DOI 10.1515/hsz-2012-0164 — Biol. Chem. 2012; 393(10): 1055–1066

Review

Matthias Bochtler*

Structural basis of the TAL effector–DNA interaction

Abstract: Phytopathogen transcription activator-like effectors (TALEs) bind DNA in a sequence specific manner in order to manipulate host transcription. TALE specificity correlates with repeat variable diresidues in otherwise highly stereotypical 34–35mer repeats. Recently, the crystal structures of two TALE DNA-binding domains have illustrated the molecular basis of the TALE cipher. The structures show that the TALE repeats form a right-handed superhelix that is wound around largely undistorted B-DNA to match its helical parameters. Surprisingly, repeat variable residue 1 is not in contact with the bases. Instead, it is involved in hydrogen bonding interactions that stabilize the overall structure of the protein. Repeat variable residue 2 contacts the top strand base and forms sequence-specific hydrogen bonds and/or van der Waals contacts. Very unexpectedly, bottom strand bases are exposed to solvent and do not make any direct contacts with the protein. This review contains a summary of TALE biology and applications and a detailed description of the recent breakthroughs that have provided insights into the molecular basis of the TALE code.

Keywords: cipher; structure; transcription activator-like effector (TALE).

***Corresponding author: Matthias Bochtler,** International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland; and Institute of Biochemistry and Biophysics of the Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warsaw, Poland, e-mail: mbochtler@iimcb.gov.pl

Introduction: virulence and avirulence in phytopathogen–host interactions

Xanthomonas, *Pseudomonas* and *Ralstonia* pathogens are locked in an evolutionary arms race with their hosts (Jones and Dangl, 2006). The bacteria produce virulence factors, termed effectors, to manipulate the host to their advantage, for example in order to facilitate colonization and spreading or to combat resistance (Gohre and Robatzek, 2008). From the perspective of the host, bacterial virulence factors signal infection. Hosts have therefore evolved mechanisms to recognize such factors (Van den Ackerveken and Bonas, 1997; Dangl and Jones, 2001; Jones and Dangl, 2006), often with the help of a single resistance (*R*) gene (Keen, 1990). A gene that normally promotes infection in susceptible plants becomes an avirulence (*avr*) gene for resistant plants (Van den Ackerveken and Bonas, 1997) and is named accordingly (Vivian and Mansfield, 1993). For example, *avrBs3* designates the bacterial gene that renders *Xanthomonas* avirulent in the presence of the plant *Bs3* gene (Bonas et al., 1989).

Transcription activator-like effector virulence/avirulence factors

Transcription activator-like effectors (TALEs) belong to the virulence/avirulence proteins and have been reviewed in detail (Boch and Bonas, 2010; Bogdanove et al., 2010; Scholze and Boch, 2011). They are made by bacteria, but injected into the plant cytoplasm and then translocated to the nucleus (Szurek et al., 2002). With hindsight, it is now clear that TALE proteins are not only 'transcription activator-like', but actually control transcription in both susceptible (Kay et al., 2007) and resistant plants (Romer et al., 2007; Romer et al., 2009b). The architecture of TALE proteins reflects their function (Bogdanove et al., 2010). Near the amino terminus, there is a type III secretion signal that directs the transfer of the protein from the phytopathogen producer into the plant cytoplasm. Next, there is a set of 34–35mer repeats that differ almost exclusively in two adjacent positions (residues 12 and 13 in conventional numbering, known as the repeat variable diresidue or RVD). Despite their stereotypical nature, the repeats mediate tight and sequence-specific DNA binding (Kay et al., 2007). In the C-terminal region, TALE proteins have

one or several nuclear localization signals for transfer to the nucleus and an activation domain for interaction with the host transcription machinery (Figure 1).

The TALE cipher correlates repeat variable diresidue amino acid and target DNA sequence

The correlation between the TALE repeat amino acid sequence and specifically recognized DNA has been investigated using at least three different assays. Early experiments were done using a system closely mimicking physiological conditions. *Xanthomonas* was used to express TALEs and to secrete them into plant cells. In these assays, the TALE concentration was relatively low. Higher concentrations of TALEs were achieved in so called 'agroassays'. In these experiments, *Agrobacterium* was used to deliver 35S promoter driven T-DNA coding for a TALE together with the promoter reporter construct. Finally, some TALE specificity data have been obtained *in vitro* using the systematic evolution of ligands by exponential enrichment (Miller et al., 2011). The data that were obtained by the three different approaches are mostly, but not completely, consistent. In essence, they show that there is a strikingly simple correlation between TALE protein and target DNA sequences (Figure 2; Boch et al., 2009; Moscou and Bogdanove, 2009). This clear-cut 'cipher' correlates RVDs with single DNA bases (in the 'top' DNA strand): HD pairs with Cyt, NG with Thy, NI with Ade and NN with Gua or Ade (single letter code for amino acids and three-letter code for bases; Boch et al., 2009; Moscou and Bogdanove, 2009). Recent studies have extended the code and shown that the NK diresidue binds Gua (Morbitzer et al., 2010; Miller et al., 2011). HG is ambiguous and can select either Cyt or Thy. The naturally frequent NS

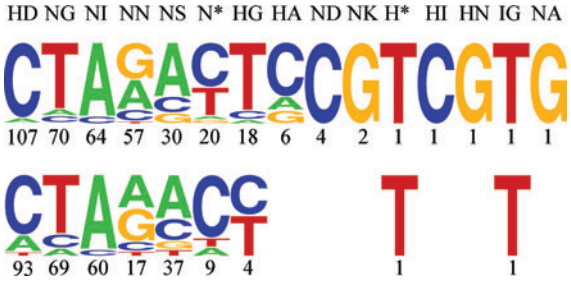

Figure 2 TALE cipher in frequency representation.
The repeat variable diresidues (RVDs) and DNA bases are given in single letter code. Amino acids: A-alanine, D-aspartic acid/aspartate, G-glycin, H-histidine, I-isoleucine, K-lysine, N-asparagine, S-serine. Bases: A-adenine, C-cytosine, G-guanine, T-thymine. Top: The data from Moscou and Bogdanove (2009) are shown in the original representation. Bottom: The data from Boch and coworkers (2009) have been reordered and are redisplayed to reflect frequencies. The number of cases for each RVD is indicated below the logo. The star corresponds to the missing RVR residue (by convention taken to be RVR2). Frequencies on the right, which are based on very few instances, are likely to change as more data become available. Some RVDs were not analyzed by Boch and coworkers.

favors Ade or Cyt, but can also pair with the other bases (Boch et al., 2009; Moscou and Bogdanove, 2009). Ten or more repeats (plus an obligatory flanking half-repeat) are required for efficient transcription activation (Boch et al., 2009). Specifically recognized DNA sequences start with Thy (Boch et al., 2009). The subsequent order of DNA bases in the top DNA strand (in 5′–3′ direction) is dictated by the order of RVDs (from N to C-terminus, Figure 2; Boch et al., 2009; Moscou and Bogdanove, 2009).

Applications of TALE proteins

The simplicity of the TALE cipher makes it possible to design artificial proteins that specifically bind a predetermined DNA sequence (Boch et al., 2009; Morbitzer et al., 2010; Miller et al., 2011; Zhang et al., 2011). Together with

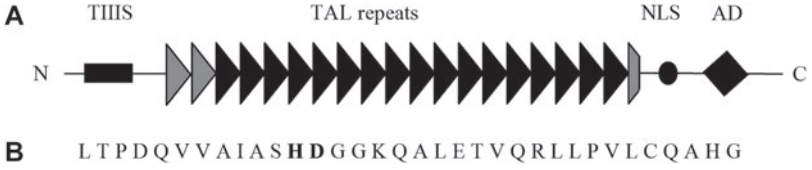

Figure 1 TALE architecture.
(A) Domain organization of a typical TAL effector protein. TIIIS stands for type III secretion signal, NLS for nuclear localization signal, and AD for activation domain. The black triangles represent canonical TALE repeats. The complete grey triangles represent two degenerate repeats that cooperate to pair with a 5′-Thy. The truncated grey triangle stands for a half-repeat at the end of the repeat domain. (B) Amino acid sequence of a (typical) single repeat. The amino acids in positions 12 and 13, which form the repeat variable diresidue (RVD), are highlighted by bold print.

methods for efficient TALE assembly (Cermak et al., 2011; Geissler et al., 2011; Li et al., 2011, 2012; Morbitzer et al., 2011; Weber et al., 2011; Zhang et al., 2011; Reyon et al., 2012; Sanjana et al., 2012), this has prompted the use of TALE proteins for various applications (Bogdanove and Voytas, 2011).

TALEs can be applied to transcription control as demonstrated in plants (Boch et al., 2009; Romer et al., 2009a,b, 2010; Morbitzer et al., 2010; Mahfouz et al., 2011; Weber et al., 2011; Li et al., 2012) and in mammalian cell cultures (Geissler et al., 2011; Miller et al., 2011; Zhang et al., 2011; Bultmann et al., 2012). For engineered transcriptional control, TALE binding sites should be chosen in a way that will minimize off-target effects. A tool for the selection of such 'orthogonal' target sequences has just been published (Garg et al., 2012).

TALEs can also be adapted to genomic editing by fusing their sequence-selective regions to unspecific FokI nuclease domains to create TALE nucleases (TALENs; Christian et al., 2010; Li et al., 2011; Miller et al., 2011). As FokI is only active as a dimer (Bitinaite et al., 1998), two fusion proteins are required that bind in a head-to-head manner to adjacent DNA regions. TALENs have been shown to initiate homologous recombination in yeast (Li et al., 2011) and to work in the genome editing of cultured cells from different vertebrates, including humans (Hockemeyer et al., 2011; Miller et al., 2011; Sander et al., 2011; Tong et al., 2012). Whole animal transgenesis with TALENs is attractive because it is applicable to species without established embryonic stem cells. The technique has already been demonstrated for both invertebrates (Wood et al., 2011; Liu et al., 2012) and vertebrates (Huang et al., 2011; Tesson et al., 2011). The list of tractable species is likely to grow quickly because it is now much easier to obtain the necessary TALENs. Computational tools are available to aid the selection of target sites and the construction of TALENs using different assembly protocols (Doyle et al., 2012). Moreover, custom-made TALENs can now be obtained from commercial sources (de Francesco, 2011). It is still too early to tell how TALENs will compare with zinc finger nucleases (ZFNs) in terms of efficiency and toxicity, but they certainly look promising (de Francesco, 2011; Mussolino et al., 2011).

Towards a structural interpretation of the TALE cipher

The first hints about the structure of TALE repeats were based on the observation that TALE and tetratricopeptide repeat (TPR) sequences were distantly similar. Strictly speaking, the term 'TPR' describes only the repeat length (34 amino acids). However, from the beginning it was coined with reference to a particular set of homologous proteins (Hirano et al., 1990; Sikorski et al., 1990). The family of TPR proteins has since grown dramatically. They all share the same key structural features. A TPR repeat consists of two α-helices with a short linker in between (Blatch and Lassle, 1999; D'Andrea and Regan, 2003). Repeats usually assemble into right-handed superhelical structures (Main et al., 2003). Most TPR proteins are thought to be involved in protein–protein interactions or in the assembly of multi-subunit complexes (Blatch and Lassle, 1999). Based on the (distant) similarities between the TALE and TPR amino acid sequences, it was suggested that TALE repeats adopt a TPR-like fold and assemble into a right-handed superhelix (Schornack et al., 2006).

The first experimental analysis of TALE structure was the nuclear magnetic resonance (NMR) structure of one and a half repeats of the TALE protein PthA2 in the absence of DNA (Murakami et al., 2010). The data showed that the analyzed TALE repeat adopted a TPR-like structure. The additional half-repeat was used to derive the relative orientation of repeats to each other. Interestingly, extension of the structure '*in silico*' led to a TALE superhelix with a pore wide enough to accommodate DNA on the inside. The dimensions of this TALE superhelix were consistent with small-angle X-ray scattering data for a larger PthA2 fragment. However, the axial spacing of the modeled TALE superhelix (≈6.8 Å) was twice the size of the axial spacing between bases in B-DNA (≈3.4 Å). This finding implied that the TALE repeats had to adopt a more compact structure in order to bind to DNA. Experimentally, dynamic light scattering experiments showed that the hydrodynamic radius of the PthA2 repeat domain decreased significantly upon the addition of a DNA duplex representing part of the PthA2-induced promoter (Murakami et al., 2010). Based on these data, a detailed model with DNA on the inside and TALE repeats in a superhelical arrangement adjusted to match the axial and angular spacings between base pairs was proposed (Scholze and Boch, 2011). Even this model, however, was not sufficiently detailed to explain the TALE cipher in terms of molecular interactions.

Crystal structures of TALE repeat domains

The detailed structure of TALE repeats and their interactions with DNA was clarified in macromolecular

crystallographic studies. A series of independent structures, obtained by two different groups (Deng et al., 2012; Mak et al., 2012), has recently provided key insights. The results of the two studies are consistent with each other, but also very complementary. Deng et al. have analyzed dHax3 (Deng et al., 2012), a variant designed based on Hax3 from *Xanthomonas campestris pv. armoraciae* (Mahfouz et al., 2011). Mak et al. have focused on the naturally occurring TALE protein, PthXo1, from the rice pathogen *Xanthomonas oryzae* (Mak et al., 2012). Deng and colleagues managed to crystallize the DNA-binding domain of dHax3 in the presence and absence of its target, so they can provide insights into the rearrangements necessary for complex formation. The main limitation of their study is that dHax3 is relatively short (11.5 repeats) and contains only a limited diversity of RVDs (HD, NG and NS). Mak and colleagues did not determine the apo-structure, but their DNA-binding domain is much longer (23.5 repeats) and contains a more diverse set of RVDs (HD, NG, HG, NI, NN and N*; unfortunately the single NS repeat is disordered in the crystal). Together, the new structures almost completely explain the TALE cipher (with the exception of the NK–Gua interaction). At the same time as the experimental papers, a prediction of the TALE–DNA interactions was published (Bradley, 2012). The model is strikingly accurate, but nevertheless superseded by the crystal structures. Hence, the following sections will be focused exclusively on the experimental data.

Apo-structure of the complete TALE DNA-binding domain

The dHax3 structure in the absence of DNA confirms key features of the PthA2 NMR model, but also differs in important respects. First, Deng et al. report that the angle between the first and second helices of the repeat is different from the angle determined by NMR. This discrepancy might result from the amino acid differences between PthA2 and dHax3. However, it could be also due to the drastic PthA2 truncation that was necessary to obtain the structure, or reflect the difficulty in accurately determining the relative position of secondary structure elements by NMR. Independent of the detailed explanation, Deng et al. suggest that TALE repeats should not be classified as TPR-like, due to the different angles between the helices within one repeat. Second, the NMR and crystallographic data assign different conformations to the RVD. According to the former, the RVD main chain adopts a helical conformation. As a result, the RVR1 and RVR2 side chains are oriented roughly parallel to each other, so that both could interact with the DNA bases of a Watson-Crick pair. In contrast to the NMR results, the crystallographic data place the RVD in the loop between repeat helices. With this arrangement, the RVR1 and RVR2 side chains point in opposite directions, so that at most one of them can interact with a specifically recognized Watson-Crick pair. Third, the crystal structure shows a clear kink in the second helix of the repeat, which was not reported for the NMR model. Fourth, the consecutive repeats of the dHax3 superhelix are separated by an axial distance of 5.4 Å, which is smaller than for PthA2, but still much larger than the 3.4 Å van der Waals distance between base pairs in B-DNA.

Gross structure of the TALE–DNA complexes

The crystal structures of the two TALE DNA-binding domains with their targets show the flexibility of the repeat arrangement. As predicted, the TALE superhelix adapts to match the helical parameters of the DNA (Figure 3; Scholze and Boch, 2011). In the case of dHax3, a direct comparison between the TALE conformations in the presence and absence of DNA is possible. The superposition shows that the repeat structure is almost identical for

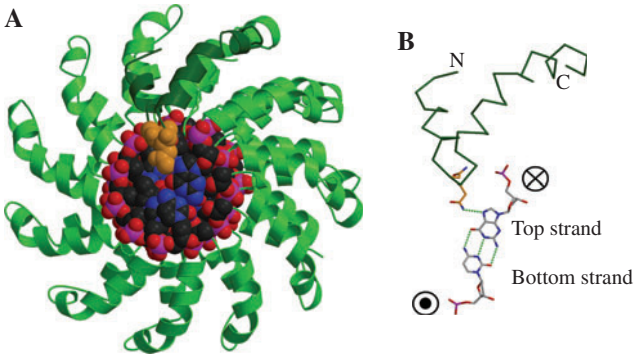

Figure 3 PthXo1–DNA interaction (top view).
(A) Fifteen repeats and DNA are shown. The repeats are presented as green ribbons and the DNA is in space filling representation. The first repeat is in dark green with repeat variable diresidue amino acids in orange in space filling representation. (B) Detail from (A), showing only the Cα trace of the first repeat, the repeat variable diresidue and its interacting base pair. Hydrogen bonds are indicated by dotted green lines. Circles with a dot/cross inside indicate the polarity of the DNA backbone (in 5′–3′ direction) towards the viewer and into the page, respectively.

residues 1 to approximately 22, and that subtle changes of the further 12 amino acids accumulate and lead to the overall more compact arrangement in complex with DNA. In contrast to the protein, the DNA does not change conformation in a major way upon complex formation and is close to perfect B-DNA in both crystal structures (except for very slight overwinding with 11 base pairs per turn). As expected, RVDs come close to the specifically recognized bases. To a first approximation, the first and second helices of each repeat point towards and away from the DNA in a radial direction, respectively. A more accurate description takes into account the axial components of helix orientation antiparallel and parallel to the top DNA strand for the first and second helices, respectively (if 'direction' is defined as 5′–3′ for DNA and N to C for protein; Figure 4). In the view along the DNA, the packing arrangements are clearly visible. Looking down the top DNA strand, the second helix of a repeat is placed approximately 'on top' of the first helix of the next one (Figure 3).

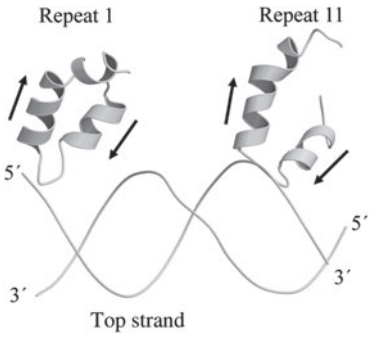

Figure 4 PthXo1–DNA interaction (side view).
Schematic representation of TALE repeats 1 and 11 bound to a 15-mer DNA. The arrows indicate polypeptide chain direction.

DNA readout from the major groove side

The crystal structures show that TALE repeats exclusively make contacts with the DNA on the major groove side (Figure 3B). This does not come as a surprise. Compared to the minor groove, it is much wider and more accessible. Moreover, hydrogen bonding interactions and shape selection on the major groove side are sufficient to distinguish all four bases. In the minor groove Ade:Thy and Thy:Ade pairs 'look alike', and the same applies for Cyt:Gua and Gua:Cyt pairs. Minor groove binding proteins therefore only tend to distinguish W (Ade or Thy) from S (Gua or Cyt), but not individual bases. There is no evidence in the TALE cipher for this type of degenerate sequence recognition (Figure 2), in agreement with the structural data.

No contacts between RVR1 and the DNA bases

Repeat variability is essentially limited to the RVDs, suggesting that both residues are involved in direct hydrogen bonding and van der Waals interactions with the DNA bases. On the other hand, RVR1 makes little contribution to specificity. Asparagine in the first position is compatible with all four bases, and histidine at least with the pyrimidines. A natural explanation of this promiscuity of RVR1 could be that N and H can act both as hydrogen bond donors and acceptors and can therefore interact with all four bases. Surprisingly, this explanation is wrong. Both PthXo1 and dHax3 crystal structures show that RVR1 does not make any direct contacts with the bases. Instead, it points away from the DNA and is involved in interprotein interactions that stabilize the repeat conformation. According to the dHax3 structure, the RVR1 donates a hydrogen bond to the carbonyl oxygen atom of Ala8 of the same repeat. A more complex picture emerges from the structure of PthXo1, which contains several instances of the IG RVD. As the I side chain does not contain polar groups, the intra-repeat hydrogen bond with Ala8 cannot be established. Moreover, in the PthXo1 structure the interactions of RVR1 are less stereotypical, even if this residue is suitable as a hydrogen bond donor. In some cases it even appears to interact with a downstream repeat rather than with an adjacent residue.

Contacts of RVR2 with the top strand DNA base

According to the PthXo1 and dHax3 crystal structures, TALE repeat sequence selectivity is mediated by the contact of RVR2 with the top strand base of target DNA. Together, the PthXo1 and dHax3 crystal structures provide illustrations for most of the RVR2 interactions with cognate bases in the TALE cipher. Inevitably, crystal structures can only show favorable contacts that promote binding. Unfavorable interactions that discourage or preclude the binding of other DNA bases are difficult to observe. Nevertheless, they contribute to selectivity at least as much as the favorable ones (due to the high penalty for van der Waals clashes, anti-selection can be more rigorous than selection). Therefore, predicted unfavorable interactions are also described here (Figure 5). In the following, RVR2 residues are ordered according to their mode of base selectivity. I and G are described first,

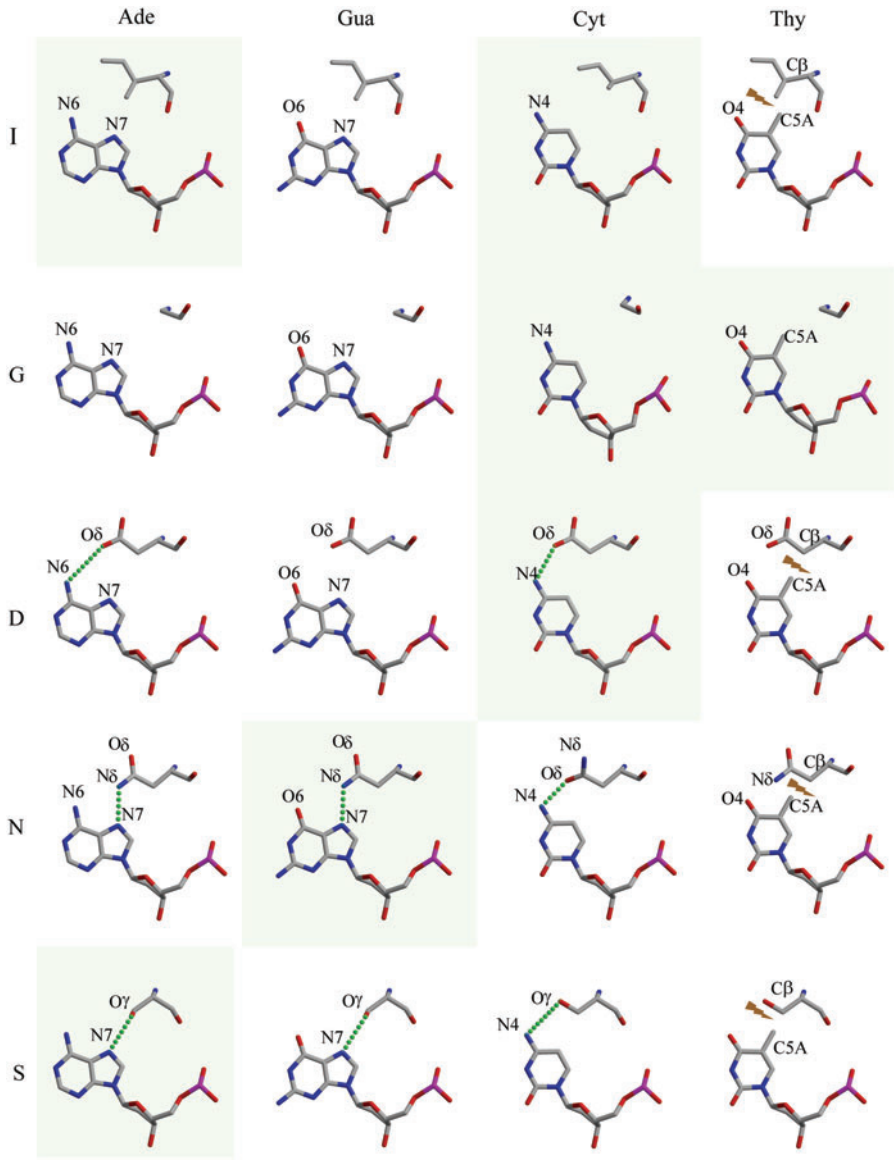

Figure 5 Structural basis of the TALE cipher.
All atom representation of RVR2 residues and interacting bases, colored according to atom type. Dotted green lines indicate hydrogen bonds. Interactions that have already been documented crystallographically are presented on a light green background. All others (which have only been modeled) are presented on a white background. Clashes are indicated by brown lightning symbols.

because their specificity is largely determined by shape complementarity (van der Waals interactions). Next, the polar residues D, N and S are discussed. For their interactions with DNA bases, shapes and hydrogen bonds are important. K is discussed last, because its interactions with DNA bases have not yet been documented by a crystal structure.

Isoleucine has a fairly bulky side chain that reaches into the outermost major groove region (in the documented interactions with Ade and Cyt). In this region, purine bases occupy the least space, Cyt somewhat more space and Thy is the most bulky. On this basis, we might expect that I would select for Ade and Gua, and to a lesser extent Cyt. However, there is a complication. In order to fit, the purine bases must lose a water molecule bound to the N7 atom. This loss has an entropic benefit, but an enthalpic cost. Taking the desolvation cost into account, binding of Cyt could be more favorable than binding of Ade and Gua. Experimentally, I is found to preferentially interact with Ade and to a lesser extent with Cyt, but barely with Gua. The reasons for this order of preferences are not yet fully understood, but at least the discrimination against

Thy is perfectly consistent with the structural data (Figure 5, row 1).

Glycine has no side chain. Interactions between G (in the HG and NG RVDs) and Thy and Cyt are documented in the crystal structures. In most but not all cases, the 5-methyl group of Thy (the carbon atom is labeled C5A in Figure 5) is in van der Waals contact with the glycine main chain. Cyt, which lacks the 5-methyl group, cannot make such a contact, which is confirmed by the crystal structure. Similarly, the glycine is too far away from the Hoogsteen edge of the purine bases to make direct van der Waals contacts (in the absence of structural changes to accommodate these bases). Thus, the strong preference of NG and HG for Thy that has been observed experimentally is clearly explained, but a mild bias towards some (Cyt, Ade) and against other (Gua) bases is not directly obvious from the structural data (Figure 5, row 2).

Aspartate acts as a hydrogen bond acceptor (at neutral pH) and therefore selects for the bases that have hydrogen bond donor atoms in the major groove. Its interaction with Cyt is documented in the PthXo1 and dHax3 crystal structures and shows the expected hydrogen bond from the N4 atom of Cyt to the Oδ atom of D. The much less frequent interaction with Ade is not yet structurally documented, but it can be predicted that D accepts a hydrogen bond from the Ade N6 atom. The Ade N6 is located more centrally in the major groove than the Cyt N4. Therefore the hydrogen bond with D is expected to be longer and weaker. D excludes binding of Thy, because its Cβ atom (and at least for some conformations also its Cγ and Oδ atoms) would clash with the 5-methyl group of Thy. Thus, the structural data are perfectly consistent with the experimental observation that D selects for Cyt and (rarely) for Ade (Figure 5, row 3).

Asparagine is isosteric to D (except for hydrogen atoms), but it can act as a hydrogen bond donor or acceptor. In the PthXo1 structure, its interaction with Gua has been observed twice and shows a hydrogen bond from the N carboxamide to the N7 atom of Gua. In both instances, the complementary base (Cyt) is found to be significantly out of the plane of the recognized Gua. Whether this is a general feature of this pairing remains to be clarified when more instances are crystallographically characterized. A hydrogen bond from the N side chain carboxamide to the N7 atom is also expected when N pairs with Ade. In contrast, a flip of the N side chain (which interchanges the Oδ and Nδ atoms) is required for pairing with Cyt. As for aspartate, steric clashes prevent pairing of N with Thy. The very rarely observed pairing of N with Thy can be attributed to a rearrangement of the main chain conformation (Figure 5, row 4).

Serine has a shorter side chain than the other RVR2 amino acids (except for glycine) and can act as a hydrogen bond donor or acceptor. So far, only the pairing of S with Ade has been crystallographically observed. S donates a hydrogen bond to the N7 atom of the purine base. Although not yet observed, the same pattern is likely to occur for the interaction of S with Gua. As for other amino acids with side chains, S should exclude Thy by steric clashes. Rare pairings of S with Thy probably involve avoidance of these clashes, presumably via rearrangements in the main chain of the protein to reposition the S side chain (Figure 5, row 5).

Lys is not found very frequently in natural TALE proteins, but the NK RVD has been shown to preferentially pair with Gua. In structural terms, K has a very long side chain and a terminal amino group that should be protonated at neutral pH and therefore is only suitable as a hydrogen bond donor. The exocyclic O6 atom of Gua is a hydrogen bond acceptor that is further away from the RVR2 main chain than the N7 atom. Although the side chain of K can adopt very twisted conformations that bring the amino group closer to the main chain, the most common rotamers are extended. It is therefore more likely that the K amino group donates a hydrogen bond to the Gua O6 atom. This interaction does not work for adenine and can therefore explain why NK, in contrast to NN, prefers Gua over Ade (interaction not shown).

In addition to the canonical RVDs that consist of two amino acids, some TALE repeats have a single RVR. The N* (the * denotes the 'missing' amino acid) prefers the pyrimidines, but is not very selective and can also accept purine bases. There are two examples of this repeat in the PthXo1 structure. Like NG, N* does not have an amino acid side chain and therefore has space to accommodate Thy. The preference of pyrimidines over purines is not obvious from the crystal structure, since the truncated loop is found at considerable distance from the top strand base (interaction not shown).

Recognition of the 5′-Thy

The base at the 5′-end of a sequence that is selectively bound by a series of TALE repeats is a Thy. This was recognized at the time of the discovery of the TALE cipher and speculatively attributed to interactions with the 0th repeat. The PthXo1 structure qualitatively supports this concept. A 0th repeat is present, but upstream there is yet another cryptic repeat (termed the -1st repeat), which is only apparent in the structure, but was not recognized at the level of the amino acid sequence. Surprisingly, the

selection of Thy is mediated by the -1st (rather than the 0th) repeat, via a favorable interaction with the indole ring of a tryptophan residue (Figure 6). In the dHax3 structure, there is a tryptophan residue at (approximately) the same place in the amino acid sequence, but this residue does not come close to the 5′-Thy. The lack of an interaction could be due to the choice of the N-terminal boundary of the dHax construct (there is only one residue upstream of the tryptophan).

Influence of methylation on the TALE code

In higher plants, cytosine is frequently modified to 5-methylcytosine, both in the CpG and CpNpG context as well as in a sequence-independent manner (Jeltsch, 2010). It is therefore worth considering the effects of 5-methylation of cytosines on the TALE code. According to the TALE cipher, Cyt pairs preferentially with HD, HG and N*. In the case of HD, the structural data predict a clash between the 5-methyl group and the side chain of D. Hence, 5-methylation should work against this pairing. Despite this, the TALE cipher also shows that the steric clash does not absolutely exclude the 5-methyl group of Thy. The discrimination against 5-methylcytosine is therefore not expected to be perfect. The TALE cipher shows that HG and N* can accept Thy, and should thus have space for a pyrimidine 5-methyl group, which is consistent with the results of molecular modeling. HG and N* should therefore bind Cyt regardless of methylation. The absence of drastic effects of 5-methylation of cytosines might also explain why this aspect of the TALE code has received very little attention so far. For applications of TALE in biotechnology, the effects of other types of DNA methylation (cytosine N4 and adenine N6, found in bacteria and some lower eukaryotes; Pace, 1997) are also of interest. Molecular modeling suggests that these modifications should be much more disruptive for the TALE code. Nevertheless, there are no experimental data to confirm the prediction, perhaps because TALEs have mostly been used in cells of plants and higher animals. Clearly, more experimental work is necessary on this neglected aspect of TALE biology.

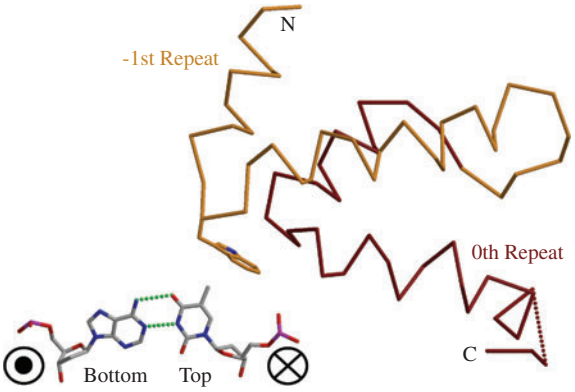

Figure 6 Structural basis of the 5′-Thy recognition.
The Cα-traces of the -1st and 0th repeats are shown in orange and brown color, respectively. The 5′-Thy and its paired Ade are shown in grey, with heteroatoms colored by atom type and hydrogen bonds indicated by green dotted lines. The tryptophan side chain, which interacts via its indole ring with the 5′-methyl group of Thy, belongs to the -1st repeat and is shown in all atom representation.

Correlations between RVR1 and RVR2

If RVR1 and RVR2 fulfill their roles completely independently of each other, it should be possible to arbitrarily combine them. In other words, the frequent HD and NI should have ND and HI counterparts. This appears to be the case, but both ND and HI are rare RVDs. ND appears to be at least roughly synonymous to HD. In the case of HI, the statistics are too poor to draw any conclusions (Figure 2). Correlations between RVR1 and RVR2 make sense if RVR1 identity affects the position and conformation of RVR2. For example, the RVR2 residues N and S must reach deeper into the major groove to reach the N7 atom of purines, than the RVR2 residue D to probe for hydrogen bonding interactions with the N4 atom of Cyt. N and S follow RVR1 residue N, whereas D follows RVR1 residue H according to the TALE cipher. Thus, we could speculate that RVR1 residues N and H promote ‘in and ‘out’ conformations of the RVR2, respectively. Unfortunately, the structural data suggest a more complex picture. Superpositions show that there is no clear-cut clustering of RVR2 Cα-positions depending on the identity of the RVR1 residue (data not shown). Whether such correlations will emerge when more crystal structures are published and the statistics of repeat conformations improve remains to be seen. For now, RVR1–RVR2 correlations cannot be fully interpreted by the structural data available.

Repertoire of RVR1 amino acids

The crystal structures suggest that a RVR1 amino acid with a hydrogen bond donor in the δ or ε position of the side chain helps to stabilize the loop conformation and therefore TALE repeat structure (although the rare IG RVD shows that this is not an absolute requirement).

This observation is consistent with the absence of D, E, S, T, C and K in the RVR1 position, because the side chains of these amino acids are too long, too short, or cannot donate a hydrogen bond at neutral pH. Amino acids R, W and Q have hydrogen bond donor atoms in the ε position, but are probably not used because of their steric bulk. The absence of Q from the RVR1 position is not easy to explain on the basis of the structural data.

Repertoire of RVR2 amino acids

RVR2 amino acid composition also poses some puzzles. S interacts with purines as a hydrogen bond donor but T or C do not play this role, at least in natural TALEs, despite potential hydrogen bond donor atoms in the γ position. The C might not be used to avoid disulfide bond formation, but the explanation is not convincing because TALE repeats contain another C elsewhere in the sequence and operate in a reducing environment. The absence of T is altogether unclear. In contrast, it is not surprising that the longer or bulkier amino acids (E, Q, W and R) are not represented in the RVR2 position. Without repositioning or structural change, these amino acids would clash with the target bases. However, the argument does not seem to apply for the flexible K. How K bypasses the apparent length restriction remains to be elucidated.

Dependence of the TALE efficacy on the repeat number

Experiments have shown that the efficiency of TALE-controlled transcription increases sharply at a threshold of around 10 repeats (Boch et al., 2009). Ten (or eleven, according to the crystal structure) base pair steps are required for a full DNA turn. It was therefore tempting to speculate that the threshold behavior was the result of direct interactions between repeats that are distant from each other in amino acid sequence, but could be close in space after a complete superhelix turn. However, such contacts are not found in either the dHax3 or the PthXo1 crystal structures. Instead, there is a gap of approximately 10 Å between the translationally-related repeats (Figure 4). This gap could only be closed by major DNA bending in solution. Such distortion seems unlikely in light of the two independent crystal structures. Thus, the dependence of TALE efficacy on the number of repeats remains unexplained.

Enhancing TALEs to recognize additional DNA features?

In the co-crystal structures of TALEs with DNA, the major groove side of the bottom DNA strand and the entire minor groove are completely solvent exposed, providing many additional interaction surfaces for sequence-specific contacts. Could two TALE DNA-binding domains bind to the same duplex DNA, with one recognizing the top and the other the bottom strand? Although this would be attractive in principle, it is unfortunately not viable, because the inner repeat helices would clash. Whether other proteins can be found to exploit the solvent-exposed hydrogen-bond donors and acceptors in the major groove of TALE-bound DNA remains to be seen. In contrast, the exposed minor groove provides attractive possibilities for enhancing the specificity and affinity of TALEs for their targets. Many years ago, Derwan and colleagues developed custom small-molecule polymers that could be designed to recognize an arbitrary DNA sequence on the minor groove side, with the aim of directing DNA cleaving or damaging agents to specific target sequences. N-methylimidazole, N-methylpyrrole, and N-methyl-3-hydroxypyrrole polyamide hairpins (Trauger et al., 1996; Kielkopf et al., 1998a,b; White et al., 1998) bind exclusively to the minor groove of DNA and should not clash with TALE DNA-binding domains. Whether TALEs and minor groove-binding polymers can really interact with DNA-independently, or whether they require mutually exclusive subtle DNA deformations, remains to be tested experimentally.

Linkage to effector domains, particularly nucleases

The current structures of natural and synthetic DNA binding domains do not shed light on the other functionally important regions of TALE proteins, such as the type III export signals, or the activation domains (which might be fully or partially unstructured in the absence of their binding partners). The current structures also do not provide an explanation for the empirical rules that have been obtained for the ‘best’ way to fuse FokI nuclease domains to TALE DNA-binding domains and for the best separation between target half sites. In the case of zinc finger proteins, fusions to FokI are most popular. However, other domains have also resulted in useful fusion proteins, both for introducing double strand breaks into DNA (Schierling et al., 2012) and for other reactions, such as

DNA recombination (Gordley et al., 2007, 2009; Gersbach et al., 2011; Prorocic et al., 2011). It can be expected that structural understanding of TALE proteins will now spur the development of artificial sequence specific enzymes with a wide variety of applications.

Acknowledgments: This review is dedicated to Robert Huber and Wolfram Bode on the occasion of their 75th and 70th birthdays. The work was begun as a prediction of the structural basis of the TALE cipher and converted into a review when the crystal structures appeared. Thomas Lahaye (Ludwig-Maximilians-Universität München, Germany) and Alfred Pingoud (Justus Liebig Universität Gießen, Germany) introduced the author to TALEs and commented critically on an early version of the manuscript. Honorata Czapinska proofread and commented on all versions of the text. The author is funded by the Foundation for Polish Science, EU European Regional Development Fund (TEAM/2010-6/1), the European Union (EC FP7, HEALTH-PROT, GA No 229676) and the National Science Centre (NCN, UMO-2011/02/A/NZ1/00052). These contributions are gratefully acknowledged.

Received March 21, 2012; accepted July 23, 2012

References

Bitinaite, J., Wah, D.A., Aggarwal, A.K., and Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. Proc. Natl. Acad. Sci. USA *95,* 10570–10575.

Blatch, G.L. and Lassle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. Bioessays *21,* 932–939.

Boch, J. and Bonas, U. (2010). Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu. Rev. Phytopathol. *48,* 419–436.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. Science *326,* 1509–1512.

Bogdanove, A.J. and Voytas, D.F. (2011). TAL effectors: customizable proteins for DNA targeting. Science *333,* 1843–1846.

Bogdanove, A.J., Schornack, S., and Lahaye, T. (2010). TAL effectors: finding plant genes for disease and defense. Curr. Opin. Plant Biol. *13,* 394–401.

Bonas, U., Stall, R.E., and Staskawicz, B. (1989). Genetic and structural characterization of the avirulence gene avrBs3 from Xanthomonas campestris pv. vesicatoria. Mol. Gen. Genet. *218,* 127–136.

Bradley, P. (2012). Structural modeling of TAL Effector-DNA Interactions. Protein Sci. *21,* 471–474.

Bultmann, S., Morbitzer, R., Schmidt, C.S., Thanisch, K., Spada, F., Elsaesser, J., Lahaye, T., and Leonhardt, H. (2012). Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. Nucleic Acids Res. *40,* 5368–5377.

Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res. *39,* e82.

Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. Genetics *186*, 757–761.

D'Andrea, L.D. and Regan, L. (2003). TPR proteins: the versatile helix. Trends Biochem. Sci. *28,* 655–662.

Dangl, J.L. and Jones, J.D. (2001). Plant pathogens and integrated defence responses to infection. Nature *411,* 826–833.

de Francesco, L. (2011). Move over ZFNs. Nat. Biotechnol. *29,* 681–684.

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.K., Shi, Y., and Yan, N. (2012). Structural basis for sequence-specific recognition of DNA by TAL effectors. Science *335,* 720–723.

Doyle, E.L., Booher, N.J., Standage, D.S., Voytas, D.F., Brendel, V.P., Vandyk, J.K., and Bogdanove, A.J. (2012). TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. Nucleic Acids Res. *40 (Web Server issue),* W117–W122.

Garg, A., Lohmueller, J.J., Silver, P.A., and Armel, T.Z. (2012). Engineering synthetic TAL effectors with orthogonal target sites. Nucleic Acids Res. May 11 [epub ahead of print].

Geissler, R., Scholze, H., Hahn, S., Streubel, J., Bonas, U., Behrens, S.E., and Boch, J. (2011). Transcriptional activators of human genes with programmable DNA-specificity. PLoS One *6,* e19509.

Gersbach, C.A., Gaj, T., Gordley, R.M., Mercer, A.C., and Barbas, C.F. 3rd. (2011). Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase. Nucleic Acids Res. *39,* 7868–7878.

Gohre, V. and Robatzek, S. (2008). Breaking the barriers: microbial effector molecules subvert plant immunity. Annu. Rev. Phytopathol. *46,* 189–215.

Gordley, R.M., Smith, J.D., Graslund, T., and Barbas, C.F. 3rd. (2007). Evolution of programmable zinc finger-recombinases with activity in human cells. J. Mol. Biol. *367,* 802–813.

Gordley, R.M., Gersbach, C.A., and Barbas, C.F. 3rd. (2009) Synthesis of programmable integrases. Proc. Natl. Acad. Sci. USA *106,* 5053–5058.

Hirano, T., Kinoshita, N., Morikawa, K., and Yanagida, M. (1990). Snap helix with knob and hole: essential repeats in S. pombe nuclear protein nuc2+. Cell *60,* 319–328.

Hockemeyer, D., Wang, H., Kiani, S., Lai, C.S., Gao, Q., Cassady, J.P., Cost, G.J., Zhang, L., Santiago, Y., Miller, J.C., et al. (2011). Genetic engineering of human pluripotent cells using TALE nucleases. Nat. Biotechnol. *29,* 731–734.

Huang, P., Xiao, A., Zhou, M., Zhu, Z., Lin, S., and Zhang, B. (2011). Heritable gene targeting in zebrafish using customized TALENs. Nat. Biotechnol. *29,* 699–700.

Jeltsch, A. (2010). Molecular biology. Phylogeny of methylomes. Science *328,* 837–838.

Jones, J.D. and Dangl, J.L. (2006). The plant immune system. Nature *444,* 323–329.

Kay, S., Hahn, S., Marois, E., Hause, G., and Bonas, U. (2007). A bacterial effector acts as a plant transcription factor and induces a cell size regulator. Science *318,* 648–651.

Keen, N.T. (1990). Gene-for-gene complementarity in plant-pathogen interactions. Annu. Rev. Genet. *24,* 447–463.

Kielkopf, C.L., Baird, E.E., Dervan, P.B., and Rees, D.C. (1998a). Structural basis for G.C recognition in the DNA minor groove. Nat. Struct. Biol. *5,* 104–109.

Kielkopf, C.L., White, S., Szewczyk, J.W., Turner, J.M., Baird, E.E., Dervan, P.B., and Rees, D.C. (1998b). A structural basis for recognition of A.T and T.A base pairs in the minor groove of B-DNA. Science *282,* 111–115.

Li, T., Huang, S., Jiang, W.Z., Wright, D., Spalding, M.H., Weeks, D.P., and Yang, B. (2011). TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res. *39,* 359–372.

Li, L., Piatek, M.J., Atef, A., Piatek, A., Wibowo, A., Fang, X., Sabir, J.S., Zhu, J.K., and Mahfouz, M.M. (2012). Rapid and highly efficient construction of TALE-based transcriptional regulators and nucleases for genome modification. Plant Mol. Biol. *78,* 407–416.

Liu, J., Li, C., Yu, Z., Huang, P., Wu, H., Wei, C., Zhu, N., Shen, Y., Chen, Y., Zhang, B, et al. (2012). Efficient and specific modifications of the Drosophila genome by means of an easy TALEN strategy. J. Genet. Genomics *39,* 209–215.

Mahfouz, M.M., Li, L., Shamimuzzaman, M., Wibowo, A., Fang, X., and Zhu, J.K. (2011). De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. Proc. Natl. Acad. Sci. USA *108,* 2623–2628.

Main, E.R., Xiong, Y., Cocco, M.J., D'Andrea, L., and Regan, L. (2003). Design of stable alpha-helical arrays from an idealized TPR motif. Structure *11,* 497–508.

Mak, A.N., Bradley, P., Cernadas, R.A., Bogdanove, A.J., and Stoddard, B.L. (2012). The crystal structure of TAL effector PthXo1 bound to Its DNA target. Science *335*, 716–719.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2011). A TALE nuclease architecture for efficient genome editing. Nat. Biotechnol. *29,* 143–148.

Morbitzer, R., Romer, P., Boch, J., and Lahaye, T. (2010). Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. Proc. Natl. Acad. Sci. USA *107,* 21617–21622.

Morbitzer, R., Elsaesser, J., Hausner, J., and Lahaye, T. (2011). Assembly of custom TALE-type DNA binding domains by modular cloning. Nucleic Acids Res. *39,* 5790–5799.

Moscou, M.J. and Bogdanove, A.J. (2009). A simple cipher governs DNA recognition by TAL effectors. Science *326,* 1501.

Murakami, M.T., Sforca, M.L., Neves, J.L., Paiva, J.H., Domingues, M.N., Pereira, A.L., Zeri, A.C., and Benedetti, C.E. (2010). The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction. Proteins *78,* 3386–3395.

Mussolino, C., Morbitzer, R., Lutge, F., Dannemann, N., Lahaye, T., and Cathomen, T. (2011). A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic Acids Res. *39,* 9283–9293.

Pace, N.R. (1997). A molecular view of microbial diversity and the biosphere. Science *276,* 734–740.

Prorocic, M.M., Wenlong, D., Olorunniji, F.J., Akopian, A., Schloetel, J.-G., Hannigan, A., McPherson, A.L., and Stark, W.M. (2011). Zinc-finger recombinase activities in vitro. Nucleic Acids Res. *39,* 9316–9328.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., and Joung, J.K. (2012). FLASH assembly of TALENs for high-throughput genome editing. Nat. Biotechnol. *30,* 460–465.

Romer, P., Hahn, S., Jordan, T., Strauss, T., Bonas, U., and Lahaye, T. (2007). Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. Science *318,* 645–648.

Romer, P., Recht, S., and Lahaye, T. (2009a). A single plant resistance gene promoter engineered to recognize multiple TAL effectors from disparate pathogens. Proc. Natl. Acad. Sci. USA *106,* 20526–20531.

Romer, P., Strauss, T., Hahn, S., Scholze, H., Morbitzer, R., Grau, J., Bonas, U., and Lahaye, T. (2009b). Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper Bs3 alleles. Plant Physiol. *150,* 1697–1712.

Romer, P., Recht, S., Strauss, T., Elsaesser, J., Schornack, S., Boch, J., Wang, S., and Lahaye, T. (2010). 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. *187,* 1048–1057.

Sander, J.D., Cade, L., Khayter, C., Reyon, D., Peterson, R.T., Joung, J.K., and Yeh, J.R. (2011). Targeted gene disruption in somatic zebrafish cells using engineered TALENs. Nat. Biotechnol. *29,* 697–698.

Sanjana, N.E., Cong, L., Zhou, Y., Cunniff, M.M., Feng, G., and Zhang, F. (2012). A transcription activator-like effector toolbox for genome engineering. Nat. Protoc. *7,* 171–192.

Schierling, B., Dannemann, N., Gabsalilow, L., Wende, W., Cathomen, T., and Pingoud, A. (2012). A novel zinc-finger nuclease platform with a sequence-specific cleavage module. Nucleic Acids Res. *40,* 2623–2638.

Scholze, H. and Boch, J. (2011). TAL effectors are remote controls for gene activation. Curr. Opin. Microbiol. *14,* 47–53.

Schornack, S., Meyer, A., Romer, P., Jordan, T., and Lahaye, T. (2006). Gene-for-gene-mediated recognition of nuclear-targeted AvrBs3-like bacterial effector proteins. J. Plant Physiol. *163,* 256–272.

Sikorski, R.S., Boguski, M.S., Goebl, M., and Hieter, P. (1990). A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. Cell *60,* 307–317.

Szurek, B., Rossier, O., Hause, G., and Bonas, U. (2002). Type III-dependent translocation of the Xanthomonas AvrBs3 protein into the plant cell. Mol. Microbiol. *46,* 13–23.

Tesson, L., Usal, C., Menoret, S., Leung, E., Niles, B.J., Remy, S., Santiago, Y., Vincent, A.I., Meng, X., Zhang, L., et al. (2011). Knockout rats generated by embryo microinjection of TALENs. Nat. Biotechnol. *29,* 695–696.

Tong, C., Huang, G., Ashton, C., Wu, H., Yan, H., and Ying, Q.L. (2012). Rapid and cost-Effective Gene Targeting in Rat Embryonic Stem Cells by TALENs. J. Genet. Genomics *39,* 275–280.

Trauger, J.W., Baird, E.E., and Dervan, P.B. (1996). Recognition of DNA by designed ligands at subnanomolar concentrations. Nature *382,* 559–561.

Van den Ackerveken, G. and Bonas, U. (1997). Bacterial avirulence proteins as triggers of plant disease resistance. Trends Microbiol. *5,* 394–398.

Vivian, A. and Mansfield, J. (1993). A proposal for a uniform genetic nomenclature for avirulence genes in phytopathogenic pseudomonads. Mol. Plant-Microbe Interact. *6,* 9–10.

Weber, E., Gruetzner, R., Werner, S., Engler, C., and Marillonnet, S. (2011). Assembly of designer TAL effectors by Golden Gate cloning. PLoS One *6,* e19722.

White, S., Szewczyk, J.W., Turner, J.M., Baird, E.E., and Dervan, P.B. (1998). Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands. Nature *391,* 468–471.

Wood, A.J., Lo, T.W., Zeitler, B., Pickle, C.S., Ralston, E.J., Lee, A.H., Amora, R., Miller, J.C., Leung, E., Meng, X., et al. (2011). Targeted genome editing across species using ZFNs and TALENs. Science *333,* 307.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G.M., and Arlotta, P. (2011). Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat. Biotechnol. *29,* 149–153.

Matthias Bochtler completed a PhD with Prof. Robert Huber at TU Munich and the Max-Planck-Institute of Biochemistry on bacterial model systems of the proteasome. He has since been a group leader at the Max-Planck-Institute of Molecular and Cell Biology running an outstation at the International Institute of Molecular and Cell Biology (IIMCB) in Warsaw, Poland, and at Cardiff University in Wales, UK. In 2011, he returned as a group leader to the IIMCB and now also holds a professorship at the neighboring Institute of Biochemistry and Biophysics (IBB) of the Polish Academy of Sciences. His current research interests center on the structural biology of protein nucleic acid interactions and on DNA methylation.