Chemical Genomics Hot Paper

DOI: 10.1002/anie.201405497

Mapping Polyamide–DNA Interactions in Human Cells Reveals a New Design Strategy for Effective Targeting of Genomic Sites**

*Graham S. Erwin, Devesh Bhimsaria, Asuka Eguchi, and Aseem Z. Ansari**

Abstract: *Targeting the genome with sequence-specific synthetic molecules is a major goal at the interface of chemistry, biology, and personalized medicine. Pyrrole/imidazole-based polyamides can be rationally designed to target specific DNA sequences with exquisite precision in vitro; yet, the biological outcomes are often difficult to interpret using current models of binding energetics. To directly identify the binding sites of polyamides across the genome, we designed, synthesized, and tested polyamide derivatives that enabled covalent crosslinking and localization of polyamide–DNA interaction sites in live human cells. Bioinformatic analysis of the data reveals that clustered binding sites, spanning a broad range of affinities, best predict occupancy in cells. In contrast to the prevailing paradigm of targeting single high-affinity sites, our results point to a new design principle to deploy polyamides and perhaps other synthetic molecules to effectively target desired genomic sites in vivo.*

A major goal at the interface of chemistry, biology, and personalized medicine is the design of small molecules that selectively target the genome to perturb, and rectify, malfunctioning gene regulatory networks. The greatest success in designing molecules with programmable DNA-binding specificity has been with polyamides.[1] Polyamides that contain N-methylpyrrole and N-methylimidazole can be rationally designed to bind to specific sequences in the minor groove of DNA.[2] Polyamides can bind to specific sequences with nanomolar affinity,[3] and unlike most protein-based DNA-binding domains, they retain their affinity and specificity when binding to methylated[4] and chromatinized[5] DNA. Polyamides can also efficiently target viral DNA for degradation,[6] and they can traverse the cell membrane and traffic to the nucleus to modulate gene expression.[7]

To comprehensively examine the specificity of DNA-binding molecules, we previously developed a high-throughput platform to monitor polyamide binding to every possible sequence variant, up to 12 base pairs (bp) in length.[3,8] The cognate site identifier (CSI) method was used to determine the specificity and affinity of several hairpin and linear polyamides. Because the CSI binding intensity is directly proportional to the association constant (K_a) for a given DNA sequence,[3,8a] the intensity data can be used to assign binding probabilities to every sequence across the genome. These genome-wide binding maps, called genomescapes,[3a] predict thousands of polyamide binding sites of maximal and varying affinity, yet only a small subset of these binding sites perturb gene expression.[7a,b,9] Moreover, analysis of differential gene expression after polyamide treatment of live cells reveals that the degree of transcriptional perturbation varies considerably from gene to gene.[7a,b,9] For example, polyamide **1**, which was designed to target the hypoxia-responsive element (HRE), competes with endogenous transcription factors for binding to HRE and thereby reduces the expression of a target gene, VEGF. Yet a different gene, ET-2, with an HRE of a similar predicted binding energy, was downregulated 10 times more than VEGF.[7a] The basis for such variable impact on the transcription of specific genes when targeting energetically similar sites remains poorly understood.

The size and packaging of the genome in the cell nucleus could impact both the specificity and the accessibility of binding sites. In chromatin, DNA is packaged into nucleosomes with 146 bp of DNA wrapped around a histone octamer. The nucleosome can occlude binding sites and interfere with binding of natural factors or synthetic molecules.[10] Genomic DNA is further compacted in the nucleus, consistent with a new report stating that the in vivo chromatin landscape determines the accessibility of a polyamide for its target sites in live cells.[11]

To investigate how genomic architecture and the local chromatin landscape influence polyamide occupancy in live cells, we mapped polyamide binding across the human genome. In stark contrast to the current paradigm of single high-affinity site targeting, we find that the occurrence of multiple clustered binding sites, even suboptimal sites that display low to moderate affinity, correlates best with polyamide occupancy in live human cells. These data suggest a new design principle for in vivo genome targeting by polyamides and perhaps other DNA-binding small molecules and therapeutic agents.

Several methods have recently been developed to study interactions between small molecules and nucleic acid targets (DNA or RNA) in a cellular environment.[12] To study polyamide binding in the genome, we devised an approach that we term the “crosslinking of small molecules for isolation of chromatin” (COSMIC). We designed and synthesized

[*] G. S. Erwin, D. Bhimsaria, A. Eguchi, Prof. A. Z. Ansari
Department of Biochemistry and The Genome Center
University of Wisconsin - Madison
Madison, WI 53706 (USA)
E-mail: ansari@biochem.wisc.edu

[**] We thank Prof. Peter Dervan, Dr. José Rodriguez-Martinez, Prof. Parmesh Ramanathan, and members of the Ansari lab for helpful discussions. This work was supported by NIH grants CA133508 and HL099773, the H. I. Romnes faculty fellowship, and the W. M. Keck Medical Research Award to A.Z.A. G.S.E. was supported by Molecular Biosciences Training Grant NIH T32 GM07215. A.E. was supported by the Morgridge Graduate Fellowship, and D.B. was supported by the NSEC grant from NSF.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201405497.

trifunctional derivatives of bioactive polyamides (**3** and **6**, Figure 1). These compounds consist of a DNA-targeting polyamide, a photo-crosslinker (psoralen), and an affinity handle (biotin). Polyamides were synthesized by Boc solid-phase synthesis (Boc = *tert*-butoxycarbonyl),[13] cleaved from the resin, and conjugated to the psoralen–biotin moiety **PB** (active ester) to give **3** and **6** (Figure 1). Both **3** and **6** were purified by reverse-phase HPLC and characterized for purity and identity by analytical HPLC and MALDI-TOF mass spectrometry (Supporting Information, Figure S1).

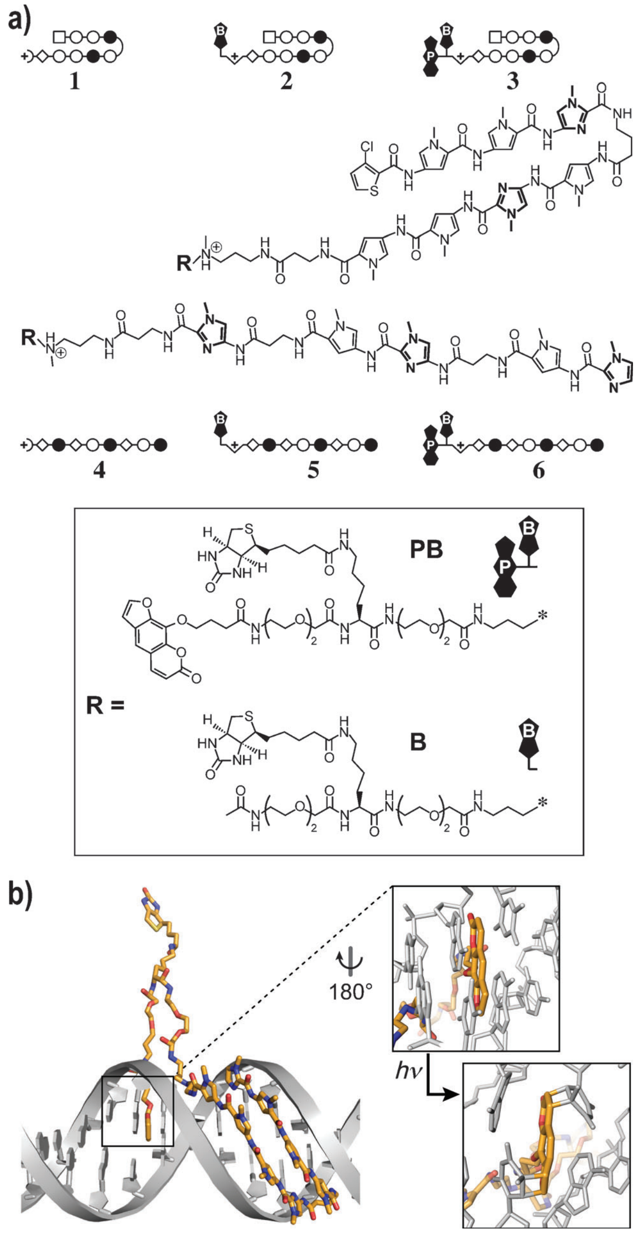

Figure 1. Trifunctional polyamides. a) Molecules employed in this study. Hairpin polyamides **1**–**3** target the DNA sequence 5′-WACGTW-3′. Linear polyamides **4**–**6** target 5′-AAGAAGAAG-3′. Rings of *N*-methylimidazole in bold for clarity. Open and filled circles represent *N*-methylpyrrole and *N*-methylimidazole, respectively. Squares represent 3-chlorothiophene, and diamonds represent β-alanine. Psoralen and biotin are denoted by P and B, respectively. b) Model of **3** crosslinking to DNA. The polyamide binds to the minor groove of DNA and psoralen intercalates between two base pairs. Upon irradiation with 365 nm light, psoralen crosslinks to thymine bases on opposite strands of DNA.

We focused on two widely-used polyamide foldamers, hairpin and linear polyamides (**3** and **6**, Figure 1), that are known to modulate transcription in live human cells.[7a,9] First, we tested whether psoralen-mediated crosslinking was driven by the polyamide. We reasoned that the high affinity of polyamides for specific DNA sequences (K_a $10^9\,\text{M}^{-1}$) would deliver the psoralen to a specific sequence of DNA, whereas the lower affinity of psoralen (K_a 10^3–$10^4\,\text{M}^{-1}$) would not detectably impact polyamide distribution at nanomolar concentrations.[14] To test this hypothesis, in vitro crosslinking experiments were performed with a 22 bp double-stranded DNA (dsDNA) that contains the cognate site of **3**. Hairpin **3** was incubated at a concentration of 40 nM with dsDNA for 1 h at 4 °C and then irradiated at 365 nm. The small molecule–DNA crosslinks were resolved by polyacrylamide gel electrophoresis under conditions that denature duplex DNA (7.5 M urea). The crosslinking to DNA required both the polyamide and psoralen groups, as well as 365 nm UV irradiation (Figure 2 a). As expected, the crosslinking was greatly reduced in 1 and 2 bp mismatches compared to match DNA, consistent with previous reports of polyamides with tethered crosslinkers (Figure S2 a,b).[7b,12c,f] Importantly, these crosslinks could be reversed with hot alkali to enable downstream analysis of the captured DNA (Figure S2).

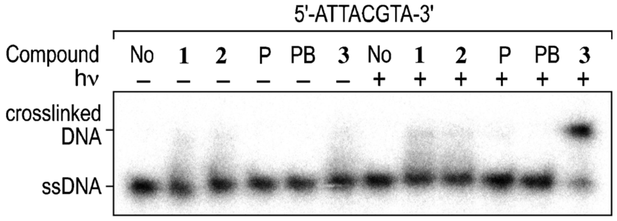

Figure 2. In vitro crosslinking of **3** to DNA. **3** (400 nM) was incubated with ^{32}P-labeled DNA containing the match sequence for **3** and a TpA site for psoralen crosslinking, 5′-ATTTACGTGTA-3′. DNA was resolved by gel electrophoresis under denaturing conditions.

We then tested whether polyamides can maintain their sequence specificity in the biochemically active, chromatinized environment of nuclei. Multiple criteria were used to select six genomic loci that would provide insight into the targeting potential of polyamides (see the Supporting Information). These loci include regions near Pol II-transcribed genes, loci with and without DNase I hypersensitivity (DNase HS), and loci with transcriptionally permissive chromatin landscapes.

COSMIC followed by quantitative polymerase chain reaction (COSMIC-qPCR) was used to determine the extent of polyamide binding to these sites. Briefly, **3** or **6** (40 nM) were incubated with nuclei from human HEK293 cells at 4 °C for 1 h and then crosslinked to DNA with 365 nm UV irradiation. DNA was sheared by sonication, and crosslinked polyamide–DNA complexes were captured with streptavidin-coated magnetic beads. After highly stringent washes under semi-denaturing conditions (4 M urea) to remove noncovalently associated DNA, psoralen crosslinks were reversed, and DNA was purified. Quantitative PCR (qPCR) was used to determine the signal of polyamide binding as a fraction of the affinity-purified (AP) DNA over a reference sample of DNA, called input DNA. It is possible that the PB moiety plays a role in the polyamide occupancy, but we did not detect this in our measurements. We applied COSMIC to examine the occupancy of **3**, the hairpin polyamide designed to target HREs, at both VEGF and ET-2 enhancers, and found that **3** targeted the HRE of ET-2 nearly twice as well as the HRE of VEGF, consistent with the greater magnitude of inhibition for ET-2 (Figure S3 b).[7a]

To understand the differential occupancy of **3** and **6** at sites with similar binding affinities, we next used genomescapes to bioinformatically score genomic loci based on the number of binding sites of varying affinities that are proximal to the targeted site [Figure 3, Eq. (1)]. Many bioinformatic methods annotate DNA-binding sites on the genome with only the highest-affinity consensus sites.[15] Yet genome-wide analyses of sites occupied by DNA-binding proteins in cooperative complexes strongly suggest that clusters of moderate-to-low binding sites should also be considered in modeling genomic occupancy profiles.[3a, 16]

We therefore summed all of the in vitro binding intensities (*Z*-score) within a genomic region of interest [Eq. (1)]

$$\sum_{i=x}^{y-l+1} Z_{\mathrm{PA}}\{\mathrm{seq}(i:i+l-1)\} \qquad (1)$$

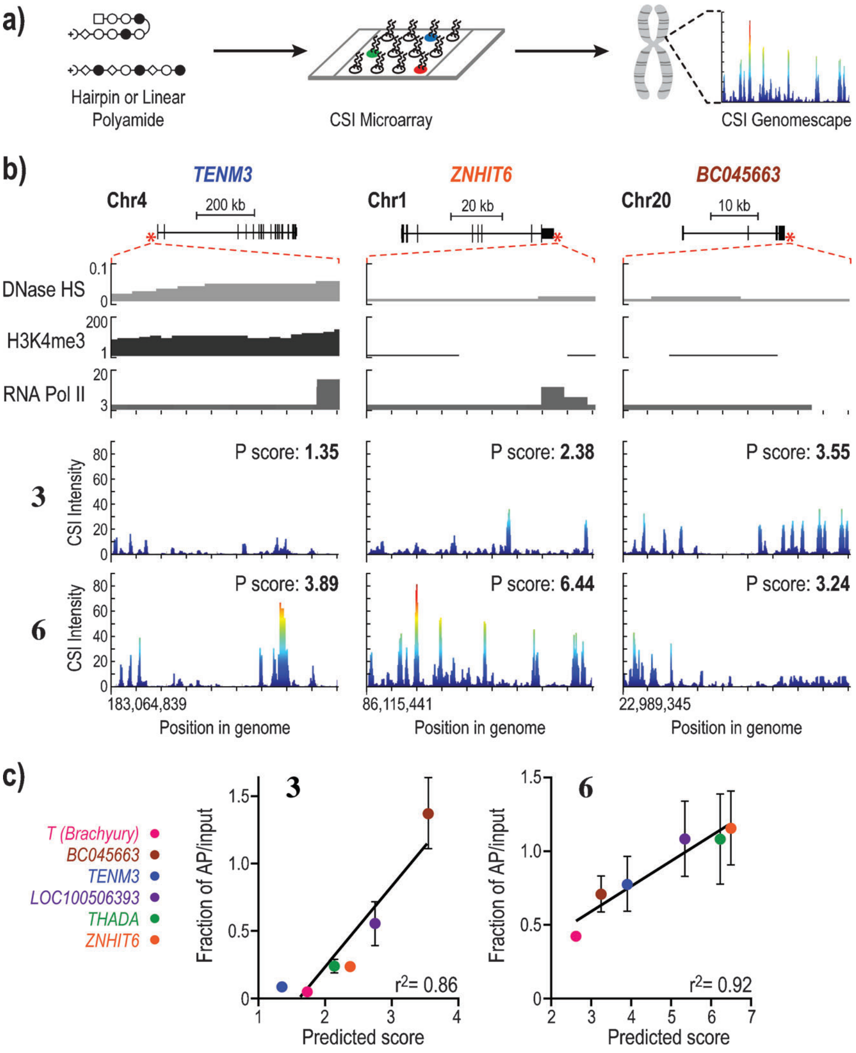

Figure 3. Polyamide binding in a chromatinized environment. a) The process to create CSI genomescapes. Each feature on the DNA microarray displays a unique sequence as a DNA hairpin, with all sequence variants of DNA represented on the array (ca. 1 million sequences). Polyamides are added to the microarray to simultaneously obtain intensity values for every DNA sequence. Genomescapes are generated by assigning an intensity level to every 12 bp sequence of the locus. b) Genomescapes of three of the six loci studied by COSMIC-qPCR (the remaining three loci are shown in Figure S3c). Each locus was bioinformatically scored for predicted binding by **3** and **6** (450 bp window shown). We also analyzed the underlying chromatin structure from ENCODE data. P score = predicted score. c) Scatterplot of predicted score and observed signal from nuclei of HEK293 cells treated with **3** or **6** (40 nM; Fraction of AP/input). AP = affinity-purified. Results are mean ± s.e.m. (standard error of the mean).

where *x* is the start of the locus (seq), *y* is the end of the locus, $Z_{\mathrm{PA}}\{\mathrm{seq}(i{:}i{+}l{-}1)\}$ is the *Z*-score of the given polyamide for the window *i* to $i+l-1$, and *l* is the length of the CSI oligo. Different loci were therefore scored by summing binding sites with CSI-derived binding energies within a window.[3a] We examined different window sizes from 10 to 2000 bp and found that 420 ± 20 bp correlated best with the COSMIC-based occupancy measurements in nuclei (Figure S3 d). Our bioinformatically predicted binding scores, derived from CSI-genomescapes, are directly proportional to the observed polyamide occupancy in nuclei (Figure 3 c).

To further examine the specificity of polyamides, we performed COSMIC analysis at an additional concentration (400 nM). We observed binding profiles that were similar to those obtained at a concentration of 40 nM (Figure S3 e). Thus, the sequence specificity of polyamides observed in vitro is preserved at the genomic level.

We next examined whether our studies with nuclei recapitulated the effects of polyamides in live cells in culture. **6** was selected for further study because linear polyamide architectures display less specificity compared to hairpin polyamides,[3] therefore **6** represents a stringent test of our bioinformatic predictions in live cells. Cellular morphology did not change after treatment with **6** (400 nM), consistent with the low toxicity of polyamides (Figure 4 a). **6** was

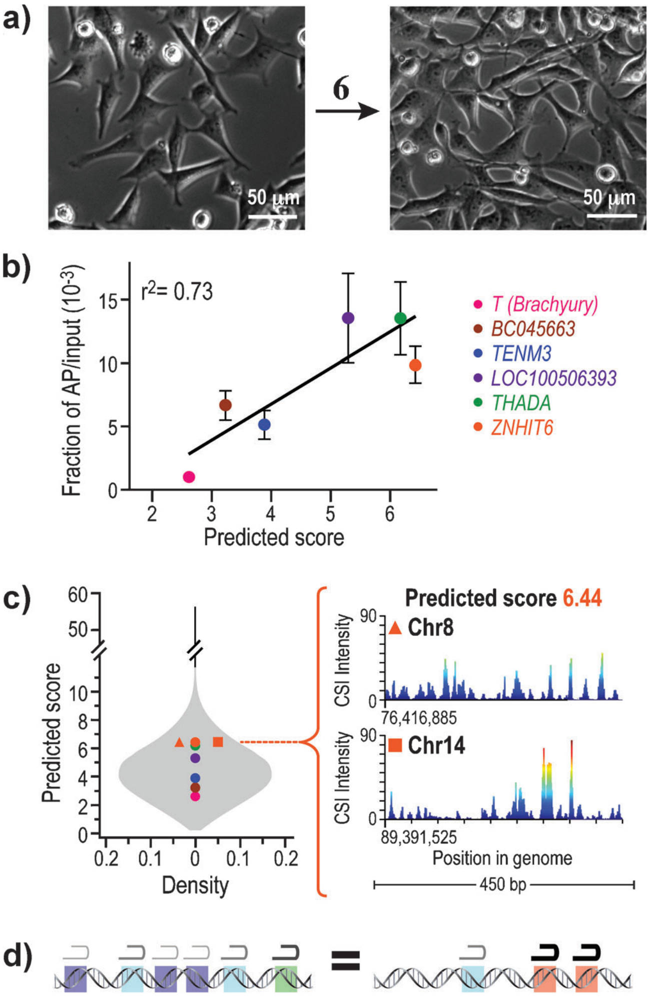

Figure 4. COSMIC-qPCR from live HEK293 cells treated with **6**. a) HEK293 cells before and after treatment with **6** (400 nм). No changes in cellular morphology were observed. b) Comparison of predicted binding signal with empirically determined signal by COSMIC-qPCR (Fraction of AP/Input). Results are mean ± s.e.m. c) Frequency distribution of predicted scores of **6** binding across the entire genome shown as a violin plot. For reference, loci studied in this work are marked as circles on the violin plot. d) Loci with multiple low- and medium-affinity sequences show similar polyamide occupancy to loci with few high-affinity sequences.

incubated for 16 h with live cells and then crosslinked to DNA, and COSMIC-qPCR was performed at the same six loci studied above. The data from live cells treated with **6** are consistent with the results found in nuclei (Figure 4b). Moreover, genomescapes and cumulative scoring over a defined window correlated well with occupancy across diverse loci in live cells. Based on our findings, we scored the entire human genome in 420 bp windows (Figure 4c and S5). Because of the limitations of histograms,[17] we displayed the distribution of scores as a violin plot. The violin plot provides a density trace to reveal patterns in the dataset. This plot shows that different genomic loci with similar predicted binding scores exhibit the diverse clustering of multiple sites of varying affinities.

The general strategy of targeting unique or specific high-affinity binding sites has been successfully used to perturb binding of a variety of DNA-binding proteins in cells.[7,9] However, computational analysis of our COSMIC data at several different genomic loci reveals that polyamide occupancy in cells is strongly correlated with multiple clustered binding sites of varying affinities. In particular, it was surprising that the level of crosslinking at such "multi-site" loci with low- and medium-affinity sequences exceeded that observed at high-affinity "single-site" loci. An approximately 400 bp genomic locus with multiple sites of varying affinity best predict levels of polyamide occupancy in cells. This observation challenges the current paradigm that guides the design of genome-targeting molecules. Such a multi-site targeting strategy would be especially effective in targeting transcription factor binding sites within enhancers, promoter elements where the transcriptional machinery assembles, and the newly discovered regulatory regions called super-enhancers that span more than a thousand base pairs.[18]

Some of the most successful therapeutic agents are molecules that bind to DNA and interfere with an array of genomic transactions.[19] We propose that COSMIC can be used to design new therapeutic strategies with the combinatorial use of DNA-targeting ligands.

Received: May 21, 2014
Published online: July 25, 2014

Keywords: chemoinformatics · COSMIC · genome targeting · molecular recognition · polyamides

[1] a) D. E. Wemmer, P. B. Dervan, *Curr. Opin. Struct. Biol.* **1997**, *7*, 355–361; b) J. W. Lown, K. Krowicki, U. G. Bhat, A. Skorobogaty, B. Ward, J. C. Dabrowiak, *Biochemistry* **1986**, *25*, 7408–7416.

[2] a) P. B. Dervan, *Bioorg. Med. Chem.* **2001**, *9*, 2215–2235; b) J. G. Pelton, D. E. Wemmer, *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 5723–5727.

[3] a) C. D. Carlson, C. L. Warren, K. E. Hauschild, M. S. Ozers, N. Qadir, D. Bhimsaria, Y. Lee, F. Cerrina, A. Z. Ansari, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4544–4549; b) J. W. Puckett, K. A. Muzikar, J. Tietjen, C. L. Warren, A. Z. Ansari, P. B. Dervan, *J. Am. Chem. Soc.* **2007**, *129*, 12310–12319.

[4] M. Minoshima, T. Bando, S. Sasaki, J. Fujimoto, H. Sugiyama, *Nucleic Acids Res.* **2008**, *36*, 2889–2894.

[5] J. M. Gottesfeld, C. Melander, R. K. Suto, H. Raviol, K. Luger, P. B. Dervan, *J. Mol. Biol.* **2001**, *309*, 615–629.

[6] a) T. G. Edwards, K. J. Koeller, U. Slomczynska, K. Fok, M. Helmus, J. K. Bashkin, C. Fisher, *Antiviral Res.* **2011**, *91*, 177–186; b) T. G. Edwards, T. J. Vidmar, K. Koeller, J. K. Bashkin, C. Fisher, *PLoS ONE* **2013**, *8*, e75406; c) G. He, E. Vasilieva, G. D. Harris, Jr., K. J. Koeller, J. K. Bashkin, C. M. Dupureur, *Biochimie* **2014**, *102*, 83–91.

[7] a) B. Z. Olenyuk, G. J. Zhang, J. M. Klco, N. G. Nickols, W. G. Kaelin, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16768–16773; b) L. A. Dickinson, R. Burnett, C. Melander, B. S. Edelson, P. S. Arora, P. B. Dervan, J. M. Gottesfeld, *Chem. Biol.* **2004**, *11*, 1583–1594; c) X. Xiao, P. Yu, H.-S. Lim, D. Sikder, T. Kodadek, *Angew. Chem.* **2007**, *119*, 2923–2926; *Angew. Chem. Int. Ed.* **2007**, *46*, 2865–2868; d) S. Janssen, O. Cuvier, M. Müller, U. K. Laemmli, *Mol. Cell* **2000**, *6*, 1013–1024; e) G. N. Pandian, Y. Nakano, S. Sato, H. Morinaga, T. Bando, H. Nagase, H.

Sugiyama, *Sci. Rep.* **2012**, *2*, 544; f) F. Yang, N. G. Nickols, B. C. Li, G. K. Marinov, J. W. Said, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 1863 – 1868; g) A. K. Mapp, A. Z. Ansari, M. Ptashne, P. B. Dervan, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3930 – 3935.

[8] a) C. L. Warren, N. C. S. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips Jr. , A. Z. Ansari, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 867 – 872; b) J. R. Tietjen, L. J. Donato, D. Bhimisaria, A. Z. Ansari, *Methods Enzymol.*, *Vol. 497* (Ed.: V. Chris), Academic Press, **2011**, pp. 3 – 30; c) K. E. Hauschild, J. S. Stover, D. L. Boger, A. Z. Ansari, *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3779 – 3782.

[9] R. Burnett, C. Melander, J. W. Puckett, L. S. Son, R. D. Wells, P. B. Dervan, J. M. Gottesfeld, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 11497 – 11502.

[10] T. K. Archer, M. G. Cordingley, R. G. Wolford, G. L. Hager, *Mol. Cell. Biol.* **1991**, *11*, 688 – 698.

[11] C. Jespersen, E. Soragni, C. James Chou, P. S. Arora, P. B. Dervan, J. M. Gottesfeld, *Bioorg. Med. Chem. Lett.* **2012**, *22*, 4068 – 4071.

[12] a) L. Anders, M. G. Guenther, J. Qi, Z. P. Fan, J. J. Marineau, P. B. Rahl, J. Loven, A. A. Sigova, W. B. Smith, T. I. Lee, J. E. Bradner, R. A. Young, *Nat. Biotechnol.* **2014**, *32*, 92 – 96; b) M. Lee, M. C. Roldan, M. K. Haskell, S. R. McAdam, J. A. Hartley, *J. Med. Chem.* **1994**, *37*, 1208 – 1213; c) N. R. Wurtz, P. B. Dervan, *Chem. Biol.* **2000**, *7*, 153 – 161; d) L. Guan, M. D. Disney, *Angew. Chem.* **2013**, *125*, 10194 – 10197; *Angew. Chem. Int. Ed.* **2013**, *52*, 10010 – 10013; e) J. D. White, M. F. Osborn, A. D. Moghaddam, L. E. Guzman, M. M. Haley, V. J. DeRose, *J. Am. Chem. Soc.* **2013**, *135*, 11680 – 11683; f) T. Bando, H. Sugiyama, *Acc. Chem. Res.* **2006**, *39*, 935 – 944.

[13] E. E. Baird, P. B. Dervan, *J. Am. Chem. Soc.* **1996**, *118*, 6141 – 6146.

[14] J. E. Hyde, J. E. Hearst, *Biochemistry* **1978**, *17*, 1251 – 1257.

[15] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, J. Taipale, *Cell* **2013**, *152*, 327 – 339.

[16] D. Panne, T. Maniatis, S. C. Harrison, *Cell* **2007**, *129*, 1111 – 1123.

[17] J. S. Simonoff, F. Udina, *Comput. Stat. Data An.* **1997**, *23*, 335 – 353.

[18] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, R. A. Young, *Cell* **2013**, *153*, 307 – 319.

[19] a) D. Wang, S. J. Lippard, *Nat. Rev. Drug Discovery* **2005**, *4*, 307 – 320; b) L. H. Hurley, *Nat. Rev. Cancer* **2002**, *2*, 188 – 200.