, 2000). Once these bacterial proteins have entered the nucleus of a host cell, they can bind to target DNA sequences and regulate gene expression.

Owing to their similarity to eukaryotic transcription factors, these bacterial effector proteins were termed transcription activator-like (TAL) effectors (also referred to as TALEs). They generally contain carboxyl (C)-terminal nuclear localization signals (NLSs), an acidic transcriptional activation domain (AD) (Schornack et al., 2006; Zhu et al., 1998), and an N-terminal type III secretion and translocation signal (T3S) (Bogdanove et al., 2010) (Fig. 1A). The presence of NLSs (2 and 3) of the AvrBs3 protein contributes to the occurrence of hypersensitive response (HR) (Van den Ackerveken et al., 1996). In contrast, NLS1 is insufficient for full activity of the AvrBs3 protein (Szurek et al., 2001). The original AD region is required for the HR-inducing activity of AvrBs3; the foreign AD can compensate for the AD-deficient AvrBs3 protein to some extent (Szurek et al., 2001).

Thorough analyses of the 113 known TALEs from *Xanthomonas*, Boch and Bonas (Boch and Bonas, 2010) revealed that these molecules have nearly identical repeats of 30–42 amino acids, with 34 amino acids being most typical. The number of repeats for each TALE ranges from 1.5 to 33.5 (most commonly 17.5). The structure of these repeats is the basis of the simple 'protein–DNA code' discussed in detail below.

## 3. Features of TALEs: recognition and specificity

TALEs are proteins with customizable DNA binding domains (Römer et al., 2007; Voytas and Joung, 2009) and are distinct from other DNA binding factors, including zinc finger proteins (ZFP) (Gommans et al., 2005), helix-turn-helix (HTH) proteins (Aravind et al., 2005), homeodomain (HD) factors (Mannervik, 1999), leucine zippers (Elhiti and Stasolla, 2009), and basic helix-loop-helix (bHLH) proteins (Jones, 2004). The extremely difference of TALEs is the 'protein-DNA code' they follow, which allows specific recognition between one TALE and one nucleotide. TALEs differ from one another in the number and order of their repetitive sequences. Each repeat consists of approximately 34 amino acids (termed a monomer), which are highly conserved except for two hypervariable residues at positions 12 and 13 (Boch et al., 2009) (Fig. 1A). These two amino acids (termed repeat variable diresidues, RVDs) determine the nucleotide-binding specificity of each repeat (Boch et al., 2009). Some amino acid repeats do not recognize a specific nucleotide sequence; instead, some are more promiscuous. For example, NN and N* can recognize either G/A or C/T, respectively (Moscou and Bogdanove, 2009). Furthermore, a single nucleotide may be recognized by multiple repeats, albeit with different efficiencies. For example, NN is more efficient when binding with G instead of NK (Christian et al., 2012). Amino acid repeats differing in the non-RVD regions have only minimal effect on the defined RVD specificity (Morbitzer et al., 2010) (summarized in Fig. 1B). Streubel et al. (2012) recently described new parameters to better characterize the specificity and efficiency of RVDs. Unlike earlier reports, the authors of this paper classified RVDs into weak, intermediate, and strong to describe their binding efficiencies; they suggested that these classifications should help researchers design more efficient TALEs (Streubel et al., 2012).

Since some repeats are nonspecific, how can one be sure that the TALE in question will actually target the sequence of interest in the genome? Previous work has shown that a minimum of 16.5 bases, with at least 6.5 continuous repeats from the N-terminus, is needed to specifically activate expression (Boch et al., 2009). While TALEs with few repeats may also activate gene expression, those with repeats less than the threshold number display severely attenuated activity. Morbitzer et al. (2010) described that many mismatching repeats at the C-terminus limit activity and may also impair TALE integrity and stability. Such problems can generally be fine-tuned by increasing the number of repeat units.