Genomic Imaging

DOI: 10.1002/anie.201107714

Enzymatically Incorporated Genomic Tags for Optical Mapping of DNA-Binding Proteins**

Soohong Kim, Anna Gottfried, Ron R. Lin, Thomas Dertinger, Andrew S. Kim, Sangyoon Chung, Ryan A. Colyer, Elmar Weinhold, Shimon Weiss,* and Yuval Ebenstein**

Affordable DNA sequencing is revolutionizing genetic research and is enabling multiple novel biomedical applications. Among the inherent properties of today's high-throughput sequencing technologies is the fact that it compiles long-range sequences from the assembly of numerous short-read data.[1] This leads to two fundamental limitations: loss of long-range contextual information on the single-genome level and difficulties coping with repetitive or variable genomic regions. Optical mapping and its variants[2–10] rely on the visualization of individual, long (50 kb–1000 kb) DNA molecules and extraction of genomic information by fluorescent labeling of the DNA. These techniques lack the resolution of sequencing but offer genomic context and therefore are attractive both in combination with sequencing to aid in sequence assembly[11–13] and for investigation of genomic structural variations on the individual chromosome level.[14,15] Such variations include deletions, duplications, copy-number variants (CNVs), insertions, inversions, and translocations, all of which have a major impact on the phenotypic variations within a population (or somatic mutations, important in cancer progression). In addition, the available information content of the genome extends beyond the sequence, and the long-range data offered by optical mapping may provide crucial information regarding the distribution of DNA-binding proteins such as transcription factors and histones along the genome.

In previous work we have shown that optical mapping can accurately reconstruct the full promoter map of the T7 bacteriophage genome.[16,17] We used fluorescent quantum dots (QDs) to tag individual RNA polymerase (RNAP) enzymes occupying their genomic binding sites and showed that all 17 promoters could be reliably detected. However, mapping relied on the fact that we were investigating intact viral genomes where the ends of the double-stranded DNA molecules served as reference points for the assignment of the RNAP to its genomic binding sites. To apply similar optical mapping technology to more complex systems, ranging from bacterial artificial chromosomes (BACs) to bacterial and mammalian genomes, it is essential to create genomic reference tags (refTags) that are independent of the DNA extremities.[18] We utilize sequence-specific methyltransferase-induced labeling of DNA (SMILing DNA) to create a unique, sequence-specific fluorescence pattern.[19] Neely et al. used a similar approach to generate a simple but long-range representation of the lambda phage DNA sequence.[10] The conceptual difference in our case is that rather than labeling the DNA with the objective of sequencing it, we use sequence-specific labeling as a reference overlay to aid in the mapping of information not necessarily encoded in the DNA sequence. In this case we use this pattern to genetically identify the region of DNA under observation and to determine the genomic location of RNAP bound to the DNA and labeled with different color probes.

We demonstrate the utility of this approach by localizing T7-RNAP on one of its target genes, but the technique can similarly be applied to mapping other DNA-binding proteins such as transcription factors or histones (see the Supporting Information), as well as for the long-range mapping of DNA methylation patterns through direct labeling of methylated DNA (with antibodies against 5-methyl cytosine for example).

We show that by choosing a DNA methyltransferase (MTase) that modifies a rare sequence in the genome we can engineer a unique "barcode" that identifies the orientation and exact identity of the observed DNA, and at the same time serves to calibrate the DNA stretching factor (by converting the observed distance between two refTags into the known distance in base pairs). Finally, the genomic location of any observable lying in between two such refTags can accurately be mapped. In fact, we observe a fivefold higher precision in the assignment of the binding site tested in this work when using the genomic markers relative to assigning the promoter location based on the DNA ends. A schematic representation of the experimental concept is depicted in Figure 1.

We used the T7 bacteriophage genome as a model system to demonstrate the utility of this approach to the assignment of T7-RNAP binding sites. Our objective was to generate

[*] S. Kim, R. R. Lin, Dr. T. Dertinger, A. S. Kim, S. Chung, Dr. R. A. Colyer, Prof. S. Weiss
Department of Chemistry and Biochemistry
University of California, Los Angeles (USA)
E-mail: sweiss@chem.ucla.edu

Dr. Y. Ebenstein
Department of Chemical Physics, Tel Aviv University
Ramat Aviv 69978, Tel Aviv (Israel)
E-mail: uv@post.tau.ac.il

A. Gottfried, Prof. Dr. E. Weinhold
Institute of Organic Chemistry, RWTH Aachen University
Landoltweg 1, 52056 Aachen (Germany)
E-mail: elmar.weinhold@oc.rwth-aachen.de

[**] This work was supported by the UCLA-DOE Institute for Genomics and Proteomics (Grant DE-FC02-02ER63421) and NIH Grant R01-GM069709. A.G. and E.W. thank Prof. Dr. Michalis Kokkinidis for providing M.BseCI and the German Science Foundation (WE 1453/4-1 and 4-2) for financial support. Y.E. thanks the Legacy Heritage fund program of the Israel Science Foundation for financial support.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201107714.

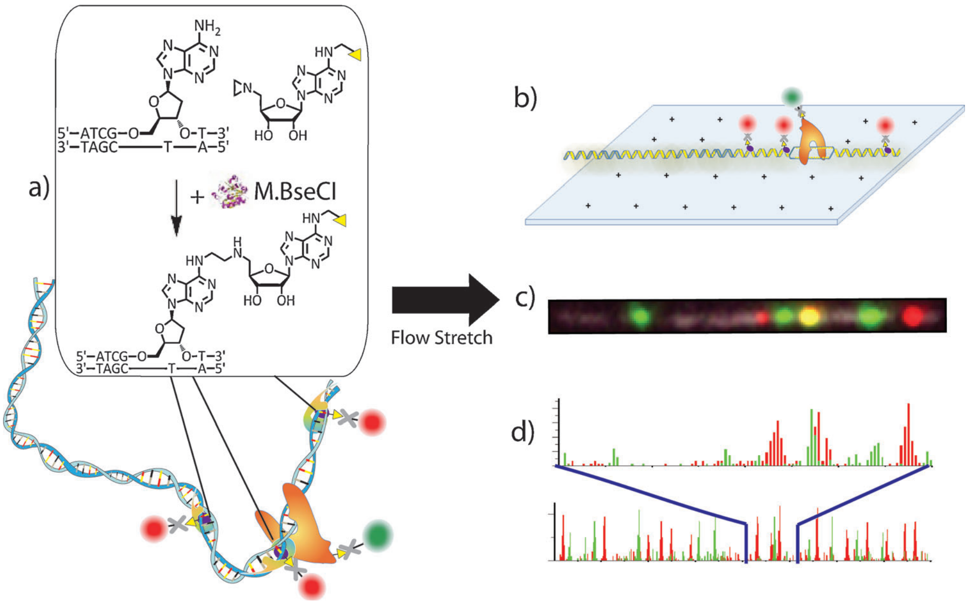

Figure 1. a) Schematic representation of QD-labeled RNAP bound to sequence-specific-labeled T7 bacteriophage DNA. The DNA methyltransferase M.BseCl recognizes specific rare sequences in the T7 genome (5′-ATCGAT-3′) and conjugates an aziridine cofactor 6BAz with an attached biotin (yellow triangle). This biotin group can be used for site-specific labeling with QDs. b) RNAP-bound T7 bacteriophage genome stretched over a polylysine surface. c) Image of flow-stretched, YOYO-1 stained T7 bacteriophage DNA (white) with QD-labeled RNAP (green) and M.BseCl refTags labeled with spectrally distinct QDs (red). Overlapping red and green signals are shown in yellow. d) Conceptual representation of a genome-wide map of promoter and DNA methyltransferase sites (bottom). Histogram of M.BseCl positions represented in red and RNAP positions represented in green (top).

a unique fluorescent pattern along the genome that would be sparse enough to not produce overlapping fluorescence signals and would also not interfere with RNAP binding. The labeling is performed by DNA MTases that naturally use the cofactor *S*-adenosyl-L-methionine (AdoMet or SAM) to introduce a methyl group onto an adenine or cytosine residue within their recognition sequence. However, upon feeding the enzyme with a modified cofactor, various chemical moieties can be covalently attached to the DNA backbone and the variety of DNA MTases from different species allows modification of a large array of possible recognition sequences including rare sequences in the human genome. We have chosen the bacterial DNA MTase M.BseCI because its recognition sequence (5′-ATCGAT-3′) appears only three times in the T7 genome and creates a well-defined, asymmetric pattern. We have used the aziridine-based AdoMet analogue 6BAz[20,21] to introduce an exposed biotin at three defined positions along the 40 kb genome (see Figure S2 in the Supporting Information). These reference points are additionally labeled with streptavidin-coated quantum dots (SAV-QDs) emitting at $\lambda = 625$ nm. The labeled DNA was extended by capillary flow between two coverslips as reported previously[16] and imaged on a fluorescence microscope. Results from this experiment are presented in Figure 2.

The efficiency of the enzymatic reaction is generally extremely high[19] and a lower limit of 95% can be estimated by a restriction enzyme assay and subsequent agarose gel electrophoresis (see Figure S3 in the Supporting Information). However the binding efficiency of the SAV-QDs was lower, with only around 40% of the 1700 tested DNA molecules displaying two or more QDs, and was most likely caused by the presence of free streptavidin in the QD solution (saturating biotin sites on the DNA). Nevertheless, this degree of binding was sufficient to generate the binding histograms shown in Figure 2 and Figure S6 in the Supporting Information. The three reference positions can be easily distinguished from one another and therefore may be independently assigned to their genomic loci. Encouraged by these results we then examined how refTags can aid in the mapping of DNA binding proteins such as RNAP. Recombinant T7-RNAP, containing a biotin tag on the N terminus,[16] was reacted with SAV-QDs emitting at $\lambda = 700$ nm at a 1:2 ratio, thus resulting in a majority of the RNAP labeled with a single QD. The enzyme was incubated with T7 genomic DNA which was prelabeled with the reference QDs. To stabilize the RNAP–DNA binding we initiated transcription by feeding RNAP with a limited set of nucleotides (G/A/UTPs), thus causing transcription to stall as a result of the lack of CTP. A detailed description of all experimental procedures is available in the Supporting Information. After staining the DNA with the intercalating dye YOYO-1, the sample was extended by flow as before and mounted on a fluorescence microscope for imaging. Fluorescence from the sample was recorded in three channels corresponding to the three emission colors from the RNAP–DNA complexes (DNA: $\lambda = 500$ nm, refTags: $\lambda = 625$ nm, RNAP: $\lambda = 700$ nm).

The sample was excited by a single blue excitation band using a Xenon lamp and an excitation bandpass filter (470/40). The three emission wavelengths were recorded on an EMCCD (Andor DU897) by rotating a filter wheel containing the appropriate emission filters (530/50, 620/40, 665 long pass). To correct for image shifts caused by the rotation of the filter wheel between the two QD channels we used a sample of QDs, emitting at $\lambda = 655$ nm, dispersed upon a clean coverslip. These QDs could be imaged through both channel filters and were used to create a transformation matrix that defined the shift between the channels at every pixel in the image. The calibration procedure is described in detail in the Supporting Information. On average, the positional error between channels was reduced from over 150 nm to below 30 nm after applying the transformation. Figure 3 shows color overlay images of selected genomes carrying both RNAPs (green) and refTags (red). RNAP signals can be traced to various binding sites along the genome. However, we found

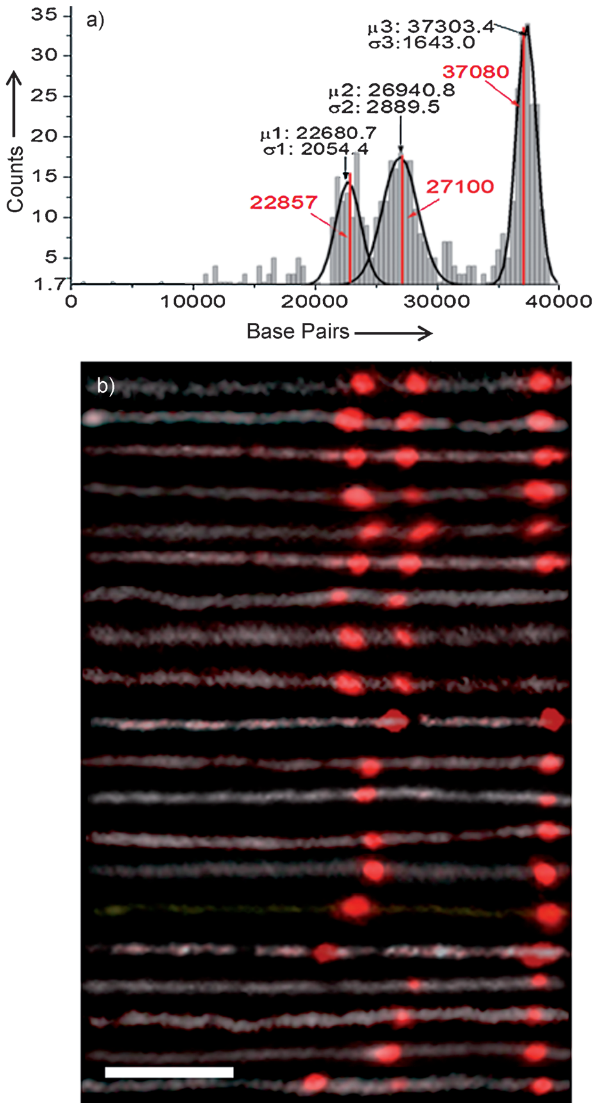

Figure 2. a) Histogram obtained from localizing refTag: QDs on stretched T7 bacteriophage genomes. Red vertical lines represent expected position of recognition sequences. b) Selected images of extended YOYO-1 stained T7 genomes (white) labeled with QDs on M.BseCI DNA MTase refTag sites (red). Scale bar: 3 μm.

that mapping precision improved only for RNAP detected between two refTags. As a result, we focused our analysis on RNAP bound to promoter Φ13, which lies between refTags one and three.

We localized the refTag and RNAP QDs by fitting a two-dimensional (2D) Gaussian to their fluorescence spots. The center of the Gaussian was extracted with nanometer accuracy and a table containing the *x* and *y* coordinates of the detected QDs was generated. Next, the transformation matrix was applied to bring RNAP and refTag locations to the same coordinate system. Using the known location of the refTags we calculate a stretching factor for each individual DNA molecule. This factor is used to convert the distance measured between the RNAP and the refTags into a genomically relevant value in base pairs. Figure 4 shows position histograms for RNAP detected on the promoter Φ13. To the left are histograms generated from the same data without using the refTags but instead relying on the DNA ends for mapping. To the right are histograms generated from data extracted using refTags. The width of the distribution, represented by the standard deviation of the fitted Gaussians, is a measure of the mapping precision while the difference between the peak of the distribution and the actual genomic promoter is the error. The width of the distribution is significantly reduced when using refTags and precision improved fivefold (from ca. 1.5 kb to ca. 310 bp when taking into account all the data and from ca. 1.2 kb to ca. 270 bp when mapping only strands longer than 80 pixels, corresponding to 70% extension). This precision compares favorably to the precision of ChIP-chip (chromatin imunoprecipitation followed by DNA-chip analysis) data and suggests that the use of refTags can overcome limitations on precision obtained when using the DNA strand end points. Moreover, it facilitates the mapping of arbitrary, long genomic fragments by uniquely identifying their genomic origin. By itself, SMILing DNA presents an attractive method for mapping structural variations in genomic DNA by visualizing the physical pattern created along long genomic stretches. Checking typical recognition sequences against the human genome shows that unique patterns composed of discrete

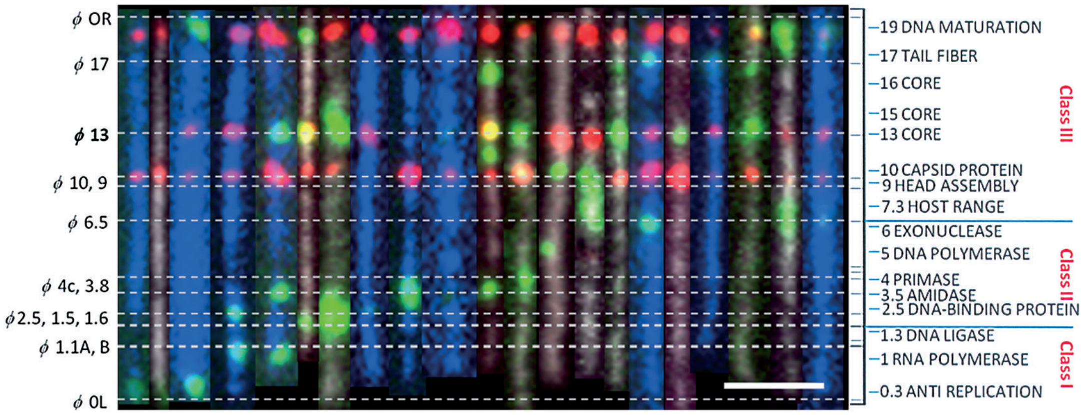

Figure 3. Sequence-specific methyltransferase-induced labeling of DNA (SMILing DNA) creates a distinct barcode which reports the preferential binding sites of RNAP on T7 bacteriophage DNA. Selected images of QD-labeled RNAP (green) bound to promoters on stretched YOYO-1 stained T7 bacteriophage DNA (white or blue) having refTag sites labeled with spectrally distinct QDs (red). Overlapping red and green signals are shown in yellow. Scale bar: 3 μm.

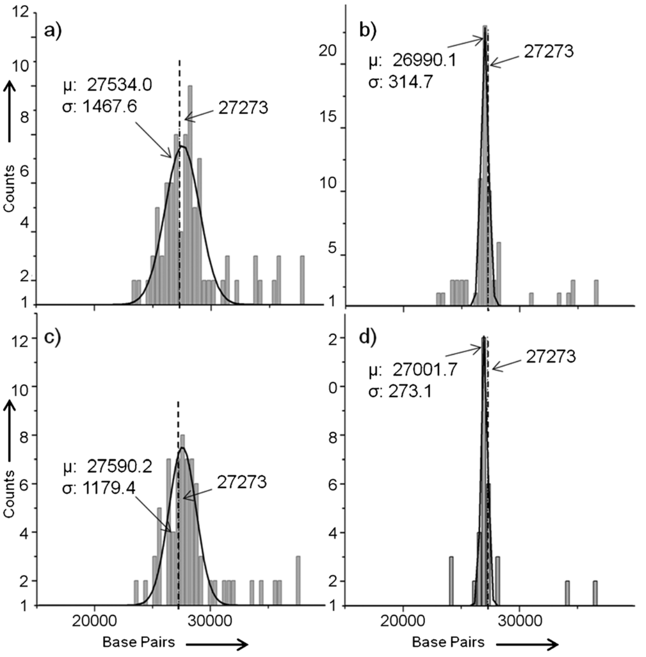

Figure 4. Histograms for localized RNAP on T7 bacteriophage using distance measurment to the DNA ends (left) versus localization by refTags (right). a), b) Histograms of all analyzed strands with Gaussian fit. c), d) Histograms based on selection of long strands (>80 pixels) with Gaussian fit. Dotted lines represent actual position of Φ13 promoter. Histograms using refTags yield a fivefold increase in accuracy as evidenced by sharp reductions in the width of promoter localization distributions.

labels can be generated. When combined with the visualization of an additional layer of information such as protein binding sites, optical mapping provides the missing contextual information lacking in bulk assays such as DNA arrays or sequencing where binding events originate in different genomes and single-cell genomic context is lost. By investigating such information over long distance scales on the single-molecule level, new information regarding the cooperative nature of certain binding proteins, as well as variations across individual chromosomes may be examined. Such analysis may give rise to subpopulations that are otherwise obscured by ensemble averaging. This analysis may be of particular relevance for diagnostic purposes where early detection of rare events may facilitate targeted and early medical intervention.

Received: November 2, 2011
Revised: December 19, 2011
Published online: February 16, 2012

Keywords: DNA · imaging agents · optical mapping · quantum dots · single-molecule studies

[1] J. Shendure, H. Ji, *Nat. Biotechnol.* **2008**, *26*, 1135 – 1145.

[2] R. K. Neely, J. Deen, J. Hofkens, *Biopolymers* **2011**, *95*, 298 – 311.

[3] X. Meng, K. Benson, K. Chada, E. J. Huff, D. C. Schwartz, *Nat. Genet.* **1995**, *9*, 432 – 438.

[4] X. Michalet, *Science* **1997**, *277*, 1518 – 1523.

[5] E. Y. Chan, N. M. Goncalves, R. A. Haeusler, A. J. Hatch, J. W. Larson, A. M. Maletta, G. R. Yantz, E. D. Carstea, M. Fuchs, G. G. Wong, et al., *Genome Res.* **2004**, *14*, 1137 – 1146.

[6] M. Xiao, A. Phong, C. Ha, T.-F. Chan, D. Cai, L. Leung, E. Wan, A. L. Kistler, J. L. DeRisi, P. R. Selvin, et al., *Nucleic Acids Res.* **2007**, *35*, e16.

[7] S. K. Das, M. D. Austin, M. C. Akana, P. Deshpande, H. Cao, M. Xiao, *Nucleic Acids Res.* **2010**, *38*, e177.

[8] W. Reisner, N. B. Larsen, A. Silahtaroglu, A. Kristensen, N. Tommerup, J. O. Tegenfeldt, H. Flyvbjerg, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13294 – 13299.

[9] H. Zohar, C. L. Hetherington, C. Bustamante, S. J. Muller, *Nano Lett.* **2010**, *10*, 4697 – 4701.

[10] R. K. Neely, P. Dedecker, J. Hotta, G. Urbanavičiūtė, S. Klimašauskas, J. Hofkens, *Chem. Sci.* **2010**, *1*, 453 – 460.

[11] G. Narzisi, B. Mishra, *PloS one* **2011**, *6*, e19175.

[12] G. Narzisi, B. Mishra, *Bioinformatics* **2011**, *27*, 153 – 160.

[13] C. Aston, B. Mishra, D. C. Schwartz, *Trends Biotechnol.* **1999**, *17*, 297 – 302.

[14] S. Caburet, C. Conti, C. Schurra, R. Lebofsky, S. J. Edelstein, A. Bensimon, *Genome Res.* **2005**, *15*, 1079 – 1085.

[15] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, et al., *Nature* **2008**, *453*, 56 – 64.

[16] Y. Ebenstein, N. Gassman, S. Kim, J. Antelman, Y. Kim, S. Ho, R. Samuel, X. Michalet, S. Weiss, *Nano Lett.* **2009**, *9*, 1598 – 1603.

[17] Y. Ebenstein, N. Gassman, S. Kim, S. Weiss, *J. Mol. Recognit.* **2009**, *22*, 397 – 402.

[18] H. Zohar, S. J. Muller, *Nanoscale* **2011**, *3*, 3027 – 3039.

[19] G. Pljevaljcić, F. Schmidt, E. Weinhold, *ChemBioChem* **2004**, *5*, 265 – 269.

[20] S. Wilkinson, M. Diechtierow, R. A. Estabrook, F. Schmidt, M. Hüben, E. Weinhold, N. O. Reich, *Bioconjugate Chem.* **2008**, *19*, 470 – 475.

[21] G. Braun, M. Diechtierow, S. Wilkinson, F. Schmidt, M. Hüben, E. Weinhold, N. O. Reich, *Bioconjugate Chem.* **2008**, *19*, 476 – 479.