

Supporting information

Lighting up individual transcription factors with quantum dots

Yuval Ebenstein^{1,2,*}, Natalie Gassman¹, Soohong Kim¹, Younggyu Kim¹, Sam Ho¹,
Robin Samuel¹, Xavier Michalet¹ and Shimon Weiss^{1,2,3,*}

¹Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA

²DOE Institute for Genomics and Proteomics, UCLA, Los Angeles, CA

³Department of Physiology, Geffen Medical School, UCLA, Los Angeles, CA

*e-mail: uv@chem.ucla.edu; sweiss@chem.ucla.edu

Brief overview of existing methods for the study of protein-DNA interactions in the context of transcription.

The coordinated regulation of gene expression in response to changes in cellular conditions is the basis for cell function. The construction of models that describe the differential expression of genes is an important step towards understanding of how living organisms behave. RNA polymerase (RNAP) is a key player in gene expression, transcribing DNA sequences (genes) into messenger RNA with the help of its associated factors. This transcriptional activity is under the control of transcription factors (TFs), and regulatory proteins such as inhibitors and enhancers,

which may bind to DNA sequences and/or interact with the transcription complex. The specific DNA sequence to which RNAP binds in order to initiate the transcription of RNA is known as a promoter. Knowledge of promoter locations is the first step towards the construction of a full transcriptional regulatory network. It gives information on the activity of specific gene expressions and defines genomic regions for the identification of regulatory elements associated with various TFs, thus facilitating the interpretation of their binding sites.

Considerable effort has been devoted to identify transcriptional networks and to map genomic regions that participate in the control of gene expression. Initial efforts have been based on *in vitro* DNA-binding and reporter assays such as electrophoretic mobility shift assays, DNA footprinting, and luciferase reporter systems. These tools allowed the identification of regulatory elements in the vicinity of selected genes and have elucidated binding motifs for candidate TFs. However, these approaches are limited by the fact that they require prior knowledge of candidate genes and permit analysis of relatively small genomic regions. *In vitro* approaches to the identification of TF binding sites have proven to suffer from both false positive and false negative results when compared to the live genomic system. This may partially be explained by the fact that protein-protein interactions may recruit TFs to sequences in the genome that differ from the optimal *in vitro* binding sites. The conformation differences between the artificial recombinant binding element and the native binding domains in their genomic context may also play an important role in dictating which binding motifs are active *in vivo*. Even in the primitive case of viral transcription, the binding of T7 bacteriophage RNAP to a recombinant promoter sequence has been altered

when the plasmid containing the binding site has been transformed from supercoiled to linear form¹.

Chromatin immunoprecipitation (ChIP) allows the *in vivo* identification of TF binding sites in the context of the entire genome and therefore avoids many of the problems associated with traditional *in vitro* approaches (for reviews see:²⁻⁴) ChIP involves chemical crosslinking of DNA–protein interactions in living cells. Commonly, formaldehyde is used to 'fix' TFs to their cognate binding sites in the genome. Formaldehyde is membrane permeable and forms a covalent bridge between proximal amino or imino groups such as a lysine in contact with a cytosine. The crosslinking provides a 'snapshot' of the cell's transcription state under well defined conditions. The crosslinked DNA is then extracted from cells and fragmented. Antibodies against specific TFs are used to immunoprecipitate TFs together with their bound DNA fragments. In this way, only the DNA associated with the protein of interest is 'fished out' of the rich protein-DNA soup, resulting in an enriched sample. Next, DNA–protein crosslinking is reversed, and enriched DNA fragments are purified for downstream analysis.

Standard ChIP assays use Southern blotting, polymerase chain reaction (PCR) or quantitative real-time PCR (qPCR) as a read-out; however, these approaches still require a specific region of interest in the genome to facilitate analysis. The introduction of micro-array technology and its combination with ChIP in the form of genomic microarrays (ChIP-chip)^{5, 6}, and more recently the direct sequencing of ChIP products using 454 or Solexa G1 sequencing platforms (ChIPseq)⁷, allows large-scale, genome-wide identification of TF-binding sites. In ChIP-chip, the

immunoprecipitated DNA is labeled with a fluorescent dye and used to probe a genomic array. Although ChIP-chip represents a high throughput approach, it requires microarrays of unbiased full genomic sequences. Whole-genome tiling arrays are available for all non-repetitive regions of the human genome, but they are expensive, and for the human genome, consist of a set of roughly 38 arrays⁸. Lower coverage platforms, although useful in many cases are not unbiased, as they use arbitrarily defined promoter regions without experimental verification and exclude other regions that might be involved in transcriptional regulation, such as those in introns. ChIPseq involves the direct sequencing of ChIP-enriched DNA. Sequencing data is matched to its respective loci in the known genomic sequence, and binding sites are determined by accumulation of sequence reads above the background level. ChIPseq is not limited by array coverage and therefore offers unbiased analysis of the entire genome. In addition, it is simple and offers higher resolution compared with array based methods.

All of the above ChIP based techniques address one TF specie at a time. Indirect assessment of cooperative binding of several different TFs is achieved by running parallel experiments on the same sample and comparing the location data for different TFs. Nevertheless, ambiguity remains regarding the exact nature of the cooperativity due to the fact that data comes from different DNA fragments. To address this issue, a second round of ChIP may be performed on an enriched sample using an antibody against a second TF of interest. This double ChIP procedure selects only DNA fragments interacting with both TFs. This approach has been termed double-ChIP, repChIP or SeqChIP^{9, 10} and although powerful, it is limited to close range interactions comparable to the DNA fragment size used in the ChIP experiment

(usually in the order of 500 bp). Regulation of gene expression however, is not limited to close range interactions between TFs and especially in eukaryotes may be influenced by extremely distant cofactors brought together by random fluctuations or directed chromatin reorganization. An example of how chromosomal interactions can influence gene expression is the folding ability of a chromosomal region that can bring an enhancer and associated transcription factors within close proximity of a gene¹¹. The Capturing Chromosome Conformation (3C) assay¹² and its offspring: ChIP-loop¹³, 3C-on chip(4C)¹⁴ and 3C-carbon copy(5C)¹⁵ are approaches to detect the frequency of interaction between any two genomic loci (for review, see ¹⁶). In these methods, long range DNA-protein interactions are captured by formaldehyde treatment. The crosslinked chromatin is subject to digestion by restriction enzymes to free the unwound chromatin from the bulk of crosslinked material. This is followed by ligation of crosslinked fragments, and then ligation frequencies are measured. In the ChIP-loop assay, immunoprecipitation enriches the sample for fragments bound by a specific protein, and restriction fragments are ligated to each other on the ChIP beads. In ChIP-loop and 3C, ligation frequencies are measured by quantitative PCR, using a unique primer set for each ligation junction analyzed. In 5C, ligation events are amplified by ligation-mediated amplification (LMA) with T7 and T3 primers, and then analyzed by large-scale sequencing or microarray. In 4C, ligation junctions are first trimmed by a frequently cutting secondary restriction enzyme, and then subjected to ligation to form circles followed by inverse PCR to amplify captured fragments. The 4C PCR product is analyzed by large-scale sequencing or microarray analysis. In general, 3C technology is particularly suited to study the conformation of genomic regions that range roughly from five to several hundred kilobases (kb) in size.

Methods

Protein Expression and Purification

T7 RNA polymerase was amplified from pDL19¹⁷ and cloned into the vector pAN-4 (Avidity) containing an N-terminal BiotinAviTag peptide. The resulting plasmid, pNG301, containing the T7 RNAP with the N-terminal BiotinAviTag and an N-terminal hexa-histidine tag was transformed into AVB101 cells (Avidity). These cells overexpress a biotin ligase, which recognizes the peptide tag and introduces a single biotin molecule at the N-terminal. An overnight culture (10 mL) from a single colony was subcultured into 1 liter of LB media with 100 µg/mL of ampicillin and 25 µg/mL chloramphenicol at 37 °C and grown to an OD₆₀₀ of 0.4-0.6. Cells were induced with 1 mM IPTG and 50 µM biotin. The cells were grown an additional 3 hours at 37°C, and then harvested by centrifugation for 15 minutes at 5000 x g, 4 °C. The cell pellet was resuspended in 35 mL of nickel binding buffer (NBB: 50 mM sodium phosphate, 10 mM Tris-HCl, 0.5 M NaCl, 5mM imidazole, 4 mM 2-mercaptoethanol (2-ME), 5 % glycerol, pH 8.0), disrupted by sonication, and the lysate was centrifuged for 30 min at 15,000 x rpm, 4 °C; all subsequent steps were maintained at 4 °C.

1 ml of Ni-NTA agarose beads (Qiagen) was equilibrated with 10 ml of NBB in a 20 ml disposable column (Econo-Pac, Bio-Rad), and the cleared lysate was passed through the column by gravity flow, followed by a 10 ml NBB (plus 10 mM imidazole) wash. Proteins were eluted with 3 column volume (CV) of NBB containing 20 mM, 40 mM and 150 mM imidazole, respectively, and fractions were collected at 1 CV each. The fractions were analyzed by SDS-PAGE, and peak fractions containing RNAP were pooled, and diluted with TGED buffer (20 mM Tris-

HCl, 0.1 mM EDTA, 1 mM dithiothreitol (DTT), 5 % glycerol) to a final NaCl concentration of 0.1 M.

The resulting sample was loaded onto a Heparin-Sepharose column (GE Healthcare) pre-equilibrated with TGED with 0.1 M NaCl, using Aktä *purifier* (GE Healthcare). Protein was eluted with a gradient from TGED + 0.1 M NaCl to 1 M NaCl over 120 min at 0.25 mL/min. Fractions containing pure RNAP were pooled and dialyzed against storage buffer (10 mM Tris, pH 7.9, 50 % glycerol, 0.1 mM EDTA, 0.1 mM DTT, 0.05 M NaCl), then stored at -20 °C. The typical yield is about 5-10 mg of purified RNAP from a 1 liter culture.

Surface Preparation

Functionalized glass coverslips were prepared as follows: glass coverslips (22 mm x 22 mm, number 1.5, Fisher) were washed with 1 % (w/v) fresh Alconox (Fisher) solution and sonicated for 15-30 min with heating. Coverslips were then washed with copious amounts of deionized water and baked at 500 °C for 2-3 hrs to remove any organic contamination. A 10 µL solution of 5-6 µg/mL polylysine (Sigma) in water was sandwiched between two cooled coverslips, and allowed to dry overnight. The sandwiched coverslips may be used for up to a week with reproducible results or stored at -20 °C for longer periods. Before imaging, sandwiched coverslips are separated and an un-treated coverslip ("out of the box") is placed on top of the polylysine surface. This top coverslip is slightly hydrophobic and serves to trap free QDs from the sample. This was verified by separately imaging the upper and lower coverslip after sample deposition.

Sample Preparation

Protein-DNA samples were prepared by reacting in a total of 10 μL , 3 nM of T7 genome (Boca Scientific) with 30-167 nM of biotinylated T7-RNAP for 20 min at 37 $^{\circ}\text{C}$ in T7 binding buffer (30 mM Hepes pH 7.0, 25 mM K-Glutamate, and 15 mM Mg-acetate). To stabilize the DNA-RNAP complex, transcription was allowed to initiate by feeding the reaction with 1 μL of 1 mM GTP, UTP, and CTP and incubating at 37 $^{\circ}\text{C}$ for an additional 5 min. When a large excess of RNAP was used (>60 nM), non-specific binding was reduced by challenging with heparin before the addition of NTPs. 0.6 μL of 1 mg/mL heparin-sepharose (GE Healthsciences) was added to a 6 μL reaction and incubated for 30 s at 37 $^{\circ}\text{C}$. The sample was then centrifuged at max speed in a benchtop centrifuge, and 5 μL was removed from supernatant. Typical reactions contained 1 μg of DNA to which 0.3 μL of a 1 mM stock solution of YOYO-1 in DMSO was added. Streptavidin conjugated QDs (Qdot 655, Invitrogen) were diluted to 100 nM and filtered by a 100 kDa pore size membrane (YM-100, Centricon) to eliminate free streptavidin. The flow-through was discarded and the QDs were resuspended in HEPES buffer to a final concentration of 20 nM. 5 μL of the QD solution was added and the sample was further incubated in the dark for 30 min. For stretching and imaging, 0.5 μL of the prepared sample is diluted into 50 μL of 100 mM HEPES pH 7.0 containing 0.1 % of n-dodecyl- β -D-maltoside (DDM). The small number of bound QDs per DNA molecule observed under the microscope was in part due to the sub-stoichiometric amount of QDs used in the experiment to reduce non-specific binding of QDs on the surface.

AFM Imaging

The reaction mixture was diluted 1:20 in 50 mM HEPES and 2.5 mM MgCl₂, and then a 10 μL droplet was deposited on freshly cleaved mica and incubated for 20 s. The sample is washed with copious amounts of deionized water and spin coated dry. Samples were scanned in tapping mode using a MFP-3D AFM (Asylum Research). For height assessment of RNAP and QD, the images were flattened and cross-sections of various features were performed. The lower than expected features of QDs are typical for AFM imaging in tapping mode¹⁸.

Fluorescence Microscopy and image processing

The sample was imaged on an inverted microscope (IX71, Olympus) with a 60X, oil immersion objective (Plan-Apo, 1.45NA, Olympus). A magnifying lens resulted in a field of view of ~50 μm². Fluorescence was excited with a Xenon arc lamp through a 470 ± 15 nm bandpass filter and images were collected with an electron multiplying CCD (EMCCD) (DU997, Andor). The resolution of 97 nm/pixel allowed high-resolution localization of fluorescence spots. Images of the YOYO-1 stained DNA were acquired with a 535 ± 25 nm band-pass filter. QD fluorescence was acquired with a bandpass filter centered at 655 nm (or as required for other QD colors). An automatic protocol was written to acquire color planes sequentially and overlay the separate planes into a multicolor image (IQ, Andor). The contrast of each color channel was modified independently to sharpen DNA edges for subsequent analysis and to enhance the image for presentation. For precise distance measurements, emission from all QDs was acquired simultaneously through a long pass filter, eliminating image shifts due to filter changes. The distances between fluorescence

spots are accumulated in a histogram, and used to construct a binding site map of the DNA template.

Image analysis

DNA length and distances between the DNA end and QDs were measured manually using a custom code written in LabView (National Instruments) or with Image J (<http://rsb.info.nih.gov/ij>). High-resolution localization of QDs by 2D Gaussian fitting was done with the NIH supported freeware: Video spot tracker (http://www.cs.unc.edu/Research/nano/cisimm/download/spottracker/video_spot_tracker.html). The localization coordinates were used to calculate the distance between QDs.

Simulations

We simulated non-uniformly extended molecules using the following (unphysical) simple algorithm. We first divide the full genome into N equal segments (typically $N = 1,000$ is sufficient and the results do not depend on the choice of N as long as $N > 100$). For each segment, we randomly assign an extension factor e_i ($i = 1, \dots, N$) taken from a Gaussian distribution centered on 1 and of standard deviation σ/\sqrt{N} . This choice ensures that the total length is normally distributed with standard deviation σ . We then compute the location of each binding site k on the extended molecule by summing the length of each segment preceding the segment on which it is located (say, segment i), and finally add e_i times the local coordinate of this binding site in segment i . We then normalize all locations by dividing those coordinates by the length of the extended molecule. These normalized coordinates are then used to obtain the

histograms shown on Fig. 5. Note that since we use normalized distances, it is equivalent to use an extension factor centered on 1 with standard deviation σ/\sqrt{N} or to use an extension factor centered on x with standard deviation $x\sigma/\sqrt{N}$. The simulation allows the user to attribute different probabilities to the different binding sites, define a global detection efficiency and add a fraction of non-specific binding.

Assessing the contribution of sequence specific genomic flags

We studied the resolution improvement expected from the incorporation of sequence specific reference flags by considering one or more promoter bound QDs as genomic flags and measuring all other locations in respect to these genomic loci. Looking at 3 DNA molecules labeled with 4 to 5 QDs, which could be assigned with high confidence to known promoters, we compared the inferred location of each promoter by three methods as described below and presented in Supporting Table 1. The average localization error of 1,336 bp was reduced to 800 bp upon introducing a single genomic flag and further improved to 354 bp when two genomic flags were considered. These results clearly demonstrate the significant improvement expected from the incorporation of sequence specific flags.

The three genomes were analyzed as if they contained 2, 1 or no flag QDs. First, the molecules were analyzed considering the distances of each QD to the molecule end (corresponding to the analysis performed in the absence of any flag). This analysis yielded an average localization error of 1,336 bp. Next, we used one bound QD as a flag and identified its position as that of its assigned promoter. The remaining QDs were localized by Gaussian fitting and their precise distances to the flag measured with nm resolution. These distances were then renormalized using the

overall length of the DNA molecule, to obtain absolute distance in bp. In this case, the average promoter localization error was reduced to 800 bp. Finally, we used two bound QDs as flags and identified their positions as that of the two corresponding promoters. Using the ratio between their observed and theoretical distance as the DNA extension factor, we located the other promoters with respect to the flags. In this case, the average localization error was reduced further to 353 bp.

1. Smeekeens, S.P. & Romano, L.J. Promoter and nonspecific DNA binding by the T7 RNA polymerase. *Nucleic Acids Research* **14**, 2811-2827 (1986).
2. Kirmizis, A. & Farnham, P.J. Genomic Approaches That Aid in the Identification of Transcription Factor Target Genes. *Experimental Biology and Medicine* **229**, 705-721 (2004).
3. Massie, E.C. & Mills, G.I. ChIPping away at gene regulation *EMBO reports* **9**, 337-343 (2008).
4. Wells, J. & Farnham, P.J. Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation. *Methods* **26**, 48-56 (2002).
5. Ren, B. et al. Genome-Wide Location and Function of DNA Binding Proteins. *Science* **290**, 2306-2309 (2000).
6. Buck, M.J. & Lieb, J.D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349-360 (2004).

7. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth* **4**, 651-657 (2007).
8. Carroll, J.S. et al. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**, 1289-1297 (2006).
9. Geisberg, J.V. & Struhl, K. Quantitative sequential chromatin immunoprecipitation, a method for analyzing co-occupancy of proteins at genomic regions in vivo. *Nucleic Acids Research* **32**, e151 (2004).
10. Scully, K.M. et al. Allosteric Effects of Pit-1 DNA Sites on Long-Term Repression in Cell Type Specification. *Science* **290**, 1127-1131 (2000).
11. Murrell, A., Heeson, S. & Reik, W. Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nat Genet* **36**, 889-893 (2004).
12. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science* **295**, 1306-1311 (2002).
13. Horike, S.-i., Cai, S., Miyano, M., Cheng, J.-F. & Kohwi-Shigematsu, T. Loss of silent-chromatin looping and impaired imprinting of *DLX5* in Rett syndrome. *Nat Genet* **37**, 31-40 (2005).
14. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-1354 (2006).
15. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat. Protocols* **2**, 988-1002 (2007).
16. Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nature Methods* **4**, 895-901 (2007).

17. He, B. et al. Rapid Mutagenesis and Purification of Phage RNA Polymerases. *Protein Expression and Purification* **9**, 142-151 (1997).
18. Ebenstein, Y., Nahum, E. & Banin, U. Tapping Mode Atomic Force Microscopy for Nanoparticle Sizing: Tip-Sample Interaction Effects. *Nano Letters* **2**, 945-950 (2002).