Published online 1 February 2021

Nucleic Acids Research, 2021, Vol. 49, No. 9 ***e49***
doi: 10.1093/nar/gkab038

Integrating quantitative proteomics with accurate genome profiling of transcription factors by greenCUT&RUN

Sheikh Nizamuddin[1,2,†], Stefanie Koidl[1,2,†], Tanja Bhuiyan[1,2], Tamara V. Werner[2,3], Martin L. Biniossek[4], Alexandre M.J.J. Bonvin[5], Silke Lassmann[2,3] and H.Th.Marc Timmers[1,2,*]

[1]Department of Urology, Medical Center-University of Freiburg, 79016 Freiburg, Germany, [2]German Cancer Consortium (DKTK) partner site Freiburg, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany, [3]Institute for Surgical Pathology, Medical Center-University of Freiburg, 79016 Freiburg, Germany, [4]Institute for Molecular Medicine and Cell Research, Medical Center-University of Freiburg, 79016 Freiburg, Germany and [5]Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht 3584CH, the Netherlands

Received October 08, 2020; Revised January 08, 2021; Editorial Decision January 11, 2021; Accepted January 16, 2021

ABSTRACT

Genome-wide localization of chromatin and transcription regulators can be detected by a variety of techniques. Here, we describe a novel method 'greenCUT&RUN' for genome-wide profiling of transcription regulators, which has a very high sensitivity, resolution, accuracy and reproducibility, whilst assuring specificity. Our strategy begins with tagging of the protein of interest with GFP and utilizes a GFP-specific nanobody fused to MNase to profile genome-wide binding events. By using a GFP-nanobody the greenCUT&RUN approach eliminates antibody dependency and variability. Robust genomic profiles were obtained with greenCUT&RUN, which are accurate and unbiased towards open chromatin. By integrating greenCUT&RUN with nanobody-based affinity purification mass spectrometry, 'piggy-back' DNA binding events can be identified on a genomic scale. The unique design of greenCUT&RUN grants target protein flexibility and yields high resolution footprints. In addition, greenCUT&RUN allows rapid profiling of mutants of chromatin and transcription proteins. In conclusion, greenCUT&RUN is a widely applicable and versatile genome-mapping technique.

INTRODUCTION

Gene expression programs are regulated by the combinatorial action of many chromatin and transcription regulatory factors through the combined binding of transcription factors or cofactors to chromatin and by histone modifications serving as binding platforms. Misregulation of transcription programs is associated with a broad range of human pathologies, for example, cancer and cardiovascular diseases (1–3). Several techniques have been developed for the genome-wide profiling of regulatory factors including ChIPseq (chromatin immunoprecipitation), ChECseq (chromatin endogenous cleavage), CUT&TAG (cleavage under targets and tagmentation) and CUT&RUN (cleavage under targets and release using nuclease) (4–7). Of these, CUT&RUN is a recently developed experimental approach for the high resolution mapping of DNA binding sites for transcription factors and chromatin proteins, and for the profiling of histone modifications across eukaryotic genomes *in situ* (6). The CUT&RUN profiling approach utilizes antibody targeting of micrococcal nuclease (MNase) fused to an immunoglobulin-binding protein (protein A or protein A/G) using unfixed permeabilized cells. Compared to traditional genome-mapping approaches like ChIPseq (chromatin immunoprecipitation followed by high-throughput DNA sequencing), CUT&RUN is independent from formaldehyde crosslinking and is characterized by low backgrounds, high spatial resolution, high reproducibility and requirement of low cell numbers (6,8). Several other techniques, for example, CUT&Tag/iACT-seq also perform well with low cell numbers (7,9). The increased signal-to-noise ratio means that

*To whom correspondence should be addressed. Tel: +49 761 270 63120; Email: m.timmers@dkfz-heidelberg.de
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

CUT&RUN profiling requires only ∼10% of the read numbers in comparison with a typical ChIPseq experiment (6). Nevertheless, both ChIPseq and CUT&RUN still depend on high specificity and high affinity antibodies, which are not available for all proteins from all species. In addition, interaction of the target protein with DNA, other proteins and post-translational modifications may occlude the epitope recognized by the antibody. Both issues are circumvented by the tagging of proteins with epitopes for which high affinity reagents are available.

Using cell lines expressing transcription factors tagged by green fluorescence protein (GFP), we explored the use of a GFP single-domain antibody (nanobody) of high affinity and high specificity fused to the catalytic domain of MNase. This approach not only circumvents antibody issues, but it also reduces CUT&RUN handling time and technical variation as steps involving binding of antibody and protein A-MNase are combined. GFP-tagged cell lines were subjected to our CUT&RUN-based approach for GFP proteins, which we name greenCUT&RUN. With greenCUT&RUN, we have developed a versatile genome profiling tool for gene-specific and basal transcription factors, which is independent of antibody availability or quality, whilst still insuring high specificity of measurements. Compared to ChIPseq and standard CUT&RUN, the greenCUT&RUN approach displays a remarkable sensitivity, resolution, accuracy and reproducibility. GreenCUT&RUN documents the experimental advantages of directly fusing MNase to single domain antibodies against epitope tags like GFP and this approach can be extended to other protein ligands. We show that greenCUT&RUN can be combined directly with quantitative mass spectrometry to achieve an integrated platform for the study of proteins regulating transcription programs and chromatin function in mammalian cells.

MATERIALS AND METHODS

Plasmid construction

The ORFs for the human NFYA, FOS, JUN and TBP proteins were obtained by PCR using the appropriate cDNA constructs followed by BP-mediated GATEWAY recombination into pDONR221 according to instructions by the manufacturer (ThermoFisher, USA). The ENTRY clones were verified by DNA sequencing and they correspond to Uniprot sequences: #P23511 for NFYA, #P01100 for FOS, #P05412 for JUN and #P62380 for TBP. The cDNAs were transferred into the pCDNA5-FRT-TO-N-GFP destination clones (pCDNA5-FRT-TO_N-GFP_β-globin for NFYA, FOS and JUN, and pCDNA5-FRT-TO_N-GFP for TBP) by LR-mediated GATEWAY recombination. We found that insertion of β-globin-intron II sequences between the GFP moiety and the cDNA leads to an increase in fusion protein expression of about three-fold. Again, the obtained constructs were verified by DNA sequencing.

The ORFs of the GFP nanobody and the catalytic domain of micrococcal nuclease (MNase) are based on studies by Kubala *et al.* (10) and Zentner *et al.* (11). ORFs were amplified by PCR with appropriate primers containing a linker region (Asp-Asp-Asp-Lys-Glu-Phe) connecting the nanobody and MNase coding regions. The PCR products were purified after agarose gel electrophoresis and fused via overlapping PCR. This nanobody-MNase fragment was cloned into the NcoI and BamHI sites of the pGEX2T-derived vector, pRP265NB, for bacterial expression. Cloned regions of this construct, pRPN265NB-enh-MNase, were verified by DNA sequencing.

Protein purification and activity assay

The GST-nanobody-MNase fusion protein was expressed in BL21(DE3) bacteria by induction with 1 mM IPTG for 3.5 h at 37°C. Bacterial cells where collected via centrifugation at 3500 *g* for 10 min at 4°C. The cell pellet was resuspended in lysis buffer (50 mM Tris–HCl pH 8.0, 20% sucrose, 300 mM KCl, 2 mM EDTA, 0.1% Triton X-100, 1 mM DTT and protease inhibitors (Complete, Roche) followed by three freeze-thaw cycles at -80°C and three cycles in a French press. The lysate was cleared by centrifugation at 83,000 *g* for 30 min at 4°C. The fusion protein was captured by glutathione-sepharose affinity chromatography (GE Healthcare) using an ÄKTA prime purification system (GE Healthcare) in lysis buffer. The column was washed with wash buffer I (50 mM Tris–HCl pH 8, 20% sucrose, 300 mM KCl, 2 mM EDTA, 1 mM DTT and protease inhibitors) followed with wash buffer II (50 mM $K.PO_4$ pH 7.0, 100 mM KCl, 1 mM EDTA, 1 mM DTT). Bound proteins were eluted with elution buffer (wash buffer II with 50 mM reduced glutathione). Peak fractions were combined and the GST fusion protein was cleaved by thrombin (Sigma) at 3 units/mg of total protein for 3 h at 37°C. PMSF was added for 15 min at 37°C to a final concentration of 0.5 mM to inactivate thrombin. Thrombin was removed by batch binding to benzamidine-sepharose (GE Healthcare) for 2.5 h at 4°C. Sepharose beads were removed by passage over a 0.45 μm filter. The cleaved proteins were concentrated by ultracentrifugation using Amicon10K filters and passed over a Superdex-75 HiLoad 16/600 gel-filtration column (GE Healthcare) in buffer A150 (20 mM $K.HPO_4$ pH7.0, 150 mM KCl, 1 mM DTT). Peak fractions still contained free GST and these fractions were further purified via cation exchange chromatography using HiTrap SP column (GE Healthcare) with a linear gradient from 80 to 1000 mM KCl in buffer B (20 mM $K.HPO_4$ pH 7.0, 0.5 mM EDTA, 1 mM DTT, 0.1 mM PMSF). The nanobody-MNase protein displayed a broad elution pattern and the peak fractions were concentrated by ultracentrifugation using Amicon 10K filters. Protein purity was determined by Coomassie staining of 15% polyacrylamide/SDS gels and protein concentrations were determined via Bradford measurement (BioRad Protein Assay) using BSA as a standard. Glycerol was added to 50% for storage at −80°C until use.

MNase activity assay

HeLa genomic DNA (1.25 μg) was incubated in reaction buffer (50 mM Tris–HCl pH 7.8, 5 mM $CaCl_2$, 100 μg/ml BSA) with the indicated amounts of nanobody-MNase and compared to commercial MNase (New England Biolabs). A control with the highest amount of enzyme tested but lacking $CaCl_2$ was performed to control for contaminating nuclease activities. Reactions were incubated 15 min at 37°C and terminated by addition of 20 mM EGTA and 10

mM EDTA. DNA products analyzed on agarose gel electrophoresis in TBE and stained with 0.5 μg/ml ethidium-bromide. MNase activity assay was conducted at five different concentrations of nanobody-MNase (30, 10, 3, 1, 0.3 ng), protA-MNase (30, 10, 3, 1, 0.3 ng) and commercial MNase (20, 6, 2, 0.6, 0.2 U).

Cell lines generation

HeLaFlp-In/T-REx cells, containing the Flp Recombination Target site and expressing the Tet Repressor (12) were grown in Dulbecco's Modified Eagle's Medium (DMEM), 4.5 g/l glucose, supplemented with 10% (v/v) fetal bovine serum, 10 mM L-glutamine and 100 U/ml penicillin/streptomycin (all purchased from Lonza), together with 5 μg/ml blasticidin S (InvivoGen, San Diego, CA, USA) and 200 μg/ml zeocin (Invitrogen, Carlsbad, CA) as selection drugs for the FRT site and the Tet repressor, respectively. To create dox-inducible expression cell lines the GFP-fusion destination vectors were co-transfected with pOG44 plasmid encoding for the Flp recombinase into HeLaFlp-In/T-REx cells using polyethyleneimine (PEI) transfection to generate stable Dox-inducible expression cell lines (12). Recombineered cells were selected by replacing zeocin by 250 μg/ml hygromycin B (Roche Diagnostics, Mannheim, Germany) 48 h after PEI transfection. Expression of GFP-tagged protein was induced by addition of 1 μg/ml doxycycline for 16–20 h.

Immunoblotting procedures

Cells were seeded in six-well dishes at 30,000 cells per well and induced with doxycycline for 22 h prior to harvesting. Cell lysates were prepared in 1× sample buffer (120 mM Tris–HCl pH 6.8, 4% SDS, 20% glycerol, 0.05% bromophenol blue). Equal amounts of cell lysates were separated by 10% SDS-PAGE and transferred onto nitrocellulase membrane. These immunoblots were developed with the appropriate antibodies and Clarity ECL reagents (Bio-Rad). Immunoblots were analyzed using ChemiDoc imaging system (BioRad). The images were subjected to linear contrast/brightness enhancement in Photoshop (CS6, 13.0.6 × 64, extended) when needed for data representation purposes. The used antibodies were directed against GFP (JL-8, Clontech), TBP (20C7, in house monoclonal) or vinculin (7F9, SantaCruz) and diluted to the appropriate concentration.

Extract preparation and GFP-tagged protein purification

Cells were seeded in 15-cm dishes (Greiner Cellstar) and grown to 70–80% confluence prior to induction with 1 μg/ml doxycycline for 22 h. GFP-protein expression was verified using ZOE fluorescence microscopy (BioRad). Next, induced cells were harvested and nuclear and cytoplasmic extracts were obtained using a modified version of the Dignam procedure (13). Protein concentrations were determined by Bradford assay (BioRad). 1 mg of nuclear extract was used for GFP-affinity purification as described (14). In short, protein lysates were incubated in binding buffer (20 mM HEPES–KOH pH 7.9, 300 mM NaCl, 20% glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% NP-40, 0.5 mM DTT and 1× Roche protease inhibitor cocktail) on a rotating wheel for 1 h at 4°C in triplicates with GBP-coated agarose beads (Chromotek) or control agarose beads (Chromotek). The beads were washed two times with binding buffer containing 0.5% NP-40, two times with PBS containing 0.5% NP-40, and two times with PBS. On-bead digestion of bound proteins was performed overnight in elution buffer (100 mM Tris–HCl pH 7.5, 2 M urea, 10 mM DTT) with 0.1 μg/ml of trypsin at RT and eluted tryptic peptides were bound to C18 stage tips (ThermoFischer, USA) prior to mass spectrometry analysis.

Mass spectrometry and data analysis

Tryptic peptides were eluted from the C18 stage tips in H_2O:acetonitril (35:65) and dried prior to resuspension in 10% formic acid. A third of this elution was analyzed by nanoflow-LC–MS/MS with an Orbitrap Fusion Lumos mass spectrometer coupled to an Easy nano-LC 1200 HPLC (Thermo Fisher Scientific). The flow rate was 300 nl/min, buffer A was 0.1% (v/v) formic acid and buffer B was 0.1% formic acid in 80% acetonitrile. A gradient of increasing organic proportion was used in combination with a reversed phase C18 separating column (2 μm particle size, 100 Å pore size, 25 cm length, 50 μm i.d., Thermo Fisher Scientific). Each MS can was followed by a maximum of 10 MS/MS scans in the data dependent mode with 90 min total analysis time. Blank samples consisting of 10% formic acid were run for 45 min between GFP and non-GFP samples, to avoid carry-over between runs.

The raw data files were analyzed with MaxQuant software (version 1.5.3.30) using Uniprot human FASTA database (14,15). Label-free quantification values (LFQ) and match between run options were selected. Intensity based absolute quantification (iBAQ) algorithm was also activated for subsequent relative protein abundance estimation (16). The obtained protein files were analyzed by Perseus software (MQ package, version 1.5.4.0 for NFYA, version 1.6.12 for FOS), in which contaminants and reverse hits were filtered out (15). Protein identification based on non-unique peptides as well as proteins identified by only one peptide in the different triplicates were excluded to increase protein prediction accuracy.

For identification of the bait interactors LFQ intensity-based values were transformed on the logarithmic scale (log_2) to generate Gaussian distribution of the data. This allows for imputation of missing values based on the normal distribution of the overall data (in Perseus, width = 0.3; shift = 1.8). The normalized LFQ intensities were compared between grouped GFP triplicates and non-GFP triplicates, using 1% and 5% in permutation-based false discovery rate (FDR) in a two-tailed *t*-test for NFYA and FOS, respectively. The threshold for significance (S0), based on the FDR and the ratio between GFP and non-GFP, samples was kept at the constant value of 3 and 1 for comparison purposes for NFYA and FOS, respectively. Relative abundance plots were obtained by comparison of the iBAQ values of GFP interactors. The values of the non-GFP iBAQ values were subtracted from the corresponding proteins in

the GFP pull-down and were next normalized on a chosen co-purifying protein for scaling and data representation purposes. All mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository under the dataset identifier PXD021089.

Standard and greenCUT&RUN protocol

The standard antibody-based CUT&RUN protocol for endogenous and GFP-tagged proteins was performed essentially as described by Skene *et al* (6). The protocol of greenCUT&RUN is summarized in Figure 1A. In brief, the greenCUT&RUN protocol is as follows. Initially, concanavalin A–coated (ConA) magnetic beads per samples (10 μl for 0.25 million cells, 20 μl ConA beads were used for 1 million cells) were incubated with 1.5 ml of binding buffer (20 mM HEPES–KOH (pH 7.9), 10 mM KCl, 1 mM $CaCl_2$ and 1 mM $MnCl_2$) on magnetic stand for 2 min and centrifuged at <100 *g* for 1 s. Supernatant was removed and same step is repeated again. Further, ConA-beads were stored in 10 μl of binding buffer on ice. In total, 1 million cells (0.25 million in case of NFYA and FOS) were freshly harvested by centrifugation at 600 *g* for 3 min. After removing supernatant, cells were washed with wash buffer (20 mM HEPES–KOH (pH 7.5), 150 mM NaCl and 0.5 mM spermidine and EDTA-free complete protease inhibitor). Washing was performed two times and supernatant was removed after centrifugation at 600 *g* for 3 min. After this, cells were resuspended in 1 ml of wash buffer and incubated with ConA-beads for 5 to 10 min. Further, ConA bound cells were permeabilized with buffer (2 mM EDTA, 0.05% digitonin wash buffer) for 4 min. To minimize cell stress, all experiments were performed at room temperature up to this stage. In the next step, cells were treated with 0.4 μg of nanobody-MNase in 100 μl digitonin buffer (0.05% digitonin in wash buffer) for 30 min at 4°C. Cells were washed 2 times and resuspended in 150 μl of digitonin buffer. Cells were kept on ice and chromatin digestion was started after adding $CaCl_2$ to 3 mM on ice. After 30 min, MNase digestions were stopped by adding 100 μl of 2× stop buffer (340 mM NaCl, 20 mM EDTA, 10 mM EGTA, 0.02% digitonin, 100 μg/ml of RNase A and 50 mg/ml glycogen). 20 pg of *Drosophila* spike-in controls were added in the reactions. Reaction mixtures were incubated for 10 min at 37°C to release DNA fragments from insoluble nuclear chromatin to the supernatant and centrifuged for 5 min at 16,000 *g* and 4°C. Supernatants (100 μl) were removed on magnetic stand and treated with 2 μl of 10% (wt/vol) SDS and 1.5 μl of proteinase K (20 mg/ml) for 10 minute at 70°C. DNA was extracted using phenol–chloroform, precipitated with 100% ethanol and dissolved in 1 mM Tris–HCl pH 8.0/0.1 mM EDTA. For the negative controls, Ca^{2+} was omitted.

Spike-in controls were prepared from the nuclear pellet of *Drosophila* S2 cells. Briefly, S2 cells were grown to confluence and harvested by trypsin digestion. After washing with PBS, cell extracts were prepared by the Dignam protocol (17). After nuclear extraction the chromatin pellet was obtained by centrifugation at 3200 *g* for 30 min at 4°C. This pellet was homogenized in HS buffer (20 mM HEPES–KOH pH 7.9, 150 mM NaCl, 1 mM EDTA, 0.34 M sucrose, 1 mM DTT, 0.5 mM PMSF) using a Turrox Tissue homogenizer at 12 krpm for 1 min on ice. The washed chromatin was obtained by centrifugation at 259,000 *g* for 30 min at 4°C. This pellet was homogenized in NI buffer (20 mM HEPES–KOH pH 7.8, 10 mM $MgCl_2$, 0.25 M sucrose, 0.1 Triton X-100 EDTA, 1 mM DTT, 0.5 mM PMSF) using a Turrox Tissue Homogenizer at 6 krpm for 1 min on ice. $CaCl_2$ was added to 3 mM and the suspension was warmed to 37°C. Micrococcal Nuclease (New England BioLabs) was added to 4 μl/ml and the digestion was allowed to proceed for 10 min at 37°C. After this, EGTA was added 10 mM and NaCl to 150 mM and the suspension was centrifuged at 260,000 *g* for 30 min at 4°C. The supernatant contained purified mononucleosomes, which was confirmed by Coomassie staining of 15% polyacrylamide/SDS gels and by DNA analysis of ∼150 bp fragments by agarose gel electrophoresis in TBE and staining with ethidium-bromide. To find enh-MNase hypersensitive regions in genome, greenCUT&RUN was performed using 0.25 million HeLa cells lacking GFP-tagged proteins. Here, 10 pg of *Drosophila* spike-in were added in the solution and only 50% of the leached out DNA were utilized in library preparation.

ChIPseq protocol

HeLa Flp-In/T-REx cells were induced for 17–24 h with 1 μg/ml doxycycline for GFP-tagged protein expression, crosslinked for 8 min using 1% formaldehyde and the reaction quenched with 125 mM glycine for 5 min. GFP-TBP and GFP-NFYA expressing cells were used for NFYA ChIP (Endo-ChIPseq) and GFP ChIP (GFP-ChIPseq), respectively. Cells were lysed in cell lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP40) and centrifuged at 3200 *g* for 5 min at 4°C. The pellet was lysed in nuclei lysis buffer (50 mM Tris–HCl pH 8.1, 10 mM EDTA pH 8.0) containing 0.5% SDS. The DNA was sonicated to 200–600 bp using a Covaris S220 instrument (settings: 12 min, PIP = 140; duty factor = 5%; CPB = 200; $T = 2.2$°C). The 40 μg of chromatin were incubated with either 2 μg of NFYA antibody (Santa Cruz, sc11753) or 5 μg of GFP antibody (GenScript, A01704) in IP buffer (16.7 mM Tris–HCl pH 8.1, 83.5 mM NaCl, 83.5 mM LiCl, 0.01% SDS, 1.10% Triton X-100, 1.2 mM EDTA pH 8.0) overnight at 4°C. Input (pre-IP) samples were prepared in parallel. All lysis and IP buffers were supplemented with EDTA-free protease inhibitors (Roche complete, 11873580001). Immunoprecipitated complexes were recovered using blocked (100 μg/ml BSA) protein-A agarose or magnetic beads for 1–4 h at 4°C and washed 2–5 times using IP wash buffer 1 (20 mM Tris–HCl pH 8.1, 50 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% Triton X-100), one time with IP wash buffer 2 (10 mM Tris–HCl pH 8.1, 250 mM LiCl, 1 mM EDTA pH 8.0, 1% NP-40, 1% Na-deoxycholate), and two times with TE buffer (10 mM Tris–HCl pH 8.1, 1 mM EDTA pH 8.0). Samples were eluted in elution buffer (100 mM NaHCO3, 1% SDS), RNase A-treated and reverse-crosslinked overnight at 65°C. DNA was de-proteinized for 1.5 h at 45°C using proteinase K and purified with the NucleoSpin Extract II kit (Machery-Nagel).

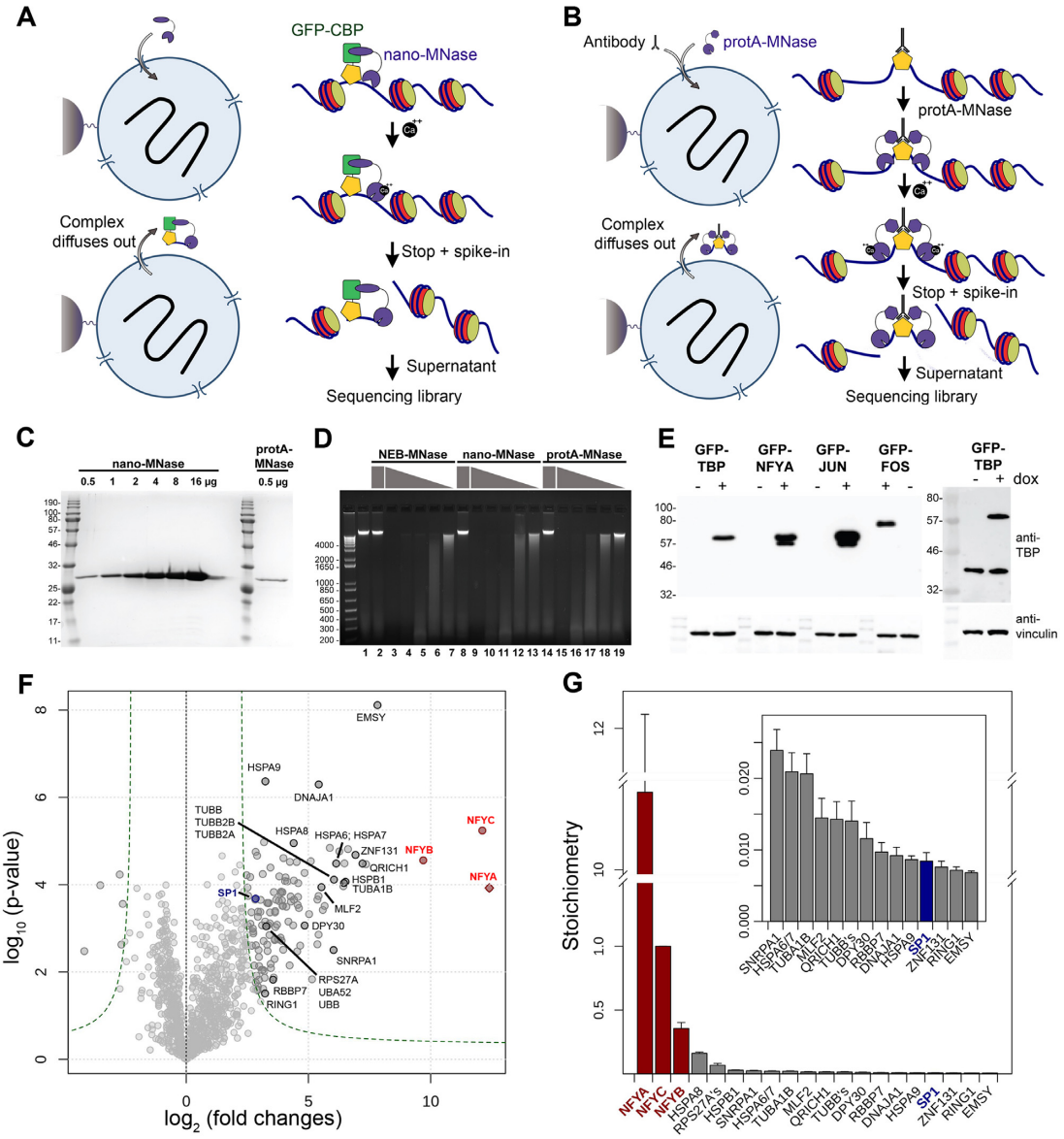

Figure 1. Experimental approach of greenCUT&RUN. Panels (**A** and **B**): Schematic of experimental strategy for greenCUT&RUN (A) and CUT&RUN (B). GreenCUT&RUN is rapid and easy protocol, which involves only three steps to complete. Panel (**C**): The GST protein fused to the GFP nanobody and MNase (nanobody-MNase) was expressed and purified from bacteria. A single band of nanobody-MNase and protA-MNase were observed by coomassie staining of protein gels (panel **C**). Panel (**D**): To access activity genomic DNA was treated with purified nanobody-MNase at different concentrations and compared with standard MNase and protA-MNase preparations. In panel D, lane 1 shows uncut genomic DNA, while lanes 2, 8 and 14 shows DNA after adding MNase in solution lacking Ca^{2+}. Lanes 3–7, 9–13 and 15–19 shows DNA fragments after activating MNase with Ca^{2+}. From lanes 3–7, 9–13 and 15–19, decreasing amount of MNases were used to access activity. Similar expression levels of endogenous and GFP-tagged TBP were observed. Panel (**E**): Expression of GFP-tagged NFYA, JUN, FOS and TBP after doxycycline induction as detected by GFP antibodies (left). Comparison of expression levels of GFP-tagged TBP with endogenous TBP (right). Panel (**F**): Volcano plot of GFP-tagged NFYA. Panel (**G**): Relative stoichiometries of top 20 proteins. In the stoichiometry calculation, NFYC is used for normalization.

Library preparation and sequencing

For CUT&RUN, purified DNA fragments were subjected to library preparation with NEB Next Ultra II and NEB Multiplex Oligo Set I/II as per manufacturer (New England Biolabs) protocol without size selection. In case of TBP, we followed the library preparation protocol of Skene *et al.* (6). For each library, DNA concentration was determined using using a Qubit instrument (Invitrogen, USA) and size distribution was analyzed on Agilent Bioanalyzer (DNA high sensitivity assay) or Agilent 4200 TapeStation (D5000 Screentape). Libraries were pooled according to desired molarity with 1% PhiX controls for run on the MiniSeq (Illumina) or without PhiX for runs on the MiSeq (2 × 75 paired end, Illumina). All NGS data have been deposited to Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) under the accession number SRP278136.

For ChIPseq, 5 ng input and less than 500 pg of ChIP purified DNA were used to prepare paired-end sequencing libraries using the NEBNext Ultra II DNA Library Prep Kit

for Illumina (New England Biolabs, E7645L) without size selection. The 75-nucleotide paired-end sequencing reads were generated (Illumina, HiSeq 3000) with 20–23 M reads per sample.

Datasets used for comparison

To compare the performance of greenCUT&RUN, ChIPseq data were downloaded from ENCODE (https://www.encodeproject.org/) for human genome version hg38 (18). Complete details are given in Supplementary Table S1.

Quality control, alignment and normalization

Initially, reads were passed through quality control filtering using Trim-galore (v0.6.3) with default parameters (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and further aligned using bowtie2 (v2.3.4.1) with option: –dovetail –local –very-sensitive-local –no-unal –no-mixed –no-discordant -I 10 -X 700(19). Reads for HeLa cells were aligned on hg38/GRCh38.p13 version of genome (https://www.gencodegenes.org/human/) while spike-in was aligned on BDGP5 version of Drosophila genome sequence (EnsEMBL v.75, www.ensembl.org) (20,21). Only in case of ChIPseq, duplicate reads were filtered out using Picard (v2.18.14) with default parameters (http://broadinstitute.github.io/picard). Flagstat program of samtools was used to calculate total number of aligned reads, corresponding to human and spike-in (22). To compare the performance of the different protocol, equal number of reads was selected randomly using sambamaba (23) and tag directories were generated using makeTagDirectory program of Homer with option: -totalReads (24). Here, spike-in normalized total number of read was used for option '-totalReads'. Suppose, total human and spike-in reads in experiment are E_h and E_S while in control are C_h and C_S, total spike-in normalized reads for experiment (E_T) and control (C_T) will be:

$$C_T = \frac{C_h}{E_s} \times C_s \text{ and } E_T = E_h$$

Moreover, in case of ChIPseq where spike-in controls were not present, total number of uniquely aligned reads was used as such. These tag-directories were further used for computing coverage and peak calling.

Coverage calculation

Coverage was calculated around 2000 bp from centre of motifs (in the case of NFYA and FOS) or ATAC peaks (in the case of TBP). Initially, whole human genome was scanned and genomic coordinates of putative functional binding sites (motifs) of NFY and FOS were identified using scanMotifGenomeWide.pl program of HOMER. Motifs present in chrY were filtered out from further analysis, and coverage (per base-pair per motifs) was computed using annotatePeaks.pl. In case of TBP, coverage was calculated around ATAC peaks (accession ID: GSE121840), as determined by Oomen *et al.* (25). Genomic coordinate of this data-set was lifted to hg38 from hg19, using UCSC genome browser utility program liftover (26). Base package of R was used to generate graphs (27).

Reproducibility among replicates

To explore reproducibility among replicates, coverage was calculated within 10kb non-overlapping genome-wide bins, with multiBamSummary program of deepTools (v2.0) (28). Further, Pearson correlation coefficient was calculated using plotCorrelation with option: –removeOutliers.

Peak calling

Different peak calling algorithms (HOMER, MACS and SEACR) were compared using their default parameters. As HOMER performed best in terms of the number of peaks and peaks with motifs it was selected for all our analyses (Supplementary Table S2). Peaks were called using findPeaks program of HOMER with default parameters, which considers both local background and the reads present in the control. The no Ca^{2+} dataset was used as a control for peak calling. In case of green and standard-CUT&RUN, a high frequency of fragments with identical ends can arise from different cells. Therefore, peak filtering based on clonal signals was disabled using –C 0 in findPeaks. To identify peaks in ENCODE ChIPseq datasets (NFYA, FOS and TBP), irreproducibility discovery rate (IDR) analysis was performed against replicated samples using homer-idr package (https://github.com/karmel/homer-idr). In case of TBP greenCUT&RUN, we generated ∼1/3 reads in control compared to experiment and could not identify peaks using HOMER. Therefore, two steps were followed to call peaks. In first step, peak were called without filtering it against control and clonal signals were kept off, using –F 0 –C 0. In second step, getDifferentialPeaks was used to filter peaks in experiment against control (fold changes ≥4 and P-value $< 10^{-4}$). In all experiments, peaks present within black-listed regions of ENCODE (https://github.com/Boyle-Lab/Blacklist/) were excluded using intersectBed program of bedtools (v2.27.1–5) (29,30). In-house R script was developed to plot coverage tracks around peaks using bigwig files returned from bamCoverage. To identify the peaks in GROseq, the R package groHMM was used (31).

Signal intensity around peaks: Heatmap

To visualize signal intensity around peaks, heatmaps were generated. Initially, spike-in normalized coverage in experiment was computed against control, using bamCompare program with option –scaleFactors 1: S, were S is equal to ratio of spike-in in control versus experiment. In the case of ChIPseq, default parameters of bamCompare were used. Next, coverage was extracted within 2000 bps from center of peaks using computeMatrix and plotted with plotHeatmap program of deeptools (28). In ATACseq, control groups are not typically run. Therefore, signal intensity was calculated without control, using bamCoverage.

Cut frequency

In-house script (https://github.com/snizam001/cutfrequency) was generated to calculate spike-in normalized cut frequency against control at base-pair resolution. Initially, motifs present near to each other within 1000 bp

were removed and coordinates of read ends were identified using bamtobed around 100 bps around motifs. For each pair-end reads, either 3′ or 5′ end which is near to motif is considered in computing cut frequency. Cut frequency (*Cf*) at j^{th} position from centre of motif is defined as follows:

$$Cf_j = 10^6 \frac{1}{N_{motifs}} \times \sum_{i=1}^{i=N_{motifs}} \left\{ \left(\frac{R_{expr_i}}{N_{expr}} \right) - \left(\frac{R_{ctrl_i}}{N_{ctrl}} \times \frac{E_s}{C_s} \right) \right\}$$

where (i) N_{motifs} is total number of whole-genome motifs, (ii) N_{expr} and N_{ctrl} are total number of reads in experiment and control, (iii) R_{expr_i} and R_{ctrl_i} are number of read ends at j^{th} position from i^{th} motif and (4) E_s and C_s are total number of spike-in reads in experiment and control, respectively.

Comparison and annotation of peaks

Unique and overlapping peaks among protocols were identified using mergePeaks program of Homer with option: -d 200 (24). To identify real unique peaks, coverage in one protocol was compared with another using getDifferentialPeaks (24). In the current study, peaks with $\log_2$ of fold changes $\geq$2 and *P*-value $\leq 10^{-4}$ in one protocol compared to other, were considered as real unique peaks. FindMotifsGenome.pl was used to find enriched motifs in peaks (24). Moreover, peaks were annotated using annotatePeaks.pl and coordinates of motifs respective to center of peak were identified. An In-house pipeline in the perl language was developed to annotate peaks lacking consensus motifs as illustrated in Supplementary Figure S1. FindMotifsGenome.pl of homer package was used to find known motifs in those peaks lacking either consensus motifs. Genomic coordinates of histone marks were obtained from ENCODE repository data-sets and windowBed of bedtools was used to classify peaks having histone marks within 400 bps from the centre of the peak (18,30).

Docking analysis

Initially, PDB files with accession IDs 4AWL and 1FOS were downloaded from https://www.rcsb.org/ for NFY-DNA, FOS/JUN-DNA complex, respectively. For both crystallographic structures, the DNA was extended on both sides taking a standard B-DNA conformation. The resulting DNA lengths were 108 and 85 bps for the NFY and FOS DNA complexes. For the MNase, we used PDB ID 2SNS, removing the bound nucleotide and keeping the calcium ion. The docking was performed using HADDOCK (version 2.4) (32) using two ambiguous distance restraints with an upper limit of 5 Å defined between the guanidium group of Arginie 35 and 87, respectively (those are coordinating the phosphate of the bound nucleotide in PDB ID: 2SNS), and any phosphate atom of the DNA, excluding the first and last three based pairs (to avoid end effects). In order to only sample the accessibility of the DNA to the MNase, a repulsion function (*repel* option in CNS (Crystallographic and NMR System) (33) with a van der Waals scaling of 0.89 was used to avoid steric clashes instead of the default electrostatics and van der Waals energy functions. Hundred thousand rigid-body docking models were generated (the flexible refinement stages of HADDOCK were skipped). Those were analyzed by extracting all contacts within a 7.5 Å distance cut off between any phosphate atom of the DNA and Arginine 35 of MNase. Based on the presence of such contacts the accessible DNA bases were identified.

RESULTS

Experimental design of greenCUT&RUN

Camelids have single chain antibodies, also known as nanobodies. Several nanobodies have been isolated against green fluorescence protein (GFP), which display a higher binding affinity compared to standard antibodies. These GFP-nanobodies have been successfully utilized in many techniques including affinity purification for quantitative proteomic profiling (10). We exploited this property in developing greenCUT&RUN (Figure 1A). In brief, unfixed cells were immobilized on magnetic Concanavalin A (ConA)-coated beads and permeabilized with digitonin. Next, the immobilized cells were incubated with nanobody-MNase (monococcal nuclease) and after washing, the cleavage reaction was initiated by adding 3 mM $CaCl_2$. The reaction was stopped by adding a Ca^{2+}-specific chelating agent and *Drosophila* mononucleosomal DNA was added as a spike-in control for later normalization. Released DNA fragments were extracted using phenol-chloroform followed by ethanol precipitation to be used for next generation sequencing (NGS) DNA libraries. Parallel reactions lacking Ca^{2+} activation of MNase were used as negative controls. In comparison to CUT&RUN (Figure 1B), greenCUT&RUN requires only a single incubation step instead of separate incubations with primary antibody and with ProtA/G-MNase (6).

To generate the nanobody-MNase fusion protein, the variable domain of a GFP-specific nanobody (V_H) was linked to the catalytic domain of micrococcal nuclease (MNase) with Asp-Asp-Asp-Lys-Glu-Phe as a linker to provide flexibility. This fusion protein was linked to glutathione S-transferase (GST) for expression in bacteria and subsequent protein purification. The nanobody-MNase was purified in three steps using glutathione-affinity chromatography followed by thrombin cleavage of the GST moiety and chromatography by gel filtration and cation-exchange. Purity of the nanobody-MNase preparation was assessed by coomassie staining of protein gels, which indicated a pure protein with the expected size of ∼29 kDa (Figure 1C). For comparison a protein A-MNase (protA-MNase) preparation (kindly provided by the Henikoff lab) was analyzed in parallel (6). The activity of nanobody-MNase preparation was assessed by digestion of genomic DNA of HeLa cells and compared with commercial MNase (from New England Biolabs) and protA-MNase. Staining of agarose gels indicate similar specific activities of all three MNases (Figure 1D).

To develop the CUT&RUN protocol for GFP-tagged proteins, we employed stable HeLa cell lines expressing GFP-fusion proteins in a doxycyclin (dox)-inducible manner (12). We focused on three DNA sequence specific transcription factors (TFs), the alpha subunit (NFYA) of nu-

clear transcription factor Y, the FOS proto-oncoprotein (FOS) of the AP-1 transcription factor and the basal transcription factor TATA-box binding protein (TBP). Fusion proteins of the expected size were detected after dox-treatment (Figure 1E). Similar expression levels of GFP-tagged and endogenous TBP were observed. While all three subnits of NFY are expressed in our HeLa cells, FOS is not expressed in HeLa cells under normal growth condition. To validate NFY complex formation, we examine the interactome of the GFP-NFYA protein by quantitative mass spectrometry (12). As expected, GFP-NFYA interacts with NFYB and NFYC (Figure 1F). The stoichiometry plot indicates that a significant part of the GFP-NFYA protein is not in a complex with NFYB and NFYC, but this should not result in spurious DNA binding as NFY complex binds to DNA as an obligate heterotrimer (34,35). Free GFP-NFYA is probably bound by protein chaperones as HSPB1, HSPA6/7 and HSPA8/9 are amongst the top interactors. Transcription factor SP1 is also amongst the top 20 interactors (Figure 1G) (36,37), which will become relevant later.

Performance of greenCUT&RUN: NFYA

To evaluate the performance of greenCUT&RUN, the results were compared with data obtained by standard CUT&RUN using antibodies against endogenous NFYA (hereafter, abbreviated as (Endo)-CUT&RUN) or against GFP for tagged NFYA (hereafter, abbreviated as (GFP)-CUT&RUN). The different CUT&RUN protocols were also compared to three independent ChIPseq experiments. The (Endo)-CUT&RUN was performed on endogenous NFYA using naive HeLa cells. The other techniques: green-CUT&RUN, (GFP)-CUT&RUN and ChIPseq, were performed on HeLa cells expressing GFP-tagged NFYA. CUT&RUN-based datasets were normalized to ∼2.1 million reads to compare their performance, whereas 7–14 million reads of the ChIPseq datasets were used (Supplementary Table S1).

The datasets were compared in different ways. First, we determined the sensitivity of detecting NFYA binding events at cognate binding sites by calculating the whole genome coverage of 1 109 981 putative NFY binding sites (AGC[CCAAT]CGG). The highest coverage was obtained with the greenCUT&RUN and (Endo)-CUT&RUN approaches (Figure 2A). CUT&RUN protocols generate shorter DNA fragments compared to sonication-based fragmentation, which may produce a higher frequency of fragments with identical ends. For this reason, we removed duplicate reads from ChIPseq, but not from CUT&RUN datasets. When ChIPseq coverage was calculated without removing duplicates, the greenCUT&RUN protocol still had a higher coverage (Supplementary Figure S2A). The (GFP)-CUT&RUN did not perform as well as compared to green and (Endo)-CUT&RUN. Hence, we performed (GFP)-CUT&RUN with other two GFP-specific antibodies ((GFP)-CUT&RUN-2 and -3), but they all showed a lower coverage of NFY binding sites (Supplementary Figure S2B). Reproducibility of greenCUT&RUN was evaluated by correlating coverage within 10-kb bins among replicates. We observed a high reproducibility (Pearson correlation coefficient $r^2 \geq 0.97$) with greenCUT&RUN similar to CUT&RUN ($r^2 \geq 0.95$) (Figure 2B) (7). Coverage around NFYA peaks exclusively identified by green-CUT&RUN is illustrated by genomic tracks (Supplementary Figure S3), which show that unique peaks identified by greenCUT&RUN are highly reproducible. Peak calling of greenCUT&RUN identified highest number of peaks (7433) (Figure 2C). In comparison to greenCUT&RUN, ∼24%, ∼57% and ∼33% fewer peaks were identified in (Endo)-CUT&RUN, (GFP)-CUT&RUN and ChIPseq, respectively. All three (GFP)-CUT&RUN experiments performed poorly compared to greenCUT&RUN (Supplementary Figure S4). As shown by the heatmaps of Figure 2C greenCUT&RUN provides the most robust NFYA occupancy compared to the other protocols. Previous studies showed that open chromatin regions are hypersensitive to sonication-based fragmentation (38). To examine a potential bias for open chromatin, we performed green-CUT&RUN on naive HeLa cells lacking the expression of any GFP-tagged protein. Coverage analysis did not show any enrichment of reads at NFYA peaks obtained by any genome mapping protocol (last column of Figure 2C). This supports the conclusion that the peaks correspond to *bona fide* NFYA binding events.

High resolution of NFYA binding as detected by green-CUT&RUN

To examine the resolution of greenCUT&RUN mapping the closest MNase cleavage site relative to the CCAAT-box was determined from the paired-end reads and the cut frequency was calculated (39). We examined whether this could reveal unique footprints of transcription factor binding at single base-pair resolution. For this analysis, 629 213 putative NFY binding sites were excluded as they map within 1000 bps of second CCAAT sequence. The ending of each paired-end read (either 3′ or 5′) near to the 'A' nucleotide of NFYA consensus motif (AG[CCAAT]CGG) was identified and used to calculate normalized cut frequency. This indicated that 31 bps (13 bp at the 5′- and 17 bps at the 3′-end from centre of motif) were well protected (Figure 2D). This result is reproduced in all three independent greenCUT&RUN replicates (Supplementary Figure S5). Of note, fragments were generated by MNase digestion and steric hindrance between MNase and the DNA-binding factors will result in a larger DNA footprint than the actual DNA sequence contacted by the binding factors (see also Discussion). The end mapping result indicates that <31 bp are protected by the NFY complex. We noted an elevated cut frequency at the motif centre, which is unique to green- and (GFP)-CUT&RUN (Figure 2D and Supplementary Figure S5). NFYA induces ∼80° bend at the DNA binding site which exposes its centre from one side (Supplementary Figure S6) (40). In the (Endo)-CUT&RUN only one flexible hinge is present between MNase and protein A moieties. In both green and (GFP)-CUT&RUN two flexible hinges are present in one between the GFP and the protein of interest, and one between the MNase and nanobody or protein A moieties (6). This may allow sufficient flexibility for the MNase to reach the exposed part of the DNA motif and cleave DNA at the centre of the motif in green-CUT&RUN. Moreover, the MNase can access a larger se-

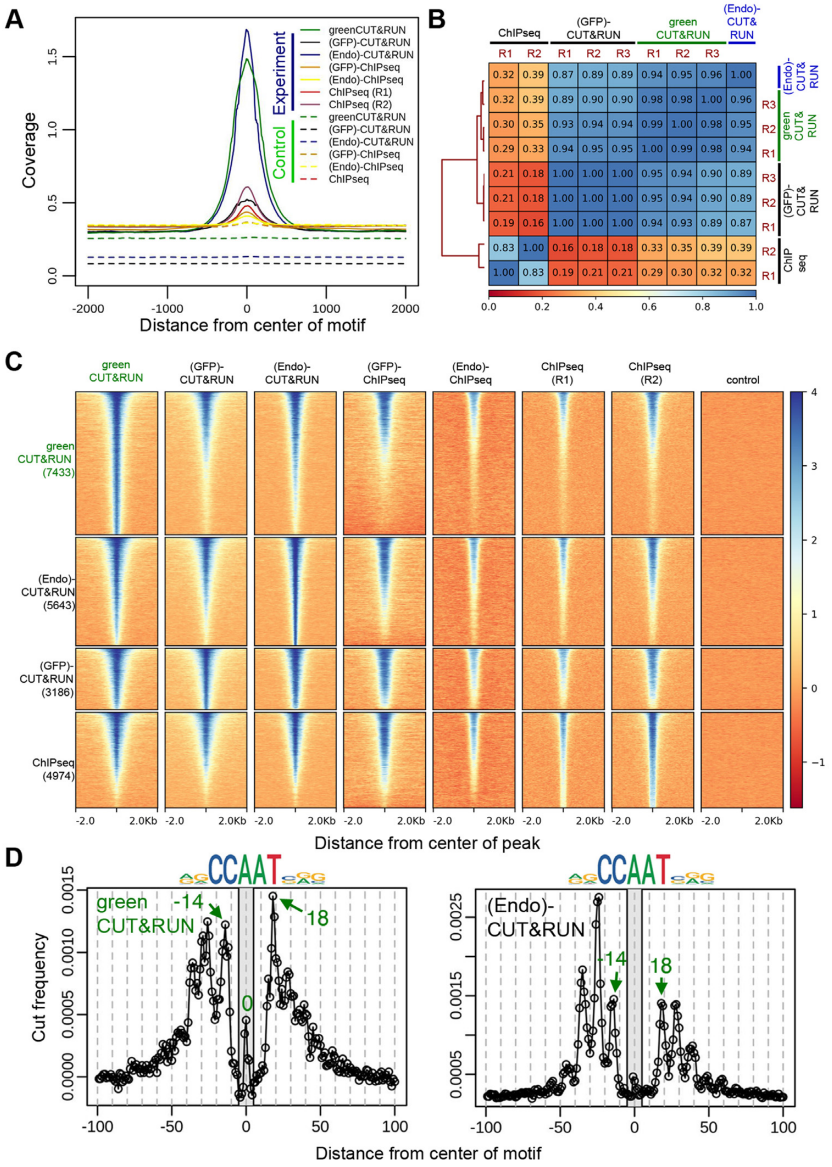

Figure 2. Performance of greenCUT&RUN for NFYA compared with other genome-mapping approaches. Panel (**A**): Coverage was calculated separately around all putative genomic binding sites of NFYA (1,109,981). Maximum coverage was obtained with the greenCUT&RUN and (Endo)-CUT&RUN protocols. Panel (**B**) shows that greenCUT&RUN data displays high reproducibility among replicates. Values represent Pearson correlations of coverage calculated for 10-kb bins. Panel (**C**): Signal intensity of genome-wide peaks in different protocols shown in the form of heatmaps. In all experiments except ChIPseq ∼2 million reads were analyzed (details are given in Supplementary Table S1). The numbers of identified peaks are given in brackets on the left. The highest number of peaks was identified by greenCUT&RUN. Panel (**D**): Normalized cut frequency of MNase around 480,768 putative binding sites of NFYA. In both green and (Endo)-CUT&RUN, 31 nucleotides are protected from MNase digestion. Numbering is relative to the central basepair of the recognition site, CCAAT. Grey area in the graph represents the width of motif.

quence around the binding sites due to the additional search space generated by tagging the protein with GFP.

Sensitivity and specificity of greenCUT&RUN for NFYA

To evaluate the detection sensitivity of binding events, NFYA peaks were compared between different protocols. In total, 3171 peaks were uniquely identified by green-CUT&RUN, which is the highest of all protocols (Figure 3A). However, it is possible that the protocols respond differently to the stringency of the criteria used for peak-calling (false discovery rate $\leq$ 0.001, fold changes $\geq$4 and *P*-value of fold change $\leq 1 \times 10^{-4}$). For example, a peak at chr1:206503584–206503784 was clearly detected by greenCUT&RUN, but other protocols only detected weak NFY binding (Figure 3B). This region contains multiple NFY consensus sites. By setting, FDR (false discovery rate) threshold at 50% and disabling filtering, the same peak was observed with other protocols. Therefore, we compared coverage of peaks between protocols and considered only those peaks as 'unique', which had a significantly higher coverage in one protocol compared to other. In total, 1965 (26.43%), true-positive unique peaks were identified in greenCUT&RUN versus (Endo)-CUT&RUN.

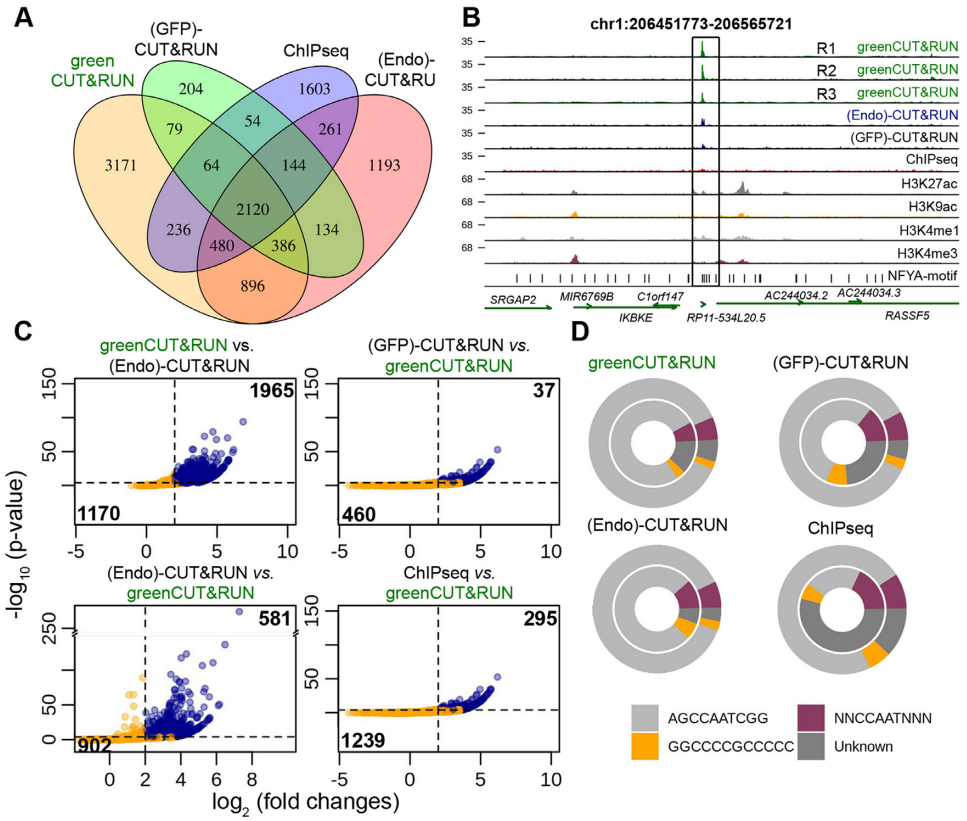

Figure 3. Sensitivity and accuracy of greenCUT&RUN for NFYA. Panel (**A**): A Venn diagram represents overlapping and unique peaks identified by the different protocols. Of which, the maximum number of unique peaks were identified with greenCUT&RUN. Panel (**B**) displays coverage around peaks. Due to the stringent criteria of the peak calling algorithm this peak was only detected in greenCUT&RUN. Panel (**C**): Coverage of unique peaks was compared among protocols and real number of unique peaks is shown in the upper left corner. Panel (**D**): Distribution of transcription factor motifs. Outer and inner circles represent motif distribution in total and unique peaks, respectively.

Only 37 (1.16% of total peaks), 581 (10.29%) and 295 (5.93%) peaks were unique in (GFP)-CUT&RUN, (Endo)-CUT&RUN and ChIPseq, respectively, when compared to greenCUT&RUN (Figure 3C). In the other two (GFP)-CUT&RUN experiments using different GFP antibodies, only 1.09% and 0.54% peaks were unique compared to greenCUT&RUN (Supplementary Figure S4C). This suggests that greenCUT&RUN is highly sensitive even at a coverage of only ∼2.1 million reads. It is interesting to note that >90% peaks could be captured with greenCUT&RUN with less stringent peak definitions, while this is not observed in other methods. These observations are consistent with the heatmaps of Figure 2C, which indicate that greenCUT&RUN covers >90% peaks of the other protocols. Next, the distribution of histone marks associated with active/inactive promoter and enhancer regions around NFYA peaks obtained in the different protocols was explored. In all cases the majority of peaks were associated with histone marks indicative of active promoter and enhancer regions (Supplementary Figure S7).

Next, the distribution of the NFY motif was explored to evaluate the specificity of protocols (Supplementary Figure S1). In total, 6453 (86.81%), 4935 (87.6%), 2721 (85.4%) and 3631 (72.99%) peaks with motif were identified by greenCUT&RUN, (Endo)-CUT&RUN, (GFP)-CUT&RUN and ChIPseq, respectively (Figure 3D and Supplementary Table S3). In the other two (GFP)-CUT&RUN experiments, ∼88% of the peaks contain the NFY motif (Supplementary Figure S4D). Of these, ChIPseq had significantly lower number of peaks with motif (χ^2 proportion test; P-value $< 2.2 \times 10^{-16}$), while greenCUT&RUN was not significantly different from (Endo)- or (GFP)-CUT&RUN (P-value = 0.192 and 0.056 respectively). In the remaining peaks, we identified CCAAT sequence within 424 (5.7% of total peaks), 372 (6.6%), 227 (7.1%) and 439 (8.8%) of the peaks of greenCUT&RUN, (Endo)-CUT&RUN, (GFP)-CUT&RUN and ChIPseq, respectively. We examined cooperative DNA binding events between NFY and other transcription factors. To investigate such binding events in peaks without NFY-motifs, the enrichment of 414 transcription factor motifs was determined (Supplementary Table S5). The consensus motif for transcription factor SP1 was significantly enriched. SP1 motifs were present in 143 (1.9% of the total peaks), 132 (2.3%), 80 (2.5%) and 285 (5.7%) peaks in greenCUT&RUN, (Endo)-CUT&RUN, (GFP)-CUT&RUN and ChIPseq, respectively (Supplementary Table S3). In total, 2051, 1740, 1312 and 1545 Peaks with both NFY and SP1-motifs were also found in the green-CUT&RUN, (Endo)-CUT&RUN, (GFP)-CUT&RUN and ChIPseq. Interestingly, by quantitative mass spectrometry for NFYA (Figure 1F and G) we iden-

tified SP1 as the top interacting transcription factor. Only a minority of peaks cannot be explained by the presence of an NFY or SP1 motif, which is the highest (12.44%) for ChIPseq (Figure 3D). These observations suggest that SP1 can mediate NFY binding at genomic locations lacking the NFY binding sequence. The ENCODE consortium generated ChIPseq datasets for NFYA and for SP1 in the HepG2 cell line (Supplementary Table S1). We examined these datasets for NFYA binding events at regions lacking CCAAT consensus motifs. Out of these 143 peaks detected by greenCUT&RUN, ninety NFYA peaks were also detected by NFYA ChIPseq in HepG2 and 82 of these overlap with the SP1 peaks (Supplementary Figure S8). This provides independent evidence for NFY piggy-backing on SP1 to selected genomic regions.

The distribution of motifs in unique peaks can be a good indicator for specificity. Therefore, we also explored the NFY consensus sequences and SP1 motif in the unique peaks of the different NFY mapping approaches. In comparison to greenCUT&RUN (1507: 76.69%), NFYA motif distribution in (Endo)-CUT&RUN (446: 76.76%) was not significantly different (P-value ~1), while in (GFP)-CUT&RUN (20: 54.05%) and ChIPseq (64: 22.69%) was significantly lower (χ^2 proportion test; P-value = 2.596×10^{-3} and $<2.2 \times 10^{-16}$). In (Endo)-CUT&RUN, the proportion of SP1 motif containing unique peaks (6.19%) was higher compared to greenCUT&RUN (3.36%) and ChIP (6.10%). The highest proportion of SP1 motif containing peaks (8.1%) was observed in (GFP)-CUT&RUN. Moreover, ChIPseq had the highest proportion of unexplained peaks (161: 54.58%) (Figure 3D and Supplementary Table S3). This suggests that greenCUT&RUN is as specific as (Endo)-CUT&RUN and that it is more specific compared to (GFP)-CUT&RUN and ChIPseq. Of note, we did not observe any spurious signal due to the overexpression of GFP-tagged NFYA (Figure 1F) in both green and (GFP)-CUT&RUN, which is consistent with the observation that NFY only binds to DNA as a trimeric complex (34,35).

Genome-wide profiling of DNA sequence specific transcription factor FOS

The performance of greenCUT&RUN was further evaluated for the FOS protein, which forms DNA binding complexes with JUN and ATF family members (41,42). In total, ~4 million reads of greenCUT&RUN were used to compare with ~6 and 25–30 million reads of ChIPseq (FOS) and ChIPseq (Jun), respectively (Supplementary Table S1). Initially, 856,556 putative AP1 binding sites were identified in the whole human genome and coverage was computed around the centre of motif (underlined in this sequence: NDA[TGASTCA]YN). Higher coverage was obtained with greenCUT&RUN protocol in comparison to ChIPseq (without and with duplicate reads) (Figure 4A and Supplementary Figure S9, respectively).

As noticed for NFYA, the reproducibility for FOS is higher for the greenCUT&RUN protocol (Pearson correlation coefficient ~1) compared to ChIPseq (Pearson correlation coefficient = 0.86) (Figure 4B). This is also illustrated in Supplementary Supplementary Figure S10, which shows unique peaks identified by greenCUT&RUN in both replicates. In total, two-fold more peaks were identified in greenCUT&RUN (39,971) compared to ChIPseq (19,881) (Figure 4C). GreenCUT&RUN provides robust occupancy maps of FOS transcription factor compared to ChIPseq as indicated by the heatmaps (Figure 4C). Similar to NFYA, a biased enrichment of reads to open chromatin region was not observed as shown by the HeLa control (Figure 4C).

Next, we explored the FOS specific MNase footprint around the AP1 motif. To exclude interference by MNase bound to near-by AP1 sites, 397 387 motifs were removed, which were present within 1000 bp of each other. After mapping end of the reads and computing cut frequency, we observed that 17bps (8 bp at both 5′- and 3′-end from the motif centre) are protected by the FOS-complex, which is similar for both replicates (Figure 4D). In comparison to NFYA, MNase digestion at FOS binding sites generated a symmetrical pattern of protection, which is reminiscent of the symmetrical nature of AP1 binding sites.

Sensitivity and specificity of greenCUT&RUN: FOS

The sensitivity of greenCUT&RUN for FOS was evaluated in terms of unique peaks. In total, 26 729 (66.9%) unique peaks were identified by greenCUT&RUN compared to 6552 (33%) in ChIPseq (Figure 5A). As predicted previously, 9020 and 3122 unique peaks were not significantly different between protocols. Due to the stringent definition these peaks were not called in any of the protocols (Figure 5B). By comparing coverage, higher percentage (17 709: 44.31%) of real unique peaks were identified in greenCUT&RUN compared to ChIPseq (3430: 17.25%) (Figure 5B) suggesting that greenCUT&RUN is highly sensitive.

To evaluate specificity, the distribution of AP1 motifs was explored. Similar to NFYA we examined whether peaks lacking the AP1 motif can be explained by cooperative binding events between FOS and other transcription factors. Quantitative mass spectrometry revealed known interactors of the JUN family (JUN, JUNB and JUND) and of the ATF family (ATF1 and ATF7) (Figure 5C and D). Of all peaks 35 556 (88.95%) contained the AP1 motif and were identified in greenCUT&RUN, which is significantly higher (χ^2 proportion test; P-value $<2.2 \times 10^{-16}$) compared to ChIPseq (15 506: 77.87%) (Figure 5E and Supplementary Table S4). In the remaining peaks, 865 and 519 peaks with TGANTCA sequence were identified, which represented 2.16% and 2.61% of the total. To further investigate binding events in the remaining peaks enrichment of 414 transcription factors motifs was computed, which showed enrichment of ATF7 motifs in a significant proportion (Supplementary Table S6). ATF7 motifs were observed in 802 (2%) and 603 (3.03%) peaks of greenCUT&RUN and ChIPseq, respectively. In the remaining peaks (2748 and 3278) we observed the ATF [TGANNTCA] consensus sequence in 292 (0.73%) and 343 (1.73%) peaks. In total, 14.76% of ChIPseq peaks did not contain AP1 or ATF motifs, which was significantly higher (χ^2 proportion test; P-value $<2.2 \times 10^{-16}$) compared to greenCUT&RUN (Supplementary Table S4).

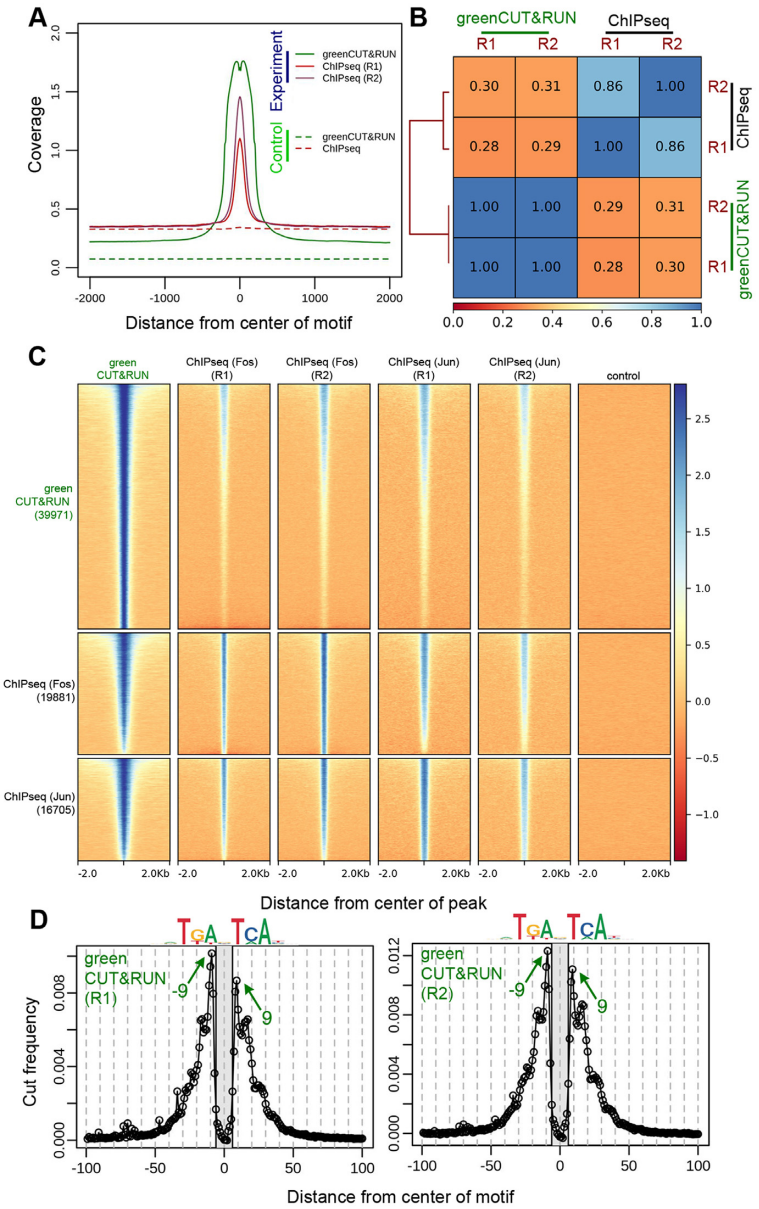

Figure 4. Performance of greenCUT&RUN for FOS. Panel (**A**): Coverage centered around the 856,556 putative AP1 binding sites. Panel (**B**): Reproducibility of biological replicates of greenCUT&RUN as determined by Pearson correlation coefficients. Panel (**C**): Signal intensity of genome-wide peaks in different protocols shown by heatmap. Panel (**D**): Normalized cut frequency of MNase around 459,169 putative AP1 binding sites. In total, 17 nucleotides are protected from MNase digestion. Gray area in the graph represents the width of motif.

Comparison of FOS binding events with histone modifications indicated that majority of the peaks were associated with histone marks of active promoter and enhancer regions (Supplementary Figure S11).

We also explored the FOS-motif, TGANTCA-sequence, ATF7-motif and TGANNTCA-sequence in the unique peaks. In comparison to ChIPseq (1233: 35.94%), these peaks obtained by the greenCUT&RUN (16 137: 91.12%) protocol had a significantly higher proportion (χ^2 test; *P*-value $<2.2 \times 10^{-16}$) of AP1 motifs. Similar to NFYA, ChIPseq for FOS displays the highest proportion of unexplained peaks (1619: 47.2%) (Figure 5E and Supplementary Table S4).

Genome-wide profiling of the basal transcription factor TBP

To complement the analysis of greenCUT&RUN with components of the basal transcription machinery for RNA polymerase II (pol II), HeLa cells expressing GFP-tagged TBP were analyzed. The DNA-library preparation protocol was modified to include shorter fragments and a nonspecific IgG-based control for (Endo)-CUT&RUN was generated. To evaluate the performance of greenCUT&RUN for TBP, the coverage was computed around 89,069 open chromatin regions, identified earlier in HeLa cells by ATAC-seq (25). In total, ∼4.0 million reads of greenCUT&RUN and of (Endo)-CUT&RUN were compared against ∼12.5 million reads of ChIPseq. We observed that both greenCUT&RUN

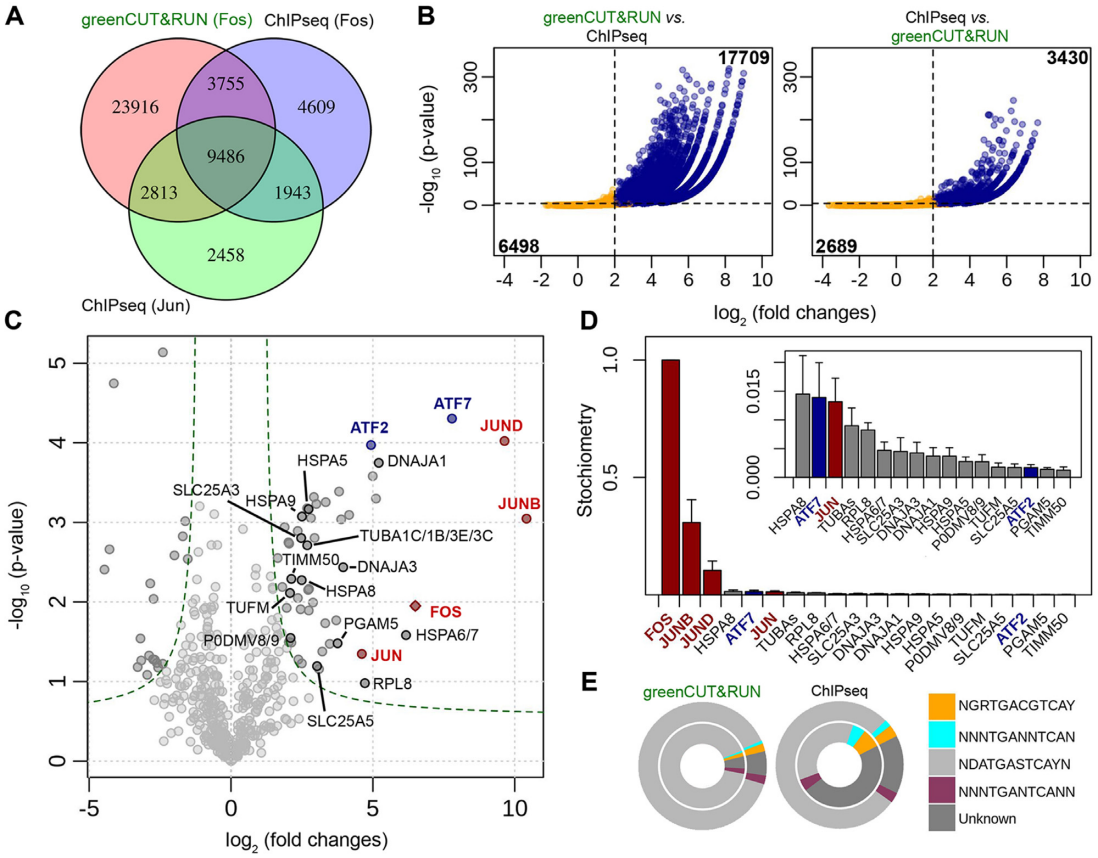

Figure 5. Sensitivity of greenCUT&RUN for FOS. Panel (**A**): A Venn diagram representing overlapping and unique peaks identified by the different protocols. Panel (**B**): Real number of unique peaks are shown in upper right part. Panel (**C**): Known and unique interactors of FOS were identified using quantitative mass spectrometry. Interestingly, ATF2/7 was identified, which binds to the TGANNTCA motif as a heterodimer with FOS. Panel (**D**) represents the stoichiometry of top 20 proteins relative to FOS. Panel (**E**): Distribution of transcription factor motifs. Outer and inner circle represents motif distribution in total and unique peaks, respectively. A large proportion of unique ChIPseq peaks were without AP1 or ATF consensus motifs.

and (Endo)-CUT&RUN performed equally well but better than ChIPseq (Figure 6A). In the controls peaks were detected in the open chromatin region both by ChIPseq and (Endo)-CUT&RUN, but not by greenCUT&RUN (Figure 6A). A high reproducibility was observed between greenCUT&RUN replicates (Figure 6B).

For TBP, 21 108, 16 165 and 16 943 peaks were identified by greenCUT&RUN, (Endo)-CUT&RUN and ChIPseq, respectively (Figure 6C). Examples of peaks unique to a specific protocol are shown as browser tracks in Supplementary Supplementary Figure S12. Analysis of heatmaps indicate that greenCUT&RUN provides the most robust occupancy of TBP compared to other protocols (Figure 6C). We observed similar coverage around transcription start sites (TSSs) for both greenCUT&RUN and (Endo)-CUT&RUN. in contrast, a lower coverage was observed by ChIPseq (Figure 6D). In the case of NFYA and FOS, the maximum number of binding events can be explained on basis of its respective DNA motifs, but this is difficult for TBP as only small fraction of TBP binding to pol II promoters can be explained by the TATA-box (43). Other factors *e.g.* H3K4me3 have been involved in the recruitment of transcriptional initiation complexes specifically for TATA-less promoters (44). It has been reported that eRNA expressing enhancers are also enriched for TBP and the basal transcription factors characteristic of cognate pol II promoters (45). Moreover, TBP can also binds to cryptic promoter sites (46). Therefore, we used multiple factors to examine TBP binding events including enhancers and promoters marked by H3K27ac, active promoters marked by H3K4me3, open chromatin regions, TATA-box and regions where active transcription occurring (POLR2A and GROseq peaks) in HeLa cells. The logo of TATA-box used is given in Supplementary Figure S13. Of the 67,008 TSSs, this TATA-box was present in 15,978 (23.85%) as reported earlier (47). Using all these factors, we found that 20 110 (95.27%), 15 324 (94.24%) and 16 786 (99.07%) peaks from greenCUT&RUN, (Endo)-CUT&RUN and ChIPseq, respectively, can be explained (Figure 6E).

Sensitivity and specificity of greenCUT&RUN: TBP

To explore the sensitivity of greenCUT&RUN to identify TBP binding events, peaks were compared to the other protocols. In total, 9486 (44.9%), 5450 (33.7%) and 8679 (51.12%) unique peaks were identified in greenCUT&RUN, (Endo)-CUT&RUN and ChIPseq, respectively (Figure 7A). Of the unique peaks observed in greenCUT&RUN versus (Endo)-CUT&RUN and greenCUT&RUN versus ChIPseq respectively, we identified 2814 (13.3% of the total)

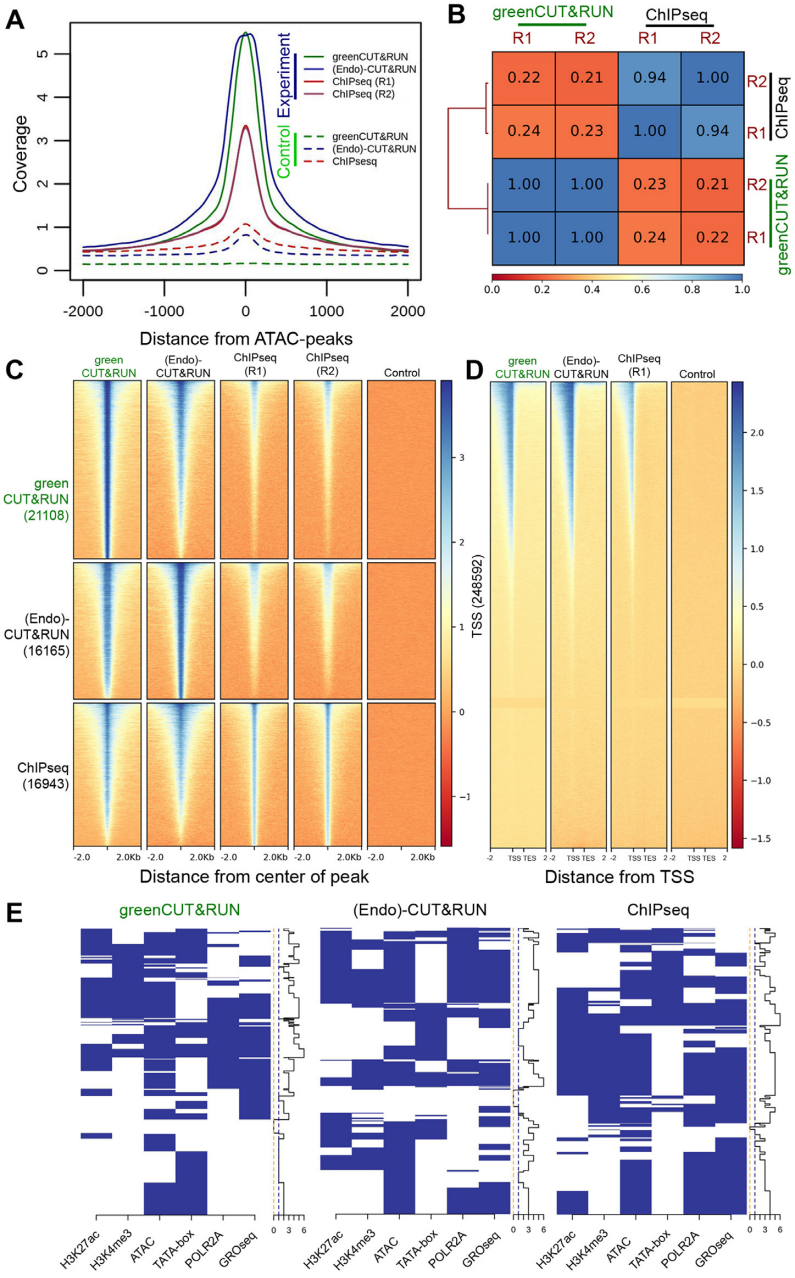

Figure 6. Performance of greenCUT&RUN for TBP. Panel (**A**): Coverage around open chromatin region (ATAC-peaks of HeLa) for each protocol. Peaks are observed around open chromatin region in IgG-based control of ChIPseq and (Endo)-CUT&RUN compared to the no Ca^{2+} control of green-CUT&RUN. Panel (**B**) shows that greenCUT&RUN data displays high reproducibility among replicates, compared to ChIPseq. Values represent Pearson correlation of coverage calculated for 10 kb bins. Panel (**C**): Signal intensity of genome-wide peaks in different protocols shown by heatmaps. In total, ∼25% more peaks were identified in greenCUT&RUN compared to ChIPseq. Replicate 2 (R2) of ChIPseq displays similar intensity compared to R1 (data not shown) and it is excluded. Panel (**D**): Metagene plot of TBP for all three techniques. Panel (**E**) displays TBP peaks overlapping with active promoters (H3K4me3) and promoters/enhancers (H3K27ac), open chromatin regions (ATAC), TATA-box and transcriptionally active regions (POLR2A and GROseq). On the right side, a staircase plot represents the number of factors overlapping with peaks.

and 10 987 (52.05%) as truly unique peaks (Figure 7B). On the other hand, only 3004 (18.58%) and 1968 (11.62%) truly unique peaks were identified in (Endo)-CUT&RUN versus greenCUT&RUN and ChIPseq versus greenCUT&RUN. This indicates that greenCUT&RUN is more sensitive for TBP than ChIPseq and similar to (Endo)-CUT&RUN. It is interesting to note that almost all ChIPseq peaks can be captured with greenCUT&RUN using a less stringent peak definition, while this is not observed for ChIPseq. This conclusion is also evident from the heatmap (Figure 6C). By analyzing the unique peaks similarly to Figure 6E, we found 2602 (92.47%), 10 003 (91.04%), 2675 (89.05%) and 1919 (97.51%) in respectively greenCUT&RUN versus (Endo)-CUT&RUN, greenCUT&RUN versus ChIPseq, Endo)-CUT&RUN versus greenCUT&RUN and ChIPseq versus greenCUT&RUN can be explained (Figure 7C).

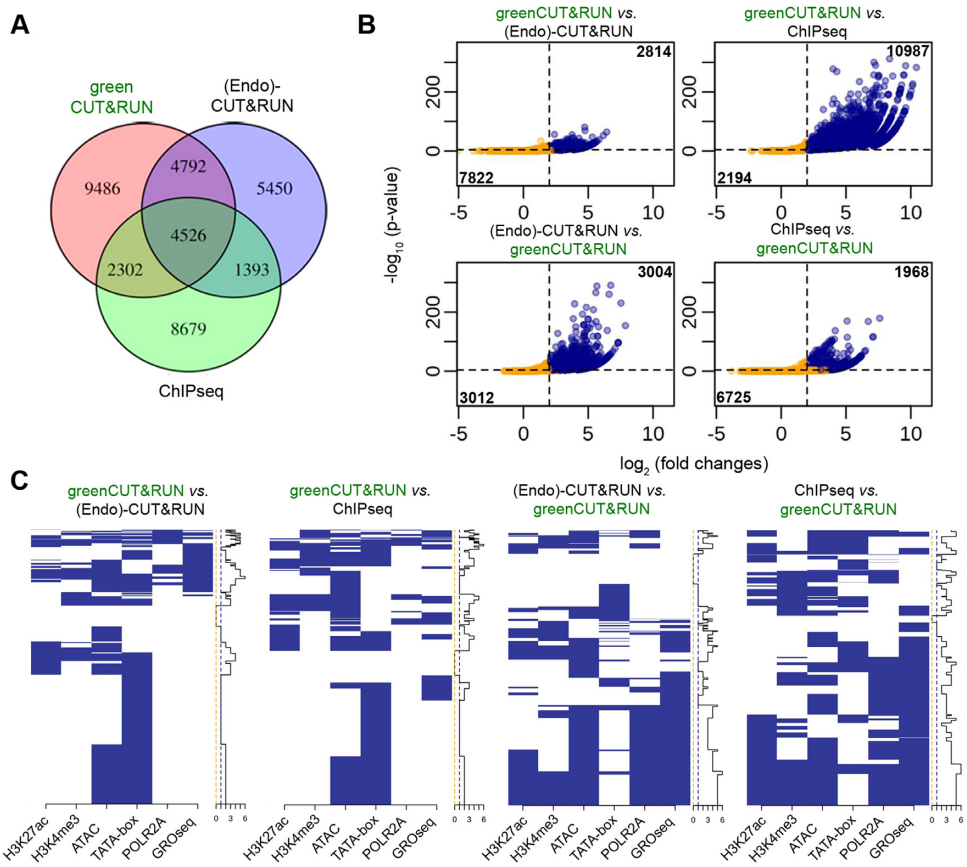

Figure 7. Sensitivity of greenCUT&RUN for TBP. Panel (**A**): A Venn diagram representing overlapping and unique peaks. Panel (**B**): The real number of unique peaks is shown. Panel (**C**): Real unique TBP peaks overlapping with active promoter (H3K4me3) and enhancers (H3K27ac), open chromatin regions (ATAC), TATA-box and transcriptionally active regions (POLR2A and GROseq). On the right side, a staircase plot represents the number of factors overlapping with peaks.

DISCUSSION

ChIPseq has been the standard technique to profile transcription factors and histone modification in a genome-wide manner. Recently, 'CUT&RUN' was developed, which does not rely on crosslinking and requires only ∼10% of the reads compared to ChIPseq (6). Both ChIPseq and CUT&RUN protocols depend on antibodies of high specificity and high affinity, which are not available for each protein in each organism. We designed 'greenCUT&RUN' to address this issue. The protocol starts by tagging the protein of interest with GFP and uses a GFP nanobody-MNase fusion protein to profile genome-wide binding events. Since the nanobody targets GFP, greenCUT&RUN eliminates the dependency on high quality antibodies. It is important to note that formaldehyde crosslinking required for ChIPseq obstructs GFP recognition by camelids nanobodies (unpublished observations). By quantitative mass spectrometry we showed that GFP-tagging did not interfere in the formation of active transcription complexes of the proteins tested.

GreenCUT&RUN reduces handling time and technical variations as steps involving binding of antibody and of protA-MNase are combined in one step. The performance of greenCUT&RUN was evaluated with three transcription factors: NFYA, FOS and TBP. In all cases the maximum number of peaks was identified with greenCUT&RUN. The accuracy of peak detection was indicated by presence of known consensus motifs. Analysis of genomic coverage indicates that greenCUT&RUN covers ∼90% of peaks obtained in other protocols, whilst this is not true with other protocols. Many techniques are biased towards chromatin accessibility and generate more reads in open chromatin regions (48). We tested this by performing greenCUT&RUN on cells lacking GFP-tagged proteins, which indicated that the nanobody-MNase has no preference for open chromatin by itself. A small proportion of the NFYA peaks determined by greenCUT&RUN lacks NFY consensus motifs. By integrating greenCUT&RUN with affinity purification mass spectrometry we identified 'piggyback' binding events of NFYA via SP1 binding to DNA. Previous studies already showed that NFY complex can physically interact with SP1 (36,37). Our experiments indicate that SP1 can bring NFYA to promoter/enhancer sequences lacking a CCAAT-box, which was confirmed by examining ENCODE ChIPseq datasets. FOS interacts with both JUN and ATF family members, which results in the binding to TGANTCA and TGANNTCA, respectively. GreenCUT&RUN identified FOS binding events on both the AP1 (TGANTCA) and ATF (TGANNTCA) con-

sensus motifs. A large proportion of unique peaks identified in ChIPseq lack the consensus motif for the transcription factor in question. These unique peaks are not replicated by other techniques, which indicate that these peaks are false-positives having originated from spurious cross-linking or antibody cross-reactivity. In both green and (Endo)-CUT&RUN a small proportion of TBP peaks only coincide with the TATA-box and they do not carry other promoter or enhancer characteristics. Interestingly, 18 of such peaks from greenCUT&RUN but not from the other datasets overlap with pol III-transcribed genes (data not shown). The other peaks in the subsets of both green and (Endo)-CUT&RUN peaks may represent cryptic pol II promoters (49–51). This would imply that both green and (Endo)-CUT&RUN are sensitive enough to capture such TBP binding events. Further experiments including TBP mutants and BTAF1 and NC2 complex genome-profiling would be required to elucidate whether these peaks indeed represent TBP binding to cryptic pol II promoters.

The unique design of greenCUT&RUN grants more flexibility to MNase and increases its DNA accessibility around genomic binding sites. In both greenCUT&RUN and (GFP)-CUT&RUN two flexible hinges are present, one between GFP and the protein of interest and other one in between the nanobody and MNase, while CUT&RUN has only a single flexible linker between protA/G and MNase. This might explain detection of the MNase hypersensitive site within the CCAAT-box by greenCUT&RUN method. The NFY complex introduces an 80° bend in the DNA, exposing center of the CCAAT-box from one side of the DNA (40). Asymmetrical MNase peaks in cut frequency around the CCAAT-box were obtained for NFYA (Figure 2C), while for FOS symmetrical peaks were observed around the TGANTCA consensus motif (Figure 4C). This is consistent with the crystallographic structures for the NFY-DNA and FOS/JUN-DNA complexes. In case of TBP no peaks in cut frequency around TSS (transcription start site)-associated TATA boxes were observed (data not shown), which may be an indication of multiple promoter binding modes of TBP. To explain the apparent protection from MNase digestion around the CCAAT and TGANTCA/TGANNTCA binding sites for NFYA and FOS, we applied a protein-DNA docking approach to probe the DNA accessibility using the crystallographic structures of NFY-DNA, FOS/JUN-DNA and MNase bound to a dinucleotide as starting point. The docking was performed with HADDOCK2.4 (32) using a single ambiguous distance restraint between the active site residue of MNase and any phosphate on the DNA. The DNA in both complexes was extended on both ends to allow for broader sampling by the MNase. Only a repulsive energy potential was used to prevent steric clashes. In the analysis of resulting 100,000 rigid-body docking models, only MNase accessible sites from the 5′-end were considered (Supplementary Figures S6 and S14). Docking of MNase unto the FOS/JUN-DNA structure revealed nuclease access to the eighth nucleotide relative to the center of binding. Bouallaga *et al.* (52) identified that 19 nucleotides including the AP1 site are protected by an AP1 transcription factor from DNase-I digestion. This study involved the JUNB/FRA2 heterodimer and the non-canonical TtAGTCA DNA motif. Both MNase docking, DNaseI footprinting and the crystallographic structures are consistent with the cut frequency results for FOS, which stress the high resolution of the greenCUT&RUN approach. In case of the NFY–DNA complex the hypersensitive site at the center of the CCAAT sequence observed was not observed by docking. This may relate to DNA bending, which is not considered in our MNase docking approach. The MNase docking results for the NFY-DNA complexes indicate a protection of 6–9 bases around the CCAAT sequence, while cut frequencies indicated a protection of 13 bases upstream and 17 bases downstream of CCAAT (Figure 2D). This may indicate that parts of the NFY complex, which are not in the crystallized form of NFY, may be shielding flanking DNA from MNase digestion. Unfortunately, it is not clear whether the MNase approaches DNA from minor or major grooves or how MNase digestion is influenced by the DNA curvatures, but it has a preference for A/T-rich DNA. Indeed, we observe this preference in all CUT&RUN-based data (Supplementary Figure S15).

Taken together, our results indicate that a very good performance, sensitivity and specificity can be achieved for genome mapping of any protein using greenCUT&RUN. The greenCUT&RUN protocol overcomes major hurdles of existing methods like ChIPseq and standard CUT&RUN, which require highly specific and sensitive antibodies against the protein of interest. GreenCUT&RUN is less time consuming and easy to process in standard molecular biology laboratories allowing broad applications. Like CUT&RUN the experimental protocol does not involve crosslinking of protein or sonication steps, which avoids common shortcomings of ChIPseq like epitope masking, preferences for open chromatin and high percentages of false-positive peaks. Similar to CUT&RUN robust and reproducible genomic profiles, which are accurate as defined by the presence of known consensus motifs, unbiased towards open chromatin regions and obtained by greenCUT&RUN at low read numbers (2–4 million). While we tested transcription factors directly binding to DNA, we expect that greenCUT&RUN would also be applicable to genomic profiling of (subunits of) large transcription and/or chromatin remodeling complexes. However, distance and accessibility to DNA of the tethered MNase moiety could affect efficient mapping in specific cases.

The GFP nanobody used in greenCUT&RUN has been also utilized in quantitative mass spectrometry, which provides an integrated platform for both genomic and proteomic profiling of target proteins. As such we identified ‘piggy-back’ DNA binding events. The greenCUT&RUN protocol provides high resolution DNA footprinting compared to other protocols, which may be due to more flexibility and accessibility of MNase around binding sites. It is important to note that greenCUT&RUN also provides a platform for the rapid analysis of transcription factor mutants. Moreover, large collections of GFP-tagged genes, both N-terminal and C-terminal, have been constructed in bacterial artificial chromosomes (53) and have been used to generate transgenic human HeLa or mouse embryonic stem cell lines (53). In conclusion, greenCUT&RUN is a widely applicable genome mapping technique with the opportunity for combination with quantitative mass spectrometry.

DATA AVAILABILITY

All NGS data have been deposited to Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) under the accession number SRP278136 (Bioproject-ID: PRJNA658159) and mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository under the dataset identifier PXD021089.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank all members of the Timmers lab for their constructive comments during this work and we are grateful to Dirk Schübeler and Luke Isbel (Friedrich Miescher Institute, Basel) for critical review of the manuscript. Parts of the sequencing were performed within the setting of the Core Facility Genomics (TVW, SL), Medical Faculty, University of Freiburg. We thank Winship Herr and Maykel Lopes for the ChIP protocol and the Deep Sequencing Facility of Max Planck Institute for Immunobiology and Epigenetics, Freiburg, Germany, for performance of ChIP quality controls, library construction and Illumina sequencing.
Author contributions: M.T. and S.K. conceived the greenCUT&RUN strategy; M.T., S.K., T.B. designed and performed the genome localization experiments; T.V.W. and S.L. contributed to the DNA sequencing experiments; M.B. performed the mass spectrometry analysis; A.B. performed the protein/DNA docking and MNase accessibility analysis; S.N., S.K. and M.T. performed the data analysis; and all the authors were involved in paper writing.

FUNDING

Deutsche Forschungsgemeinschaft [SFB850 to M.T., S.L., SFB992 to M.T., S.L., TI688/1-1 to M.T.]; Collaborative Center for X-linked Dystonia; Bonvin acknowledges financial support from the European Union Horizon 2020 project BioExcel [823830 to A.M.J.J.]. Funding for open access charge: DKTK.
Conflict of interest statement. None declared.

REFERENCES

1. Robertson,A.G., Kim,J., Al-Ahmadie,H., Bellmunt,J., Guo,G., Cherniack,A.D., Hinoue,T., Laird,P.W., Hoadley,K.A., Akbani,R. *et al.* (2017) Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, **171**, 540–556.
2. Dreijerink,K.M.A., Timmers,H.T.M. and Brown,M. (2017) Twenty years of menin: emerging opportunities for restoration of transcriptional regulation in MEN1. *Endocr. Relat. Cancer*, **24**, T135–T145.
3. Ganesan,M., Nizamuddin,S., Katkam,S.K., Kumaraswami,K., Hosad,U.K., Lobo,L.L., Kutala,V.K. and Thangaraj,K. (2016) c.*84G>A mutation in CETP is associated with coronary artery disease in south Indians. *PLoS One*, **11**, e0164151.
4. Sanborn,A.L., Rao,S.S., Huang,S.C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *PNAS*, **112**, E6456–E6465.
5. Schmid,M., Durussel,T. and Laemmli,U.K. (2004) ChIC and ChEC; genomic mapping of chromatin proteins. *Mol. Cell*, **16**, 147–157.
6. Skene,P.J. and Henikoff,S. (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, **6**, e21856.
7. Kaya-Okur,H.S., Wu,S.J., Codomo,C.A., Pledger,E.S., Bryson,T.D., Henikoff,J.G., Ahmad,K. and Henikoff,S. (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.*, **10**, 1930.
8. Hainer,S.J., Bošković,A., McCannell,K.N., Rando,O.J. and Fazzio,T.G. (2019) Profiling of pluripotency factors in single cells and early embryos. *Cell*, **177**, 1319–1329.
9. Carter,B., Ku,W.L., Kang,J.Y., Hu,G., Perrie,J., Tang,Q. and Zhao,K. (2019) Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat. Commun.*, **10**, 3747.
10. Kubala,M.H., Kovtun,O., Alexandrov,K. and Collins,B.M. (2010) Structural and thermodynamic analysis of the GFP:GFP-nanobody complex. *Protein Sci.*, **19**, 2389–2401.
11. Zentner,G.E., Kasinathan,S., Xin,B., Rohs,R. and Henikoff,S. (2015) ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nat. Commun.*, **6**, 8733.
12. van Nuland,R., van Schaik,F.M., Simonis,M., van Heesch,S., Cuppen,E., Boelens,R., Timmers,H.M. and van Ingen,H. (2013) Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain. *Epigenet. Chromatin*, **6**, 12.
13. Dignam,J.D., Martin,P.L., Shastry,B.S. and Roeder,R.G. (1983) Eukaryotic gene transcription with purified components. *Methods Enzymol.*, **101**, 582–598.
14. Spruijt,C.G., Baymaz,H.I. and Vermeulen,M. (2013) Identifying specific protein-DNA interactions using SILAC-based quantitative proteomics. *Methods Mol. Biol.*, **977**, 137–157.
15. Tyanova,S., Temu,T. and Cox,J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.*, **11**, 2301–2319.
16. Schwanhausser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
17. Carey,M.F., Peterson,C.L. and Smale,S.T. (2009) Dignam and Roeder nuclear extract preparation. *Cold Spring Harb. Protoc.*, **2009**, doi:10.1101/pdb.prot5330.
18. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
19. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
20. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.*, **22**, 1760–1774.
21. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
22. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
23. Tarasov,A., Vilella,A.J., Cuppen,E., Nijman,I.J. and Prins,P. (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, **31**, 2032–2034.
24. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
25. Oomen,M.E., Hansen,A.S., Liu,Y., Darzacq,X. and Dekker,J. (2019) CTCF sites display cell cycle-dependent dynamics in factor binding and nucleosome positioning. *Genome Res.*, **29**, 236–249.
26. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.*

(2006) The UCSC genome browser database: update 2006. *Nucleic Acids Res.*, **34**, D590–598.
27. R Development Core Team. (2013) IN: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
28. Ramirez,F., Dundar,F., Diehl,S., Gruning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.
29. Amemiya,H.M., Kundaje,A. and Boyle,A.P. (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.*, **9**, 9354.
30. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
31. Chae,M., Danko,C.G. and Kraus,W.L. (2015) groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*, **16**, 222.
32. van Zundert,G.C.P., Rodrigues,J., Trellet,M., Schmitz,C., Kastritis,P.L., Karaca,E., Melquiond,A.S.J., van Dijk,M., de Vries,S.J. and Bonvin,A. (2016) The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.*, **428**, 720–725.
33. Brunger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.S., Kuszewski,J., Nilges,M., Pannu,N.S. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D. Biol. Crystallogr.*, **54**, 905–921.
34. Kahle,J., Baake,M., Doenecke,D. and Albig,W. (2005) Subunits of the heterotrimeric transcription factor NF-Y are imported into the nucleus by distinct pathways involving importin beta and importin 13. *Mol. Cell. Biol.*, **25**, 5339–5354.
35. Ly,L.L., Yoshida,H. and Yamaguchi,M. (2013) Nuclear transcription factor Y and its roles in cellular processes related to human disease. *Am. J. Cancer Res.*, **3**, 339–346.
36. Liang,F., Schaufele,F. and Gardner,D.G. (2001) Functional interaction of NF-Y and Sp1 is required for type a natriuretic peptide receptor gene transcription. *J. Biol. Chem.*, **276**, 1516–1522.
37. Roder,K., Wolf,S.S., Larkin,K.J. and Schweizer,M. (1999) Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. *Gene*, **234**, 61–69.
38. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
39. Liu,N., Hargreaves,V.V., Zhu,Q., Kurland,J.V., Hong,J., Kim,W., Sher,F., Macias-Trevino,C., Rogers,J.M., Kurita,R. *et al.* (2018) Direct promoter repression by BCL11A controls the fetal to adult hemoglobin switch. *Cell*, **173**, 430–442.
40. Nardini,M., Gnesutta,N., Donati,G., Gatta,R., Forni,C., Fossati,A., Vonrhein,C., Moras,D., Romier,C., Bolognesi,M. *et al.* (2013) Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell*, **152**, 132–143.
41. Halazonetis,T.D., Georgopoulos,K., Greenberg,M.E. and Leder,P. (1988) c-Jun dimerizes with itself and with c-Fos, forming complexes of different DNA binding affinities. *Cell*, **55**, 917–924.
42. Hai,T. and Curran,T. (1991) Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *PNAS*, **88**, 3720–3724.
43. Shi,W. and Zhou,W. (2006) Frequency distribution of TATA Box and extension sequences on human promoters. *BMC Bioinformatics*, **7**, S2.
44. Vermeulen,M., Mulder,K.W., Denissov,S., Pijnappel,W.W., van Schaik,F.M., Varier,R.A., Baltissen,M.P., Stunnenberg,H.G., Mann,M. and Timmers,H.T. (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, **131**, 58–69.
45. Lam,M.T., Li,W., Rosenfeld,M.G. and Glass,C.K. (2014) Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.*, **39**, 170–182.
46. Gómez-Navarro,N., Jordán-Pla,A., Estruch,F. and Pérez-Ortín,J.E. (2016) Defects in the NC2 repressor affect both canonical and non-coding RNA polymerase II transcription initiation in yeast. *BMC Genomics*, **17**, 183.
47. Yang,C., Bolotin,E., Jiang,T., Sladek,F.M. and Martinez,E. (2007) Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, **389**, 52–65.
48. Auerbach,R.K., Euskirchen,G., Rozowsky,J., Lamarre-Vincent,N., Moqtaderi,Z., Lefrancois,P., Struhl,K., Gerstein,M. and Snyder,M. (2009) Mapping accessible chromatin regions using Sono-Seq. *PNAS*, **106**, 14926–14931.
49. Koster,M.J. and Timmers,H.T. (2015) Regulation of anti-sense transcription by Mot1p and NC2 via removal of TATA-binding protein (TBP) from the 3′-end of genes. *Nucleic Acids Res.*, **43**, 143–152.
50. Koster,M.J., Yildirim,A.D., Weil,P.A., Holstege,F.C. and Timmers,H.T. (2014) Suppression of intragenic transcription requires the MOT1 and NC2 regulators of TATA-binding protein. *Nucleic Acids Res.*, **42**, 4220–4229.
51. Xue,Y., Pradhan,S.K., Sun,F., Chronis,C., Tran,N., Su,T., Van,C., Vashisht,A., Wohlschlegel,J., Peterson,C.L. *et al.* (2017) Mot1, Ino80C, and NC2 Function Coordinately to Regulate Pervasive Transcription in Yeast and Mammals. *Mol. Cell*, **67**, 594–607.
52. Bouallaga,I., Teissier,S., Yaniv,M. and Thierry,F. (2003) HMG-I(Y) and the CBP/p300 coactivator are essential for human papillomavirus type 18 enhanceosome transcriptional activity. *Mol. Cell. Biol.*, **23**, 2329–2340.
53. Poser,I., Sarov,M., Hutchins,J.R., Hériché,J.K., Toyoda,Y., Pozniakovsky,A., Weigl,D., Nitzsche,A., Hegemann,B., Bird,A.W. *et al.* (2008) BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods*, **5**, 409–415.