Recognition of Nucleobases

DOI: 10.1002/anie.201400436

Programmable and Highly Resolved In Vitro Detection of 5-Methylcytosine by TALEs**

*Grzegorz Kubik, Moritz J. Schmidt, Johanna E. Penner, and Daniel Summerer**

Abstract: *Gene expression is extensively regulated by specific patterns of genomic 5-methylcytosine (mC), but the ability to directly detect this modification at user-defined genomic loci is limited. One reason is the lack of molecules that discriminate between mC and cytosine (C) and at the same time provide inherent, programmable sequence-selectivity. Programmable transcription-activator-like effectors (TALEs) have been observed to exhibit mC-sensitivity in vivo, but to only a limited extent in vitro. We report an mC-detection assay based on TALE control of DNA replication that displays unexpectedly strong mC-discrimination ability in vitro. The status and level of mC modification at single positions in oligonucleotides can be determined unambiguously by this assay, independently of the overall target sequence. Moreover, discrimination is reliably observed for positions bound by N-terminal and central regions of TALEs. This indicates the wide scope and robustness of the approach for highly resolved mC detection and enabled the detection of a single mC in a large, eukaryotic genome.*

In 5-position methylated cytosine (mC) is an epigenetic DNA modification with vital roles in the regulation of gene expression, genome stability, and disease.[1] Methods for the direct assessment of the status and level of mC at user-defined genomic loci are thus of broad interest for diagnosis and therapy.[2] This requires effective strategies for the discrimination between mC and cytosine (C) without sequence constraints.[3] Indirect, chemical discrimination has been described based on the differential redox reactivity of the C5–C6 double bond in mC and C[4] and on the selective deamination of C with bisulfite.[5] However, these approaches, in addition to harsh reaction conditions, suffer from limited resolution[4] or a detrimental reduction in sequence complexity.[3] Proteins capable of the direct recognition of mC could circumvent these drawbacks. However, currently employed proteins are either not sequence-selective[6] or exhibit sequence constraints.[7] Single-molecule techniques using protein nanopores and DNA polymerases offer high resolution,[8] but are technically demanding and, for locus-specific analysis, require sequence-enrichment methods that are independent of mC-erasing amplification steps.[9] Hence, there is a need for proteins that can be programmed to recognize any desired DNA sequence and directly report the status and level of mC modification.

A few years ago, transcription-activator-like effectors (TALEs) emerged as a new scaffold for the design of DNA-binding domains with user-defined sequence selectivity.[10] TALEs consist of multiple repeats that each selectively recognize one nucleotide through one of two variable amino acids (repeat variable di-residue, RVD). This recognition follows a simple code[11] with the RVDs NG, HD, NI, and NN preferentially binding T, C, A and G nucleobases, respectively. RVD NG binds T (or mC) via the 5-methyl group through a hydrophobic interaction with the C_α-methylene moiety of glycine (Figure 1 A). RVD HD interacts with C through a hydrogen bond between the aspartate carboxy group and the C 4-amino group.[12] Interactions of both NG with mC[12c] and HD with C can be perturbed by the respective converse methylation status of the bound nucleotide. Indeed, the mC sensitivity of TALE proteins has been reported as a restriction of TALE-based in vivo genome engineering and transcriptional regulation; however, only limited mC sensitivity has been reported in vitro.[12c,13]

We aimed to get detailed insights into this sensitivity and to exploit it for programmable, locus-specific mC detection in vitro. We designed and expressed TALE_97, a construct targeting a 17 nt sequence of the zebrafish (*Danio rerio*) *hey2* gene.[14] To evaluate the general mC-discrimination ability of TALE_97, we analyzed its binding to DNA containing its target sequence ("97", Figure 1 B) with either C or mC at six positions opposite six NG or HD RVDs in electromobility shift assays (EMSA, Figure 1 C).

No difference in binding was observed when either C- or mC-containing nucleotides were placed opposite NG RVDs (data not shown). In contrast, for HD RVDs, binding was observed for DNA containing C, but not for DNA containing mC, consistent with previous studies (Figure 1 C).[12c,13b]

We next asked, whether this discrimination could be exploited for programmable and highly resolved mC detection. For that purpose, we aimed to control DNA synthesis by DNA polymerase through mC-dependent, inhibitive TALE binding. Newly synthesized DNA molecules could then serve as a signal that could be quantified by selective and sensitive DNA detection methods (Figure 2 A).

[*] M. Sc. G. Kubik, M. Sc. M. J. Schmidt, M. Sc. J. E. Penner, Dr. D. Summerer
Department of Chemistry and Konstanz Research School Chemical Biology, University of Konstanz
Universitätsstrasse 10, 78457 Konstanz (Germany)
E-mail: daniel.summerer@uni-konstanz.de

[**] We acknowledge support by the Zukunftskolleg of the University of Konstanz and the Konstanz Research School Chemical Biology. We thank Axel Meyer for zebrafish, Pamela Bjorkman for plasmid pET32a-TRXtag, Keith Joung for plasmids TALEN1257, 1258, and 1297, and Adam J. Bogdanove and Daniel F. Voytas for TALE assembly plasmids. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (SPP1623). TALE: transcription-activator-like effector.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201400436.

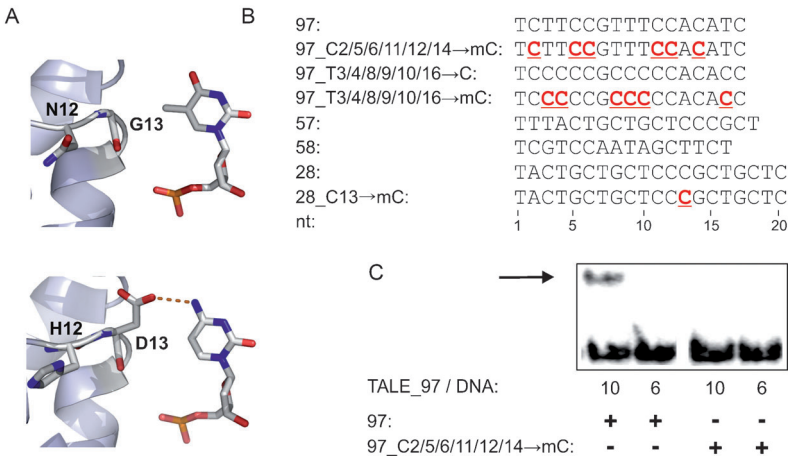

Figure 1. Recognition of thymine, C, and mC in DNA by TALEs. A) Interaction of RVD NG with thymine (top) and of RVD HD with C (bottom) in a crystal structure (pdb entry 3V6T).[12a] The hydrogen bond is shown as a dotted line. B) Sequences of the target DNA and its variants containing mC. C) EMSA using 6.3 nM 5′-^{32}P-labeled DNA containing either C or mC as shown in (B) and 63 or 37.8 nM TALE_97 resulting in a TALE/target DNA ratio of 10 and 6, respectively. TALE–DNA complex is marked with an arrow.

We hybridized a DNA oligonucleotide template consisting of a region of the *hey2* gene with sequence 97 (Figure 1B) at the 3′-terminus (t97) with a reverse complement, 5′-^{32}P-labeled primer (p97_rev), and incubated the complex with or without TALE_97. We then added dNTP and the Klenow fragment of *E. coli* DNA polymerase I (3′-5′-exo$^-$, KF(exo−)), incubated the mixture, and subsequently resolved it by denaturing polyacrylamide gel electrophoresis (PAGE). Primer extension was observed when TALE_97 was absent, but was completely abolished in its presence (Figure 2B, lanes 1 and 2), indicating that this protein is indeed an effective inhibitor of KF(exo−)-catalyzed DNA synthesis. We next asked whether we could combine this inhibition potential of TALE_97 with its observed discrimination of mC opposite HD RVDs (Figure 1C) to enable mC-dependent DNA synthesis. We performed a primer extension reaction like that described above with a template that contained only three to one mC residues at the C positions 2, 6, and 14 (Figure 2B, lanes 3–7). No inhibition of primer extension was observed for positions 2 and 6, demonstrating effective discrimination of single mC positions by central and N-terminal RVDs. Inhibition was visible for position 14 which is close to the TALE C-terminus (see discussion below).

An ideal mC-detection method should not only allow the analysis of the status but also of the level of mC modification.[1,2,15] We performed primer extension reactions with TALE_97 and mixtures of t97 and t97_C6→mC as templates, resulting in mC-modification levels between 0 and 100% (Figure 2C). Dependence of primer extension efficiency on the mC-modification level of the single position was strictly linear (insert in Figure 2C), enabling quantitative, highly resolved (< 10%) analysis.

To quantify the mC-discrimination ability of TALE_97, we performed primer extension reactions as described above with varying TALE/target DNA ratios. Primer extension was abolished from a TALE_97/target DNA ratio of about 50:1 ($IC_{50} = (83 \pm 8)$ nM for 125 mU KF(exo−), Figure 2D and Figure 2 in the Supporting Information). In contrast, no significant effect was observed for t97_C6→mC even at the highest ratio, confirming highly effective discrimination.

To get a comprehensive view on the application scope of TALE proteins for locus-specific mC detection, we tested two additional TALEs based on the scaffold used for TALE_97 (TALE_57, $IC_{50} = (24 \pm 7)$ nM for 500 mU KF(exo−) and TALE_58, $IC_{50} = (67 \pm 4)$ nM for 25 mU KF(exo−), Figures 3 and 4 in the Supporting Information). These proteins recognize sequences in the *hey2* and *gria3a* gene of zebrafish that are both absent from t97 (Figure 1B).[14] We performed primer extensions as described above on the cognate templates t57 and t58 with a single C or mC at seven and four scattered positions, respectively (Figures 6 and 7 in the Supporting Information). Discrimination of a single mC was observed in all cases, regardless of the target sequence and the direct sequence context around the mC. As with TALE_97, for a single position close to the C-terminus inhibition was observed, potentially reflecting the large distance to the KF(exo−)-binding site or polarity of TALE binding[16] (Figure 6 in the Supporting Information). However, for all central and N-terminal mC positions, a success rate of 100% was observed. This indicates the broad applicability of TALEs for mC detection.

For locus-specific mC-detection, inhibition of KF(exo−) must be sequence-selective under conditions that are compatible with KF(exo−)-catalyzed DNA synthesis. In primer extension reactions using TALE_97, _57, or _58 with templates t97 or t97_C6→mC, only TALE_97 effectively discriminated the single mC position (Figure 2E). Moreover, effective mC discrimination of TALE_97 was observed in the presence of a saturating amount of complex off-target DNA (3×10^4-fold mass excess of salmon sperm DNA over t97 or t97_C6→mC, Figures 8–10 in the Supporting Information). This indicates the highly selective binding of TALE_97 to t97 under the required assay conditions.

We next aimed to transfer our approach to genomic DNA (gDNA) samples. Since this requires selective TALE binding to highly diluted targets in complex DNA, we constructed TALE_28, which targets a 20mer sequence (sequence 28 in the *hey2* gene, Figure 1B). This provides high affinity ($IC_{50} = (19 \pm 4)$ nM for 500 mU KF(exo−), Figure 5 in the Supporting Information) and sequence uniqueness. We extracted gDNA from zebrafish fin and determined an mC-modification level of 0% at position 28_C13 (Figure 1B) by bisulfite sequencing (Figures 11–14 in the Supporting Information). We then enzymatically methylated a part of this gDNA and confirmed an mC level of 100% at 28_C13→mC (Figure 14 in the Supporting Information). We individually hybridized 100 ng of both gDNAs with an excess of 5′-biotinylated primer Bio-p28_rev (0.8 nM) and performed primer extensions in the

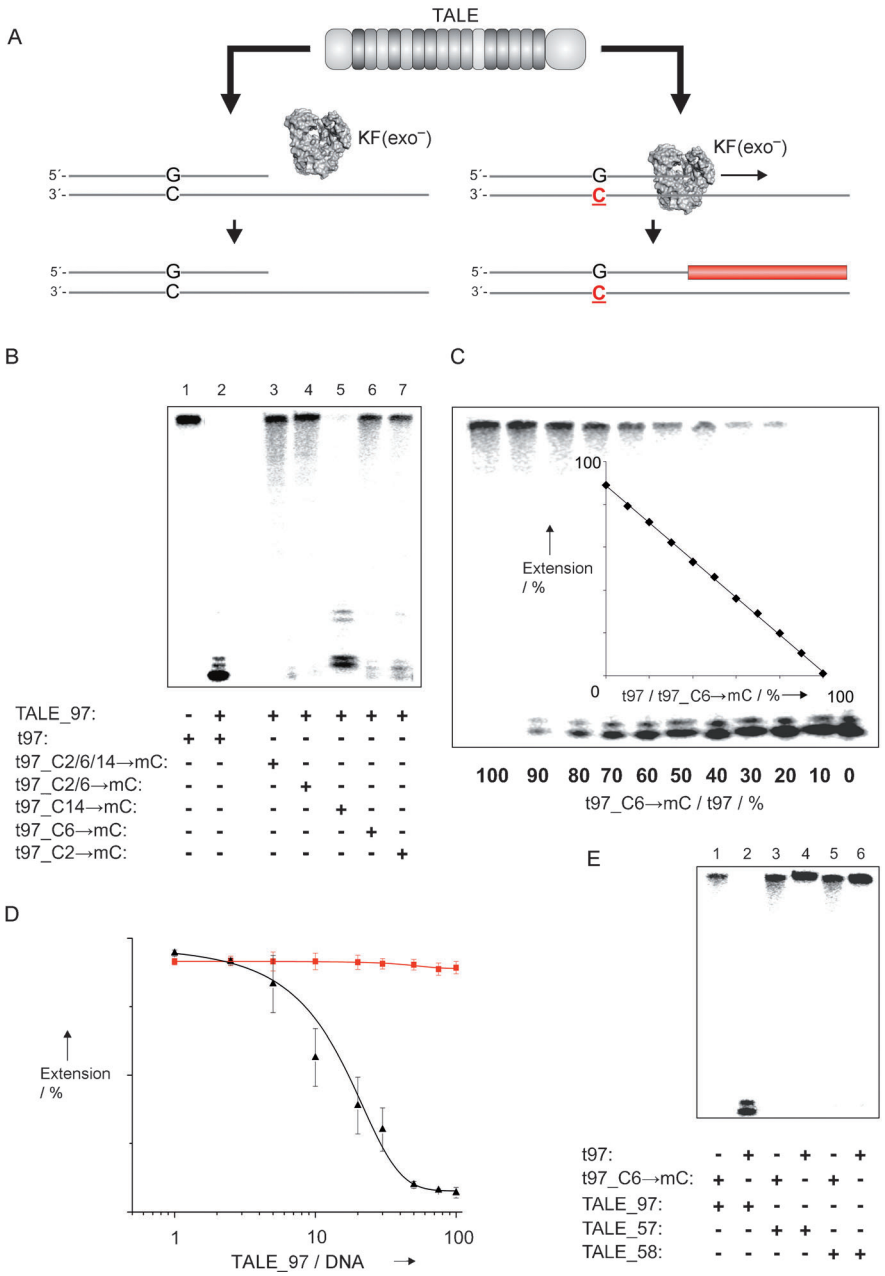

Figure 2. Highly resolved mC detection by TALE-controlled DNA replication. A) Overview. Binding of TALE to a primer-template inhibits extension by KF(exo−) in the presence of C opposite RVD HD, but not in the presence of mC (red), leading to the synthesis of DNA (red bar). B) Dependence of KF(exo−) inhibition on the number and position of mC. PAGE analysis of primer extensions containing 125 mU KF(exo−), 8.3 nM primer-template with a single mC in the presence or absence of 833 nM TALE_97. Extended primers are found on the top, not extended primers on the bottom. C) Analysis of mC level at a single position. D) Concentration dependence of KF(exo−) inhibition for t97 (black triangles) and t97_C6→mC (red squares). E) Sequence selectivity of KF(exo−) inhibition.

absence or presence of TALE_28 (1160 nM) and KF(exo−) (25 mU). We then incubated the mixtures with streptavidin beads, washed the beads under denaturing conditions to remove non-biotinylated DNA, and used them in qPCRs that targeted the primer extension product (Figure 3A, Figure 15 in the Supporting Information). Significant formation of primer extension product was only observed in presence of KF(exo−) in both cases, indicating effective hybridization and extension as well as selective isolation of biotinylated primer (Figure 3B, columns 2–5).

No significant difference in the formation of primer extension products was observed for the two gDNAs in the absence of TALE_28 (Figure 3B, columns 4 and 5; that is, methylation did not significantly affect primer extension (or qPCR efficiency, see also Figure 16 in the Supporting Information). However, a twofold reduction in product formation was observed for non-methylated gDNA compared to methylated DNA in the presence of TALE_28, indicating significant KF(exo−) inhibition at the target locus only for non-methylated gDNA (Figure 3B, columns 6 and 7). This shows that TALE_28 is able to discriminate the single mC position, and that this discrimination can be coupled with a selective, sensitive, and scalable assay read-out.

In summary, we demonstrated that TALE proteins can be employed to detect mC in a complex, eukaryotic genome. Our method offers the direct, conversion-free detection of mC with high resolution. Apart from Watson–Crick hybridization, TALEs offer the only mode of truly programmable recognition of long DNA sequences based on a simple one-to-one code. However, unlike nucleic acids, TALEs discriminate mC. The discrimination we described is significantly stronger than that previously observed in vitro and exhibits a broad scope with respect to target sequences and mC positioning within the complex. TALEs thus represent a potentially general alternative to hybridization-based nucleic acid probes in a wide range of genomic analysis techniques and offers the direct and simultaneous read-out of both the genetic and epigenetic information of DNA.

Received: January 15, 2014
Revised: March 25, 2014
Published online: May 6, 2014

Keywords: epigenetics · 5-methylcytosine · programmable protein scaffolds · protein–DNA interactions · TAL effectors

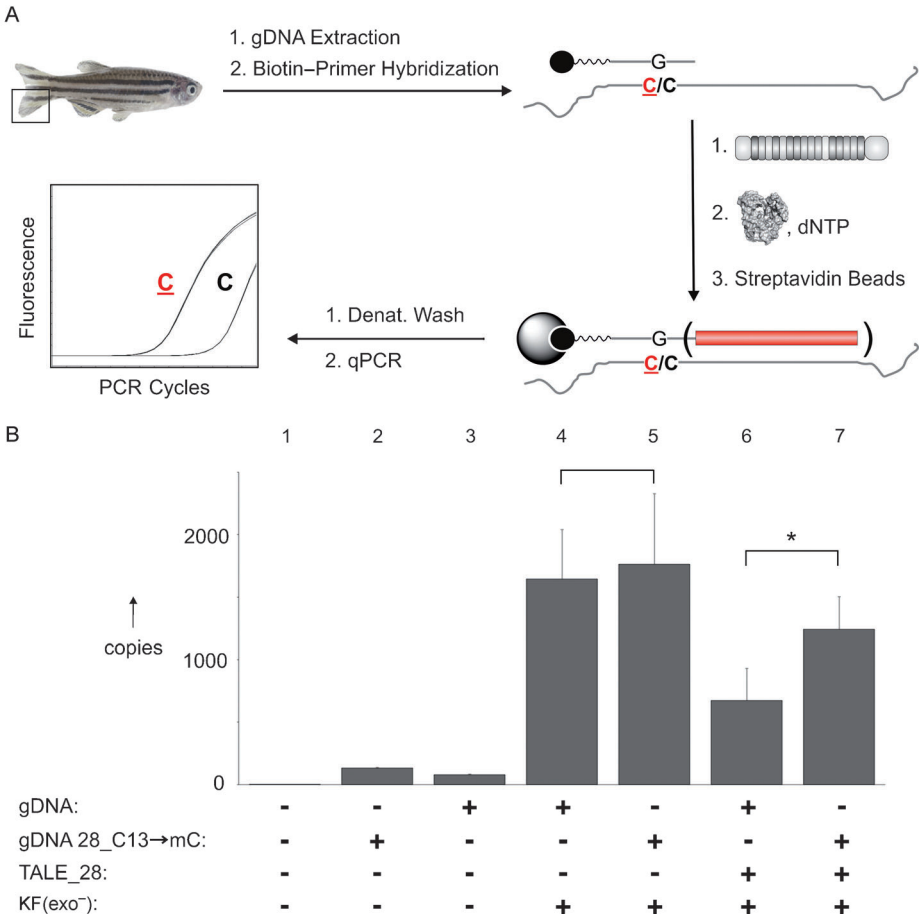

Figure 3. Highly resolved mC-detection in genomic DNA. A) Overview. Biotin is shown as a black dot, streptavidine bead as a gray sphere, other elements as in Figure 2. B) mC-dependent, TALE-controlled primer extension. *: $p < 0.05$ from Student T test (see the Supporting Information).

[1] J. A. Law, S. E. Jacobsen, *Nat. Rev. Genet.* **2010**, *11*, 204–220.

[2] M. Rodríguez-Paredes, M. Esteller, *Nat. Med.* **2011**, *17*, 330–339.

[3] a) E. Olkhov-Mitsel, B. Bapat, *Cancer med.* **2012**, *1*, 237–260; b) P. W. Laird, *Nat. Rev. Genet.* **2010**, *11*, 191–203.

[4] a) T. Wang, T. Hong, T. Tang, Q. Zhai, X. Xing, W. Mao, X. Zheng, L. Xu, J. Wu, X. Weng, S. Wang, T. Tian, B. Yuan, B. Huang, L. Zhuang, X. Zhou, *J. Am. Chem. Soc.* **2013**, *135*, 1240–1243; b) S. Bareyt, T. Carell, *Angew. Chem.* **2008**, *120*, 187–190; *Angew. Chem. Int. Ed.* **2008**, *47*, 181–184; c) K. Tanaka, K. Tainaka, T. Kamei, A. Okamoto, *J. Am. Chem. Soc.* **2007**, *129*, 5612–5620; d) K. Tanabe, H. Yamada, S. I. Nishimoto, *J. Am. Chem. Soc.* **2007**, *129*, 8034–8040.

[5] M. Frommer, L. E. Mcdonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, C. L. Paul, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1827–1831.

[6] a) M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, D. Schubeler, *Nat. Genet.* **2005**, *37*, 853–862; b) T. A. Down, V. K. Rakyan, D. J. Turner, P. Flicek, H. Li, E. Kulesha, S. Graf, N. Johnson, J. Herrero, E. M. Tomazou, N. P. Thorne, L. Backdahl, M. Herberth, K. L. Howe, D. K. Jackson, M. M. Miretti, J. C. Marioni, E. Birney, T. J. Hubbard, R. Durbin, S. Tavare, S. Beck, *Nat. Biotechnol.* **2008**, *26*, 779–785; c) D. Zilberman, M. Gehring, R. K. Tran, T. Ballinger, S. Henikoff, *Nat. Genet.* **2007**, *39*, 61–69.

[7] a) B. Khulan, R. F. Thompson, K. Ye, M. J. Fazzari, M. Suzuki, E. Stasiek, M. E. Figueroa, J. L. Glass, Q. Chen, C. Montagna, E. Hatchwell, R. R. Selzer, T. A. Richmond, R. D. Green, A. Melnick, J. M. Greally, *Genome Res.* **2006**, *16*, 1046–1055; b) M. P. Ball, J. B. Li, Y. Gao, J. H. Lee, E. M. LeProust, I. H. Park, B. Xie, G. Q. Daley, G. M. Church, *Nat. Biotechnol.* **2009**, *27*, 361–368; c) R. A. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. A. Brandenburg, J. A. Jeddeloh, B. Wen, A. P. Feinberg, *Genome Res.* **2008**, *18*, 780–790; d) S. H. Cross, J. A. Charlton, X. Nan, A. P. Bird, *Nat. Genet.* **1994**, *6*, 236–244; e) L. G. Acevedo, A. Sanz, M. A. Jelinek, *Epigenomics* **2011**, *3*, 93–101; f) C. I. Stains, J. L. Furman, D. J. Segal, I. Ghosh, *J. Am. Chem. Soc.* **2006**, *128*, 9761–9765; g) L. Zhang, K. E. Szulwach, G. C. Hon, C. X. Song, B. Park, M. Yu, X. Lu, Q. Dai, X. Wang, C. R. Street, H. Tan, J. H. Min, B. Ren, P. Jin, C. He, *Nat. Commun.* **2013**, *4*, 1517.

[8] a) J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid, H. Bayley, *Nat. Nanotechnol.* **2009**, *4*, 265–270; b) B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach, S. W. Turner, *Nat. Methods* **2010**, *7*, 461–465; c) D. Summerer, *ChemBioChem* **2010**, *11*, 2499–2501.

[9] a) T. J. Albert, M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, T. A. Richmond, C. M. Middle, M. J. Rodesch, C. J. Packard, G. M. Weinstock, R. A. Gibbs, *Nat. Methods* **2007**, *4*, 903–905; b) D. Summerer, H. Wu, B. Haase, Y. Cheng, N. Schracke, C. F. Stahler, M. S. Chee, P. F. Stahler, M. Beier, *Genome Res.* **2009**, *19*, 1616–1621; c) A. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, C. Nusbaum, *Nat. Biotechnol.* **2009**, *27*, 182–189; d) D. Summerer, *Genomics* **2009**, *94*, 363–368.

[10] a) J. Boch, U. Bonas, *Annu. Rev. Phytopathol.* **2010**, *48*, 419–436; b) A. J. Bogdanove, D. F. Voytas, *Science* **2011**, *333*, 1843–1846.

[11] a) M. J. Moscou, A. J. Bogdanove, *Science* **2009**, *326*, 1501; b) J. Boch, H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt, U. Bonas, *Science* **2009**, *326*, 1509–1512.

[12] a) D. Deng, C. Yan, X. Pan, M. Mahfouz, J. Wang, J. K. Zhu, Y. Shi, N. Yan, *Science* **2012**, *335*, 720–723; b) A. N. S. Mak, P. Bradley, R. A. Cernadas, A. J. Bogdanove, B. L. Stoddard, *Science* **2012**, *335*, 716–719; c) D. Deng, P. Yin, C. Yan, X. Pan, X. Gong, S. Qi, T. Xie, M. Mahfouz, J. K. Zhu, N. Yan, Y. Shi, *Cell Res.* **2012**, *22*, 1502–1504.

[13] a) S. Bultmann, R. Morbitzer, C. S. Schmidt, K. Thanisch, F. Spada, J. Elsaesser, T. Lahaye, H. Leonhardt, *Nucleic Acids Res.* **2012**, *40*, 5368–5377; b) J. Valton, A. Dupuy, F. Daboussi, S. Thomas, A. Marechal, R. Macmaster, K. Melliand, A. Juillerat, P. Duchateau, *J. Biol. Chem.* **2012**, *287*, 38427–38432; c) Y. Kim, J. Kweon, A. Kim, J. K. Chon, J. Y. Yoo, H. J. Kim, S. Kim, C.

Lee, E. Jeong, E. Chung, D. Kim, M. S. Lee, E. M. Go, H. J. Song, H. Kim, N. Cho, D. Bang, S. Kim, J. S. Kim, *Nat. Biotechnol.* **2013**, *31*, 251 – 258; d) A. Dupuy, J. Valton, S. Leduc, J. Armier, R. Galetto, A. Gouble, C. Lebuhotel, A. Stary, F. Paques, P. Duchateau, A. Sarasin, F. Daboussi, *PLoS One* **2013**, *8*, e78678.

[14] J. D. Sander, L. Cade, C. Khayter, D. Reyon, R. T. Peterson, J. K. Joung, J. R. J. Yeh, *Nat. Biotechnol.* **2011**, *29*, 697 – 698.

[15] R. Bonasio, S. J. Tu, D. Reinberg, *Science* **2010**, *330*, 612 – 616.

[16] J. F. Meckler, M. S. Bhakta, M. S. Kim, R. Ovadia, C. H. Habrian, A. Zykovich, A. Yu, S. H. Lockwood, R. Morbitzer, J. Elsaesser, T. Lahaye, D. J. Segal, E. P. Baldwin, *Nucleic Acids Res.* **2013**, *41*, 4118 – 4128.

