First publl. in: ChemBioChem 11 (2010), 18, pp. 2499–2501, DOI: 10.1002/cbic.201000569

High-Throughput DNA Sequencing Beyond the Four-Letter Code: Epigenetic Modifications Revealed by Single-Molecule Bypass Kinetics

Daniel Summerer*[a]

Throughout all domains of life, nucleic acid-based epigenetic modification of DNA is established as a regulatory mechanism that can affect different processes ranging from the replicative mismatch repair to the transcriptional activity of genes. Modifications can be preserved during cell divisions and play important roles in development—with consequences of aberrant regulation of modification such as the promotion of diseases including cancer.[1–3]

Several types of modification that are associated with varying biological roles, depending on the organism and cell type are known. Methylation of cytosine nucleobases at position 5 (mC, Figure 1 A) is the best-studied epigenetic mark in eukaryotes. In vertebrates, it is often found in clusters of CpG dinucleotides (CpG islands) at or close to transcriptional start positions, usually with repressive function.[4,5] Additionally, 5-hydroxymethylcytosine (hmC) has recently been identified as a further modification in certain neurons and embryonic stem cells, with a function that is not yet understood.[6,7] Finally, in bacteria and archaea *N*-6-methyladenine (mA), among others, has been found to participate in restriction-modification systems and mismatch-repair strand discrimination.[8]

Because of mC's central role in human disease, the development of methods for its analysis has been pursued with great enthusiasm. Many approaches for direct methylation detection have been described that use various forms of methylation-sensitive enzyme reactions or enrichment/detection with antibodies,[9] chromatography,[6,10] mass spectrometry[7] or nanopore amperometry.[11] However, indirect methylation detection based on bisulfite conversion has become the most successful strategy.[12] Here, genomic DNA is treated with sodium bisulfite at high temperatures, which results in the deamination of C but not mC into uracil; this represents the conversion of a chemical modification into a genetic mark that can subsequently be resolved by various methods of DNA single-base analysis,[13–15] or sequence analyses including Sanger- or Pyrosequencing.[16,17] Recently, the benefits of next-generation sequencing technologies have also started to have an impact on bisulfite sequencing and enabled genome-wide mC methylome (or "bisulfitome") mapping in a number of organisms.[18–20] Nevertheless, the bisulfite approach also has serious drawbacks, including DNA fragmentation caused by the harsh conversion conditions and a reduction in the sequence complexity of DNA to three nucleotides (A, G, T(U)), which complicates primer design or the mapping of sequence information. Moreover, the approach cannot be used to differentiate between A and mA or between mC and hmC. Methods for the direct analysis of methylation have the potential to circumvent this bottleneck, and a simultaneous analysis of the primary sequence and the state of different

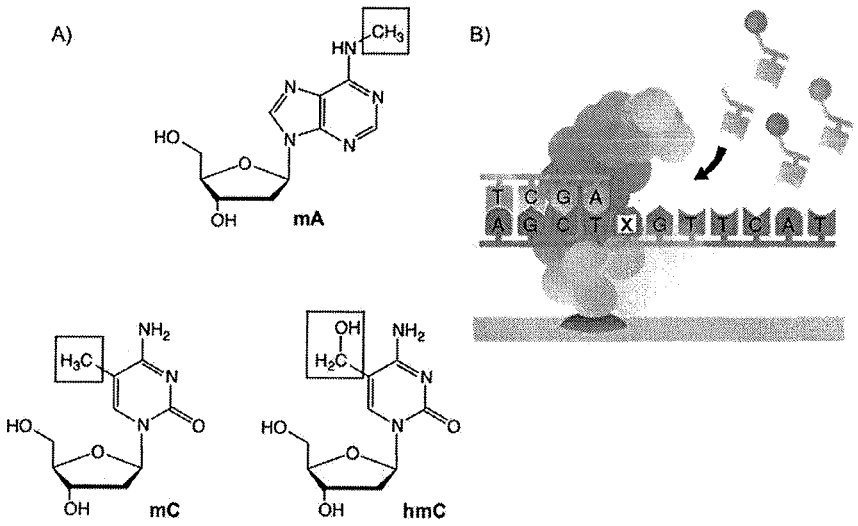

Figure 1. A) Structures of deoxynucleotides with nucleic acid-based epigenetic modifications. B) Principle of SMRT DNA sequencing, which allows for simultaneous analysis of primary sequence and epigenetic modification. X = mC, hmC or mA. Adapted from ref. [21] with permission. Copyright Macmillan Publishers Ltd. (2010).

[a] *Dr. D. Summerer*
Febit Biomed GmbH
Im Neuenheimer Feld 519, 69120 Heidelberg (Germany)
Fax: (+49) 6221-6510-329
E-mail: daniel.summerer@febit.de

types of epigenetic modification patterns in high throughput would be especially valuable.

Recently, Flusberg et al. have demonstrated an exciting approach that allows for the direct, bisulfite-free detection of methylation and overcomes many of the previous bottlenecks by using single-molecule, real-time (SMRT) DNA sequencing.[21] SMRT sequencing makes use of so-called zero-mode waveguides (ZMWs), zeptolitre-sized nanostructures that allow for fluorescence detection with minimized background. Single DNA polymerases can be immobilized in ZMWs and can perform processive primer extension on single-sample DNA molecules.[22] By using four deoxynucleoside triphosphates (dNTPs) with specific fluorescent labels on the γ-phosphate, template-specific incorporation of a canonical dNTP results in the temporal localization of the label in the ZMW and subsequent release after (labelled) pyrophosphate dissociation. Fluorescence monitoring allows the identification of consecutively incorporated dNTPs by colour and the identity of the individual templating nucleotides (Figure 1 B). Besides the order of dNTP incorporation, two additional kinetic parameters can be extracted from obtained fluorescence traces: the pulse width, which is a function of all kinetic steps of dNTP incorporation until pyrophosphate release and the interpulse duration (IPD), which is dependent on dNTP binding and DNA polymerase translocation.

It was known that IPDs are sensitive to DNA sequence and secondary structure.[22] However, Flusberg et al. have demonstrated that they are also affected by the presence of epigenetic modifications in the template strand. They performed comparative sequencing of synthetic templates containing only unmodified nucleotides or two separated mC, hmC or mA bases in typical sequence contexts found in nature (i.e., mC and hmC located in CpG sites and mA located in GATC sites). In most cases, bypass synthesis over a modified nucleotide was associated with an increased IPD (Figure 2) compared to unmodified nucleotides. IPD ratios between methylated and control samples built for all individual positions of >300 accumulated measurements revealed kinetic differences at several positions at and close to the modification site, thus suggesting multiple critical interactions with the DNA polymerase throughout the complex. Comparisons between the two identical methylated nucleotides in each template revealed a sequence dependence of the signatures. This complicates the de novo identification of different methylation types; however, some characteristics of the IPD ratio signatures could be identified that differed between the types. By taking into account multiple parameters (i.e., mean IPDs and pulse widths at nine different positions around the methylated site) direct differentiation of C, mC and hmC was indeed possible in this specific context.

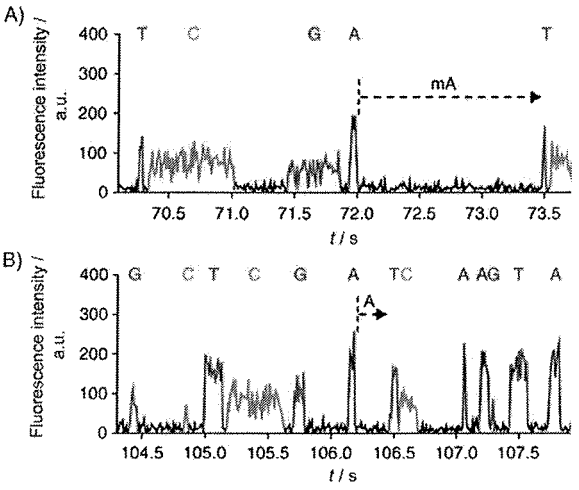

Figure 2. Fluorescence traces of SMRT sequencing showing the impact of mA on IPD (dashed arrow). A) Trace of sequenced template containing a single mA. B) Control trace of unmodified sequenced template. Adapted from ref. [21] with permission. Copyright Macmillan Publishers Ltd. (2010).

Importantly, these analyses were performed on populations of identical molecules. However, in real samples, only a fraction of a certain locus might be methylated, and accumulation of reads would result in mixed signatures that could complicate analysis. The authors reason that by using circularized templates, a single sample molecule could be repetitively analysed by several laps of strand-displacement synthesis. The total fraction of methylated bases at a given site in a sample could then be obtained by counting methylation states over multiple sample molecules. This, however, requires sufficient call rates with a relatively low number of repetitive reads, since the information has to be obtained from a single primer extension-event and the processivity of the DNA polymerase is limited. To test feasibility, an analysis of the dependence of call rates for mA on the number of accumulated reads was performed on an approximately 200-bp circular template. Taking into account the IPD of only the site of incorporation opposite mA, >85% of mA could be called with a false-positive rate of 5% when using five accumulated reads. This could be improved by accumulating more laps of sequencing by increased processivity of the DNA polymerase or the use of smaller circles. Furthermore, analysis of additional sites that are affected in their IPD by the presence of mA could yield improvement.

The approach was then extended to a more complex sample, a 3.7 kb stretch of a *C. elegans* fosmid isolated from a dam$^+$ *E. coli* strain containing 13 GATC sequences that are expected to contain mA. Comparative sequencing with a whole-genome-amplified control sample (in which mA is removed) revealed similar IPD characteristics to those found in the synthetic templates. IPD ratios (mA/A) varied little between different GATC sites, with no significant dependence on the GC content of the sequence context.

These data suggest that faithful mA detection for various sequence contexts could be in reach with relatively little further development effort. However, as the authors point out, differentiation of mC and hmC might require more substantial improvements in biochemical aspects and the use of algorithmic approaches to include multiple template positions, owing to their more complex kinetic signatures. Given the importance of these modifications in eukaryotes, a successful implementation would be highly valuable. The use of the method for de novo calling of different methylation states is not possible yet, and, with the observed complex kinetic signatures and dependencies on sequence contexts and secondary structures,

such a development appears highly challenging. However, the use of collected reference data with known states from previous sequencings for comparison could be straightforward for resequencing projects.

Overall, the approach represents an exciting first study to remove the bottlenecks of bisulfite sequencing and allow the simultaneous detection of different epigenetic modifications in a high-throughput setup. The potential usability of completely untreated genomic DNA samples together with long read lengths makes the approach a highly promising setup for an unbiased analysis of whole-methylome profiling in complex organisms. Moreover, the possibility of exploiting single-molecule bypass kinetics for the analysis of chemical DNA modifications might be transferable to any modification that sufficiently impacts DNA polymerase kinetics—ranging from DNA damage to further epigenetic modifications.

Acknowledgements

The author acknowledges Stephen W. Turner for providing graphic material for adaptation.

Keywords: DNA methylation · epigenetics · high-throughput sequencing · nucleosides · polymerases

[1] A. Razin, R. Shemer, *Hum. Mol. Genet.* **1995**, *4*, 1751–1755.
[2] P. A. Jones, S. B. Baylin, *Nat. Rev. Genet.* **2002**, *3*, 415–428.
[3] K. D. Robertson, *Nat. Rev. Genet.* **2005**, *6*, 597–610.
[4] S. Saxonov, P. Berg, D. L. Brutlag, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 1412–1417.
[5] K. R. Pomraning, K. M. Smith, M. Freitag, *Methods* **2009**, *47*, 142–150.
[6] S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929–930.
[7] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, A. Rao, *Science* **2009**, *324*, 930–935.
[8] M. G. Marinus, J. Casadesus, *FEMS Microbiol. Rev.* **2009**, *33*, 488–503.
[9] C. A. Eads, R. V. Lord, S. K. Kurumboor, K. Wickramasinghe, M. L. Skinner, T. I. Long, J. H. Peters, T. R. DeMeester, K. D. Danenberg, P. V. Danenberg, P. W. Laird, K. A. Skinner, *Cancer Res.* **2000**, *60*, 5021–5026.
[10] M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel, T. Carell, *Angew. Chem.* **2010**, *122*, 5503–5505; *Angew. Chem. Int. Ed.* **2010**, *49*, 5375–5377.
[11] J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid, H. Bayley, *Nat. Nanotechnol.* **2009**, *4*, 265–270.
[12] P. W. Laird, *Nat. Rev. Genet.* **2010**, *11*, 191–203.
[13] M. Bibikova, Z. Lin, L. Zhou, E. Chudin, E. Wickham Garcia, B. Wu, D. Doucet, N. J. Thomas, Y. Wang, E. Vollmer, T. Goldmann, C. Seifart, W. Jiang, D. L. Barker, M. S. Chee, J. Floros, J.-B. Fan, *Genome Res.* **2006**, *16*, 383–393.
[14] C. A. Eads, K. D. Danenberg, K. Kawakami, L. B. Saltz, C. Blake, D. Shibata, P. V. Danenberg, P. W. Laird, *Nucleic Acids Res.* **2000**, *28*, E32.
[15] M. Strerath, A. Marx, *Angew. Chem.* **2005**, *117*, 8052–8060; *Angew. Chem. Int. Ed.* **2005**, *44*, 7842–7849.
[16] 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.
[17] J. M. Dupont, J. Tost, H. Jammes, I. G. Gut, *Anal. Biochem.* **2004**, *333*, 119–127.
[18] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, S. E. Jacobsen, *Nature* **2008**, *452*, 215–219.
[19] C. A. Bormann Chung, V. L. Boyd, K. J. McKernan, Y. Fu, C. Monighetti, H. E. Peckham, M. Barker, *PLoS One* **2010**, *5*, e9320.
[20] N. Li, M. Ye, Y. Li, Z. Yan, L. M. Butcher, J. Sun, X. Han, Q. Chen, X. Zhang, J. Wang, *Methods* **2010**, *52*, 203–212.
[21] 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.
[22] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, *Science* **2009**, *323*, 133–138.