BIOCHEMISTRY

DNB-based on-chip motif finding: A high-throughput method to profile different types of protein-DNA interactions

Zhuokun Li[1]*, Xiaojue Wang[1,2]*, Dongyang Xu[1]*, Dengwei Zhang[1,2]*, Dan Wang[1,2,3], Xuechen Dai[1,2], Qi Wang[1,2], Zhou Li[1], Ying Gu[1], Wenjie Ouyang[1], Shuchang Zhao[1,4], Baoqian Huang[1,5], Jian Gong[6], Jing Zhao[1], Ao Chen[1], Yue Shen[7,8], Yuliang Dong[1], Wenwei Zhang[1], Xun Xu[1,2,3†], Chongjun Xu[1,6,9†], Yuan Jiang[1,2,6†]

Here, we report a sensitive DocMF system that uses next-generation sequencing chips to profile protein-DNA interactions. Using DocMF, we successfully identified a variety of endonuclease recognition sites and the protospacer adjacent motif (PAM) sequences of different CRISPR systems. DocMF can simultaneously screen both 5′ and 3′ PAMs with high coverage. For SpCas9, we found noncanonical 5′-NAG-3′ (~5%) and 5′-NGA-3′ (~1.6%), in addition to its common PAMs, 5′-NGG-3′ (~89.9%). More relaxed PAM sequences of two uncharacterized Cas endonucleases, VeCas9 and BvCas12a, were extensively characterized using DocMF. Moreover, we observed that dCas9, a DNA binding protein lacking endonuclease activity, preferably bound to the previously reported 5′-NGG-3′ sequence. In summary, our studies demonstrate that DocMF is the first tool with the capacity to exhaustively assay both the binding and the cutting properties of different DNA binding proteins.

INTRODUCTION

DNA and proteins are the two most important biological macromolecules in organisms, and their interactions play crucial roles in many living cell activities, such as gene expression, DNA replication, viral infection, etc. DNA-protein interactions are necessary to translate the encoded genetic information for use by the cells. A major function of protein-DNA interactions is in the regulation of DNA architecture. These associations between DNA and protein primarily involve binding interactions. Proteins can interact with DNA in either the major or minor groove and in a sequence-specific or secondary structure-dependent manner, often inducing large structural changes in DNA. In both prokaryotes and eukaryotes, some proteins such as nucleases bind and subsequently cleave scissile phosphodiester bonds in nucleic acids, which is essential for biological processes like DNA repair or cell defense (*1*).

Current methods for studying DNA-protein interactions include, but are not limited to, electrophoretic mobility shift assays, chromatin immunoprecipitation, DNase footprinting, biolayer interferometry, biosensor surface plasmon resonance, photonoic crystal biosensors, and some luminescence detection methods for DNA binding proteins and biolayer interferometry (*2*–*7*). But these methods can be laborious, are not high throughput, and can only be used in cases where the DNA remains intact after the interaction occurs. Protein-binding microarrays (PBMs), on which proteins bind to double-stranded oligonucleotides, have been used to study transcription factor (TF) DNA binding site preferences (*3*). Although PBMs are considered high throughput, this technology is limited by the number of features that can be placed on an array. The complete catalog of 10-mers (10^6 features) is the current approximate limit for array technology. However, many DNA binding proteins such as zinc finger proteins have recognition sites longer than 10 base pairs (bp). SELEX (systematic evolution of ligands by exponential enrichment) (*4*) can achieve higher throughput in combination with next-generation sequencing (NGS) techniques. Although SELEX is a very useful in vitro technique, it is biased toward high-affinity binding motifs and does not disclose the full spectrum of binding preferences (*8*, *9*). SELEX also cannot be used for proteins with nuclease activity or in situations in which the DNA does not remain intact after interaction with proteins. The NGS Illumina chip, which contains billions of double-stranded DNA (dsDNA) features on its surface, has been used to quantitatively analyze RNA-protein interactions (*10*). Recently, Jung *et al.* (*11*) used a chip-hybridized association mapping platform (CHAMP) to study protein-DNA binding preferences.

Revolutionary NGS technologies have decreased the cost of genome sequencing, and they provide several hundred millions or billions of short DNA sequences on the surface of a flow cell for the study of protein-DNA interactions. In addition to Illumina's HiSeq, NextSeq, and NovaSeq platforms, Beijing Genomics Institute (BGI's) BGISEQ-500 and MGISEQ-2000 sequencing platforms have been extensively used in applications such as exon sequencing (*12*), single-cell RNA sequencing (*13*), small noncoding RNA analysis (*14*), and noninvasive prenatal testing (NIPT) (*15*). The technology underlying BGISEQ-500 or MGISEQ-2000 combines DNA nanoball (DNB) nanoarrays (*16*) with polymerase-based stepwise sequencing (DNBSEQ).

DNB-based on-chip motif finding (DocMF) is similar to HT (high-throughput)–SELEX or CHAMP, but unlike other methods for studying DNA protein interactions, DocMF can provide information about protein binding at high-throughput scales and in situations involving DNA strand cleavage. DocMF uses the DNBSEQ technology

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China. [3]Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen 518120, China. [4]School of Basic Medicine, Qingdao University, Qingdao 266000, China. [5]School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China. [6]Complete Genomics Inc., 2904 Orchard Pkwy, San Jose, CA 95134, USA. [7]Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics, BGI-ShenzhenShenzhen 518083, China. [8]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, 518083, China. [9]MGI, BGI-Shenzhen, Shenzhen 518083, China.
*Joint authors.
†Corresponding author. Email: yjiang@completegenomics.com (Y.J.); cxu@completegenomics.com (C.X.); xuxun@genomics.cn (X.X.)

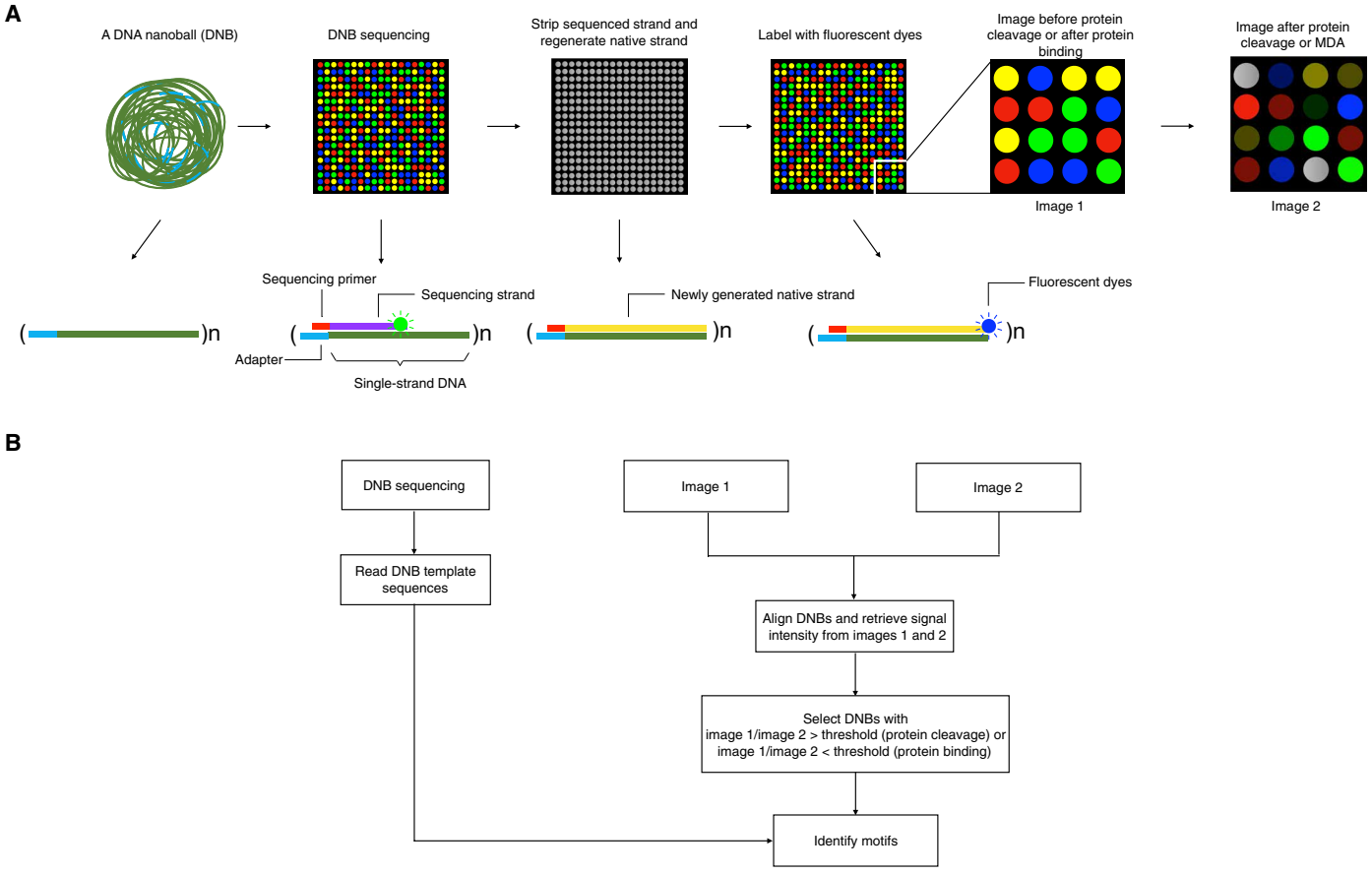

Fig. 1. DocMF overview. (**A**) Biochemistry and illustration and (**B**) bioinformatics workflow.

(*16*) and sequential imaging to detect cleavage/binding motifs involved in protein-DNA interactions (Fig. 1A). From the sequence information and the fluorescent signal change of individual DNBs, we can find all sequences that interact with the protein and identify the specific motifs via bioinformatics analysis (Fig. 1B). In this report, we successfully identified the recognition sites of different types of restriction endonucleases. We also detected the protospacer adjacent motif (PAM) sequences of SpCas9 (5′-NGG-3′, 5′-NAG-3′, and 5′-NGA-3′) (*17*) and two novel CRISPR endonucleases (VeCas9 5′-NNARR-3′ and BvCas12 5′-TYTN-3′) using a universal DNB pool for these different CRISPR-Cas systems. We also queried the DNA binding preferences of dCas9 (*18*), a mutant Cas9 lacking endonuclease activity, using a slightly modified DocMF workflow. Our identification of the dCas9 binding sites NGG agrees with previous reports (*18*). In conclusion, our platform provides a high-throughput method to interrogate a wide variety of protein-DNA interactions. The utilities of our DocMF platform can be extended to other applications, such as on-chip identification of off-target sites for a CRISPR-Cas system, single-stranded DNA cleavage sites, or TF binding motifs.

MATERIALS AND METHODS

DNB library pool

For endonuclease assays, a synthetic oligo containing 50 random bases flanked by DNBSEQTM adapter sequences was purchased from Sangon Biotech Co. Next, 2 ng of the oligo was polymerase chain reaction (PCR) amplified using nine cycles of the PCR step in the MGIEasy Universal DNA Library Prep Set (MGI Tech Co. Ltd.). The PCR product was purified by bead purification according to kit instructions and quantified using the Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific).

For PAM identification experiments and dCas9 binding experiments, the sequences of synthetic oligos used for library preparation are summarized in table S1, and all oligos were provided by the China National Gene Bank. The dsDNA used to make single-strand circles was prepared by PCR amplification using the MGIEasy Universal DNA Library Prep Set (MGI Tech Co. Ltd.). In this PCR, PAM_oligo_2-1 and PAM_oligo_2-2 were mixed and incubated at 95°C for 3 min and at 4°C for 10 min using a Bio-Rad S1000TM thermocycler. This mixture served as the PCR template, and PAM_oligo_1 and PAM_oligo_3 were used as primers. PCR was performed for 30 cycles according to the kit protocol. The PCR product was purified and quantified using the Qubit dsDNA High Sensitivity Assay Kit. Single-stranded circle (ssCir) preparation using 1 pmol of PCR product from the previous step was performed according to the circularization step in the MGIEasy Universal DNA Library Prep Set.

Nuclease-free water (Ambion) was added to 6 ng of ssCir DNA to achieve a 20-μl solution. Then, 20 μl of Make DNB buffer from the BGISEQ-500RS DNB Make Load Reagent Kit (MGI Tech Co. Ltd.) was added, and the mixture was incubated at 95°C for 1 min, 65°C

for 1 min, and 40°C for 1 min. Then, 40 μl of Make DNB enzyme mix V2.0 and 2 μl of Make DNB enzyme mix II V2.0 were added and incubated at 30°C for 20 min. The reaction was terminated by the addition of 20 μl of DNB reaction stop buffer and immediate mixing.

DocMF protocol

DNBs were loaded on the BGISEQ500 V3.1 chip by adding 30 μl of BGISEQ500 DNB loader. Single-end sequencing runs for 55 bases were performed as instructed in the BGISEQ-500RS High-throughput Sequencing Set (SE100).

After sequencing, the labeled strand synthesized in sequencing was stripped off by 100% formamide (Sigma-Aldrich), a native complementary strand was synthesized on the BGIseq500 sequencer using dNTP Mix II (MGI Tech Co. Ltd.), and dNTP Mix I (MGI Tech Co. Ltd.) was used to add the final fluorescent dNTP.

DNB-protein interactions were assessed on chip using BGISEQ500 DNB loader. For PAM identification, the first images were acquired using imaging reagent (MGI Tech Co. Ltd.) before treating DNBs with a protein of interest. For protein binding site identification, the first imaging was performed after treating DNBs with protein using the same protocol.

For PAM identification using DocMF, a second round of imaging was performed after protein-DNB interaction using the same imaging reagent. For protein binding motif identification, multiple displacement amplification (MDA) was performed after the first imaging, and then a second round of imaging was performed.

Endonuclease restriction site characterization using DocMF

In this experiment, all restriction enzymes (Eco RI, Bpu 10I, Age I, Nme AIII, Mlu I, and Bgl I) were purchased from NEB. The DNB library was prepared according to the description in the "DNB library pool" section. After native complementary strand synthesis, an image was captured. Fifty units of the selected endonuclease was then pumped into the slide and incubated for 2 hours or overnight at the manufacturer-recommended temperature for each endonuclease. The slide was washed with sequencing buffer, and the image after endonuclease digestion was captured.

PAM identification using DocMF

In this experiment, the VeCas9 and BvCas12 genes were subcloned into the pET-28a vector to include an N-terminal His6 tag. Both genes were expressed in the *Escherichia coli* BL21 (DE3) strain.

The templates for guide RNA (gRNA) transcription were prepared by PCR. PCR templates, except for spCas9, are oligos ordered form IDT. For spCas9, plasmid PX458 was used as template. Primers used in these reactions were ordered from the China National Gene Bank. The oligos used in these reactions are listed in table S1. PCR was performed using KAPAHiFi PCR HotStart Readymix (Roche) with an annealing temperature of 50°C and an extension of 15 s for 30 cycles (melting temperature of one of two primers is 47°C). The PCR products were purified using XP clean beads (Beckman Coulter) at a 2:1 ratio of beads to reaction volume. PCR products were quantified using the Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific).

For gRNA preparation, purified dsDNA from the previous step was incubated with T7 RNA polymerase overnight at 37°C using the MEGAscript T7 Transcription Kit (Thermo Fisher Scientific), and RNA was purified with the MEGAclear Transcription Clean-up Kit (Thermo Fisher Scientific).

CRISPR RNA (crRNA) was prepared by annealing two synthetic oligos (BvCas12-crRNA-F and BvCas12-crRNA-R) with complementary sequence ordered from the China National Gene Bank. Obtained dsDNA was incubated with T7 RNA polymerase overnight at 37°C using the MEGAscript T7 Transcription Kit (Thermo Fisher Scientific). crRNAs were purified using RNAXP clean beads (Beckman Coulter) at a 2:1 ratio of beads to reaction volume with an additional 1.8× supplementation of isopropanol (Sigma-Aldrich). All DNA oligo sequences used in this study are available in table S1. All RNA sequences in this study are available in table S2.

The DNB-protein reaction mix was composed of 0.1 μM protein of interest, 3 μM corresponding gRNA or crRNA, 1 μl of ribonuclease (RNase) inhibitor (Epicentre), and nuclease-free water (Ambion) to a final volume of 300 μl. For Cpf1/Cas12, the reaction buffer was NEB buffer 2.1. For Cas9, the buffer was NEB 3.1. The reaction mixture was loaded into the BGISEQ500 V3.1 chip by BGISEQ500 DNB loader and incubated for 4 hours at 37°C.

dCas9 binding protocol

In this experiment, DNB-protein interactions were assessed on chip using BGISEQ500. For the experimental lane, the DNB-protein reaction mixture was composed of 0.1 μM dCas9 *Streptococcus pyogenes* [New England Biolabs, Inc. (NEB)], 3 μM corresponding gRNA, 1 μl of RNase inhibitor (Epicentre), 1X NEB buffer 3.1 (NEB), and nuclease-free water (Ambion) to a final volume of 300 μl. Before loading into the chip, the reaction mixture was incubated at room temperature for 15 min. Then, the reaction mixture was loaded into the BGISEQ500 V3.1 chip by the BGISEQ500 DNB loader and incubated for 4 hours at 37°C. For the control lane, 1X NEB buffer 3.1 (NEB) was used. A first imaging step was performed on the BGIseq500 sequencer (MGI) using imaging reagent (MGI). After the first imaging, the MDA reaction was performed using 100 nM phi29 DNA polymerase (NEB), 1× phi29 buffer (NEB), and 400 μM dNTP (NEB) with incubation at 30°C for 30 min. After the MDA reaction, a second imaging step was performed using the same protocol as the first imaging step.

In vitro nuclease validation test

The cleavage substrate was amplified using overlap PCR (PrimeSTAR GXL DNA Polymerase from Takara) from the random PAM plasmid library (see the Supplementary Materials and Methods). A 300-bp fragment containing the spacer region was amplified by a pair of primers (M13R300 and PAMveriR, VeCas9-PAMveriR1-R10), and a 500-bp fragment containing part of the spacer region and the total specific PAM sequence was amplified using another pair of primers (M13F500 and VeCas9-PAMveriF1-18, BvCas12-PAMveriF1-14, PAMveriF); the two fragments were subjected to overlap PCR to obtain the substrate (M13F500 and M13R300). All designed oligos are shown in table S3.

VeCas9 and BvCas12 were purified as described in the Supplementary Materials and Methods. The required gRNA sequence was obtained by small RNA sequencing. The cleavage substrate (100 ng) and a final concentration of 100 nM effector (VeCas9 or BvCas12) and gRNA were used in a 20-μl reaction system. The reaction buffer system is Cas9 nuclease protein buffer (Abm). The reaction mixture was incubated at 25°C for 10 min before the addition of the cleavage substrate, and the cleavage reaction was conducted at 37°C for 1 hour. The reaction product (10 μl of total product) was detected using a 1% agarose tris-acetate-EDTA (TAE) gel running at 150 V for 30 min.

Processing DNB-based data

For each DNB, we obtained its read sequence and fluorescence intensity in image 1 and image 2 (Fig. 1). We used the fold change of fluorescence intensity (FFI) to quantify the cleaving/binding effect for each DNB

$$\mathrm{FFI} = \frac{\mathrm{F1}}{\mathrm{F2}}$$

Data analysis

Restriction endonuclease

To properly identify the exact length of restriction sites (RSs) for different restriction endonucleases, a "seed assembly" method was used. All 4- to 8-mer sequences were extracted from the 40–nucleotide (nt) positive reads. Sequences with 1.5-fold or greater enrichment relative to the reference genome were regarded as seeds. Seeds with the same length were assembled with preference for the longest sequences. The length of the consensus sequences in all longest sequences was regarded as the predicted length of the enzyme recognition site.

After obtaining the length of the enzyme's RS (L), we calculated the site rates of all possible L-mers using the following algorithm:

Suppose n is the initial total number of positive reads,

Step 1: Calculate the frequencies of all possible L-mers among positive reads via Eq. 1 and only select the L-mer with the largest frequency. We designate this L-mer with the largest frequency as M. The site rate of M is determined and equals its frequency.

$$\text{Frequency of an } L-\text{mer} = \frac{\text{No. of positive reads that contain the } L\text{–mer}}{n} \quad (1)$$

Step 2: Remove the reads that contain M from positive reads and repeat step 1 until the site rates of all possible L-mers are obtained or there are no positive reads left.

Cas9, VeCas9, and BvCas12: PAM identification and visualization

After selecting eligible reads with the correct protospacer, the 7-nt sequences at the 5′ and 3′ ends of the protospacer were extracted from these reads. Counting the total number for each unique 7-nt sequence, the relative read frequency was computed via Eqs. 2 and 3

$$K\,(\text{read frequency}) = \frac{N_2}{N_1} \quad (2)$$

$$\text{relative read frequency} = \frac{K \times 10^3}{S} \quad (3)$$

N_1 is the total number of one particular 7-nt sequence for image 1, N_2 is the total number of that 7-nt sequence for image 2, and S is the sum of whole K. By comparing the relative read frequencies at the 5′ and 3′ ends, we found that the read frequency of a partial 7-nt sequence at one end was much higher than its counterpart at the other end, indicating that the nuclease could recognize and cut these 7-nt sequences. Thus, the overall read frequency at the lower frequency end was regarded as background noise and an internal control. Because the read frequency of the control group, at the 5′ end or the 3′ end, showed a normal distribution, we applied the "three sigma rule" to define the cutoff shown in Eq. 4

$$\text{Cutoff} = \mu + 3\sigma \quad (4)$$

where μ is the mean of the distribution of control, and σ is its SD. A small portion of erroneous data could be excluded via this approach. The positive 7-nt sequences with higher read frequency were used to generate sequence logos by ggseqlogo (*19*). Similarly, Krona plots were plotted using these 7-nt sequences with their read frequencies and further modified by Adobe Illustrator to produce a PAM wheel (*20*). To more accurately define the cutting efficiency of each 7-nt sequence, Fisher's exact test (FET) was adopted as the main statistical method to prioritize the positive 7-nt sequence using the following formula (Eq. 5)

$$p = \frac{\binom{n}{m}\binom{N-n}{M-m}}{\binom{N}{M}} \quad (5)$$

N represents the total number of image 1 for each 7-nt sequence, n represents the total number of image 2, M represents the number of unique 7-nt sequences in image 1, and m represents the corresponding number in image 2. Benjamini and Hochberg (BH) correction was used to control the false discovery rate as the multitest adjustment method. All data were processed by Python and Excel.

For the frequency plot, each base frequency per site of 7-nt sequence in both images 1 and 2 was computed. Then, the base frequency of image 1 was subtracted from that of image 2 to find whether the nuclease was functional.

dCas9

To analyze the data for dCas9, we used the "relative binding strength" (RBS) shown in the following equation to evaluate the binding strength for each 7-nt sequence

$$\text{Relative binding strength} = (K_e \times S_e - K_c \times S_c) \times 10^4 \quad (6)$$

K_e is the read frequency for one particular 7-nt sequence in the experimental group, and K_c is its read frequency in the control group. S_e is the sum of all K_e in the experimental group, and S_c is the sum of all K_e in the control group. The three sigma rule was also adopted to define the cutoff because the RBS at the 5′ end approximately fits a normal distribution. With this approach, we can extract the positive 7-nt sequences with a higher RBS. Python was used to split sequencing reads and extract 7-nt sequences, and data processing and plotting were completed using Excel and R, respectively.

RESULTS

Overview and optimization of DocMF system

The novel DocMF system measures protein-DNA interactions by examining the fluorescence signal change via on-chip sequential imaging before and after protein interaction with DNBs that contain DNA targets (Fig. 1 and movie S1). The DNBs are composed of sequencing adapters and inserts of random sequences to cover the full range of protein binding sites. Hundreds of millions of DNBs are first loaded onto the BGISEQ500 chips in a patterned array, and the insert regions are sequenced at a fixed length using the DNBSeq workflow (*16*). After obtaining the unique sequence information for each DNB, we reform single-stranded DNBs (ssDNBs) by stripping off the dye-labeled strand synthesized in sequencing. Subsequently, a native complementary strand is resynthesized to form dsDNA and end-labeled with fluorescent dyes. For DNA-cleaving proteins, a first image is acquired to record the location and the signal intensity of

individual DNBs, i.e., reads. The protein of interest binds to its dsDNA targets and cleaves corresponding DNBs, leading to signal reduction or elimination of these DNBs during a second round of imaging (Fig. 1A). Specific motifs can be identified from the sequences of selected DNBs with signal elimination or reduction greater than a threshold (Fig. 1B) and verified in subsequent molecular assays. A slightly modified DocMF workflow is used to characterize protein-DNA binding preferences. In this protocol, DNBs are first imaged after incubating end-labeled dsDNA with DNA binding proteins for initial signal intensity. In the following step, an additional polymerase reaction called MDA is performed to replace the labeled strand. MDA leads to signal loss during the second imaging step as illustrated in fig. S1. However, if a protein of interest binds to its DNA targets and inhibits MDA, the signal from the DNBs containing protein binding sites would remain unchanged or be less affected than that of the control lane, which does not include protein incubation but retains the other steps.

To ensure sequential imaging is feasible, it is crucial that the stripping step does not affect spatial information or damage the DNB structure. The BGISEQ500 chips used in this experiment are patterned arrays. Therefore, the sequential imaging does not affect the registration of DNB locations to the same extent that the CHAMP method using Miseq chips is affected (*11*). In addition, we tested a variety of stripping buffers and found that the formamide buffer had the least impact on DNB integrity, only breaking the hydrogen bonds between dsDNA without affecting DNB stability or detaching DNBs from the surface. Figure S2 shows that the sequencing quality scores, including Q30 (92.83 versus 90.32), Lag (0.15 versus 0.15), and RunOn (0.15 versus 0.12), remained unaffected after stripping with formamide buffer. In comparison, the stripping buffer with NaOH significantly decreased the Q30 from more than 90% to nearly 0.

After obtaining two images before and after protein-DNA interaction, we directly compared the raw signal intensity fold change of each DNB. If the protein cleaves DNA, the DNBs that have significant signal reduction can be retrieved (Fig. 1B), and the corresponding sequences are analyzed for motif identification (Materials and Methods). To measure protein-DNA binding interactions, we obtained the binding sequence information from these DNBs with minimal signal fold change compared with the control.

DocMF can characterize a broad range of endonuclease restriction sites (RSs)

After the system was established, we tested six restriction endonucleases (Eco RI, Bpu 10I, Age I, Nme AIII, Mlu I, and Bgl I) with different RS features. The type II restriction enzymes cleave DNA adjacent to or within their recognition sites (*21*), which have been extensively studied. The selected enzymes contain restriction sites (RSs) ranging from 6 to 11 bp and comprising normal palindromic sequences, nonpalindromic sequences, and degenerate bases. The DNB library contains a pool of synthetic random DNA fragments with a length of 50 nt. Forty of the 50 nucleotides of these random sequences were read using the BGISEQ-500RS High-throughput Sequencing Set (SE100). Images were taken before and after on-chip incubation of these enzymes for 2 hours or overnight. DNBs with FFI > 2 threshold (positive reads) were identified to screen for motifs (see Materials and Methods).

The exact length of restriction sites (RSs) (*L*) was obtained via a "seed assembly" method (Materials and Methods). We then calculated the site rates of all *L*-mers among positive reads (Materials and Methods). Using this method, for each of these six restriction endonucleases, we obtained the site rates of all *L*-mers (*L* is the predicted length of an RS) and drew a boxplot for the *L*-mers with the top 50 largest site rates (Fig. 2A). The motifs (colored orange in Fig. 2A) corresponding to the outliers in the boxplot, with the sum of site frequency of these outliers (colored green in Fig. 2A) >80%, were regarded as the DocMF-predicted restriction sites (RSs). Thus, for Eco RI, Bpu 10I, Age I, Nme AIII, and Mlu I, we obtained their restriction sites (RSs) from the outliers shown in Fig. 2A, because the site rate sums of these outliers were all larger than 80%. For Bgl I,

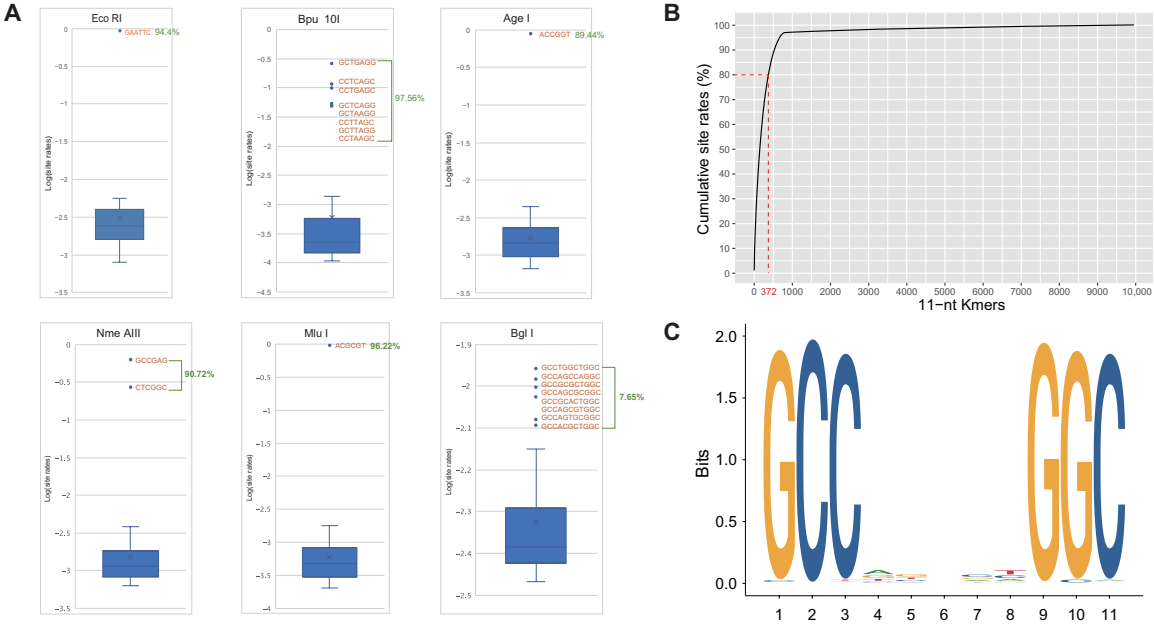

Fig. 2. Restriction endonuclease cut site identification using DocMF. (**A**) Box plots for the motifs with the top 50 $\log_{10}$(site rates). Outliers' DNA sequences (orange) and the sum of outliers' site rates (green) are shown. (**B**) Cumulative site rates for Bgl I. (**C**) A sequence logo representation of the 372 motifs for Bgl I.

however, the site rate sum of the outliers was only 7.65%, which is too small to be predicted as restriction sites (RSs) using DocMF. Thus, for Bgl I, we used the 11-mers (because 11 is the predicted length of the RS) with the top 372 largest site rates to predict the restriction sites (RSs); the site rate sum of these 372 motifs was larger than 80% (Fig. 2B). A sequence (*19*) representation (Fig. 2C) of these 372 sequences revealed that the RS for Bgl I was GCCNNNNNGGC, which agreed with the reported RS. These results demonstrated that our system coupled with an optimized bioinformatics method could reliably identify the DNA recognition site for different types of restriction enzymes.

DocMF can accurately identify the 5′-NGG-3′ PAM of SpCas9

CRISPR-Cas effectors are RNA-guided endonucleases that use a PAM as a DNA binding signal. The PAM is a short DNA sequence, normally less than 7 bp, that sits near the target DNA (termed protospacer) of the CRISPR-Cas system (*22*, *23*). One widely used PAM identification approach transforms plasmids carrying randomized PAM sequences into *E. coli* in the presence or absence of the CRISPR-Cas locus. The frequency of a functional PAM sequence is significantly lower when the Cas protein is present (*24*).Thus, this PAM depletion assay (*24*) requires two sets of libraries for either 5′ or 3′ PAM identification and corresponding negative controls. The library size also needs to be large to cover most, if not all, PAM sequences. In addition, the plasmid depletion assay (*24*) is time-consuming and low throughput. In contrast, the DocMF system can simultaneously screen both 5′ and 3′ sequences for PAMs in a single experiment, generating coverage that is multiple orders of magnitude greater than the traditional method. One of the two PAM regions that is not recognized by the protein is used as an internal negative control.

In a proof-of-concept study, we evaluated the accuracy of DocMF by assessing the PAM requirements of SpCas9, the most widely used CRISPR-Cas system, from *S. pyogenes*. SpCas9 cleaves the dsDNA after binding to corresponding RNA, and this cleavage is reported from PAM depletion assays to be dependent on a 5′-NGG-3′ PAM sequence (*25*). The PAM DNB library used in DocMF is shown in Fig. 3A. The synthetic oligo region contained a known 23-nt SpCas9 protospacer sequence (colored orange in Fig. 3A) flanked by 5′ and 3′ PAM regions with 15 random nucleotides each (colored green in Fig. 3A). The sequence information of both PAM regions was obtained by a single-end sequencing of 50 nt.

The signal fold change was compared before and after SpCas9 on-chip incubation for 4 hours. Of 494,866,059 reads (DNBs), we obtained 366,913 DNBs that exhibited a fold change greater than 3 and could potentially be cleaved by SpCas9. The 7-nt sequences at both 5′ and 3′ PAM regions were retrieved from these DNBs for further analysis. The frequency of all 16,384 (4^7) PAM combinations for both 5′ and 3′ PAMs was calculated and plotted against the individual sequence in Fig. 3B. SpCas9 endonuclease was reported to only bind to the 3′ end of the target sequence. Therefore, 5′-randomized 7-nt sequences were used as the internal negative control. We applied the three sigma rule (*26*) to the 5′ sequences to define the cutoff for positive PAM signals. In other words, at the cutoff of 0.11, approximately 99.7% of data from the 5′ PAM region fell into background noise. This statistical cutoff resulted in 944 3′ PAM sequences that were preferably cut by SpCas9. A sequence logo (*19*) representation of the sequences revealed that SpCas9 preferred a 5′-NGG-3′ motif, although approximately 5.8% (55 of 944) of 5′-NAG-3′ and 1.6% (15 of 944) of 5′-NGA-3′ could also be recognized (Fig. 3C), which is in line with previous findings (*27*–*29*).

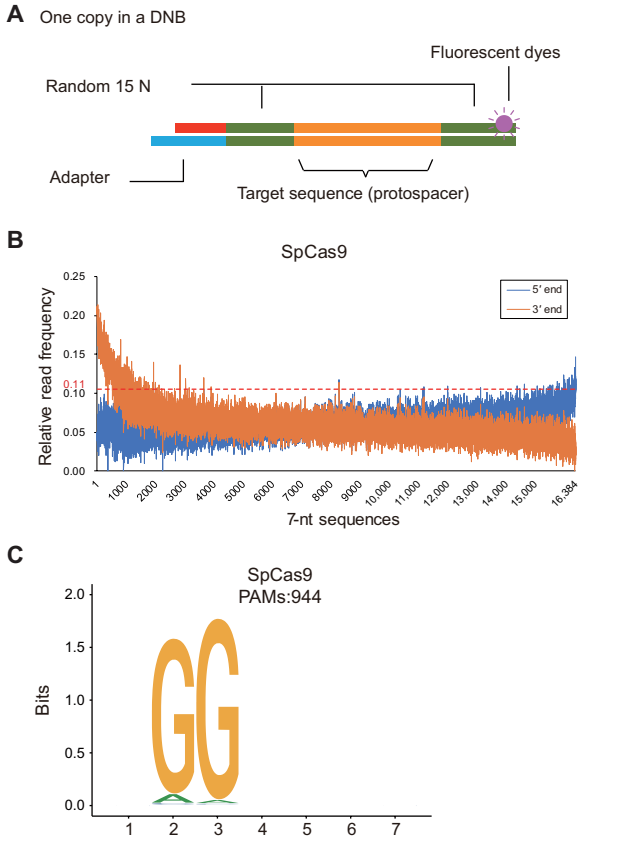

Fig. 3. PAM identification for SpCas9 using DocMF. (**A**) PAM DNB library preparation illustration. The synthetic oligo region contains a known 25-nt SpCas9 protospacer sequence (orange) flanked by 5′ and 3′ PAM regions with 15 random nucleotides each (green). Hundreds of copies of each random PAM-flanked protospacer are incorporated per DNB, and only copy is demonstrated. (**B**) The relative read frequency at both the 5′ end and the 3′ end for SpCas9. The *X* axis is all combinations of 7-nt sequences sorted by the difference between two ends in descending order. (**C**) PAM sequence for SpCas9.

DocMF enables sensitive in vitro detection of PAMs in different CRISPR-Cas systems

To further demonstrate the utility of DocMF in finding PAM sequences, we extended our study to two previously uncharacterized CRISPR-Cas systems, VeCas9 from *Veillonella* genus and BvCas12 from *Butyricimonas virosa* (*24*). VeCas9 has a Cas9 effector protein of 1064 amino acids, and BvCas12 protein is 1245 amino acids in length. Both proteins were expressed and purified as described in the Supplementary Materials and Methods. The crRNA and tracrRNA of VeCas9 were identified through small RNA sequencing, whereas the crRNA of BvCas12 was predicted in silico based on the previously reported Cas12a/Cpf1 orthologs (fig. S3). To interrogate the diversity of their PAM sequences, we conducted DocMF on VeCas9 and BvCas12 using the same DNB PAM library (Fig. 3A) used in the SpCas9 study. For VeCas9 experiments, we included three individual gRNA designs, crRNA:tracrRNA, sgRNA-1 with SpCas9 structure, and a truncated sgRNA-2 (fig. S3).

Before using DocMF, a PAM depletion assay (*24*) was first performed on VeCas9 for methodology comparison (Supplementary Materials and Methods). As shown in fig. S4 (A and B), with 4.62 Gb of

sequencing data, we observed 508 sequences with a threshold of 3 for *Alicyclobacillus acidoterrestris* C2c1 (AacC2c1), a positive control in the depletion assay, and correctly identified the reported PAM, 5′-TTN-3′ (*24*). However, with even more sequencing data (7.33 Gb) for VeCas9, 0 and 74 distinct sequences were found with thresholds of 3 and 0.8, respectively (fig. S4, C and D). The results for VeCas9 were quite similar to our negative control sets (data not shown), and thus, we failed to detect correct PAM sequences for VeCas9 using the traditional depletion method. The failure could be attributed to either weak VeCas9 protein expression or function in *E. coli* cells. In addition, the low sensitivity (~20× coverage for each 7-nt PAM sequence) of the *E. coli* depletion assay could only aggravate the problems.

Using DocMF, DNBs with signal fold change above threshold were selected for further analysis. In the read frequency plot (Fig. 4, A and B), the 5′ PAM region of VeCas9 (with sgRNA-1) and the 3′ PAM region of BvCas12 showed no protein binding pattern, and their corresponding 3 SDs (0.09 for VeCas9 and 0.075 for BvCas12) were used to set cutoff lines. As a result, 4947 and 5580 unique PAM sequences were determined to be cleaved by VeCas9 and BvCas12, respectively. Both CRISPR-Cas systems conveyed large PAM families as illustrated in consensus sequences and sequence logo, two common PAM reporting schemes (Fig. 4, C to E and G) (*22*, *23*). The consensus sequences of VeCas9 were revealed as 5′-NNARRNN-3′, or NYARRMY for an even more dominant set of PAM sequences by frequency plot (Fig. 4C), while sequence logo reported 5′-NNNRR-3′ PAM sequences (Fig. 4E). VeCas9 with the other gRNAs showed a similar pattern (fig. S5). Over 99% of PAMs with the short sgRNA-2 were found with at least one of the other two RNAs, indicating the high reproducibility of this DocMF method. Slight difference in PAM of BvCas12 was also observed between PAM reporting methods. Consensus sequence and sequence logo reported 5′-TYTN-3′ (Fig. 4D) or YYN (Fig. 4G), respectively. However, these two reporting systems ignored the correlation among all seven positions and might introduce some incorrect active PAMs if randomly combining each position.

To interrogate the relative activity of each PAM, two methods were applied, FET and the PAM wheel. FET was introduced to sort the PAM sequences. FET is a widely used test to determine whether the difference between two groups is significant. Therefore, one particular PAM with a smaller *P* value according to FET indicated that its relative read frequency, or cutting efficiency, was more significant

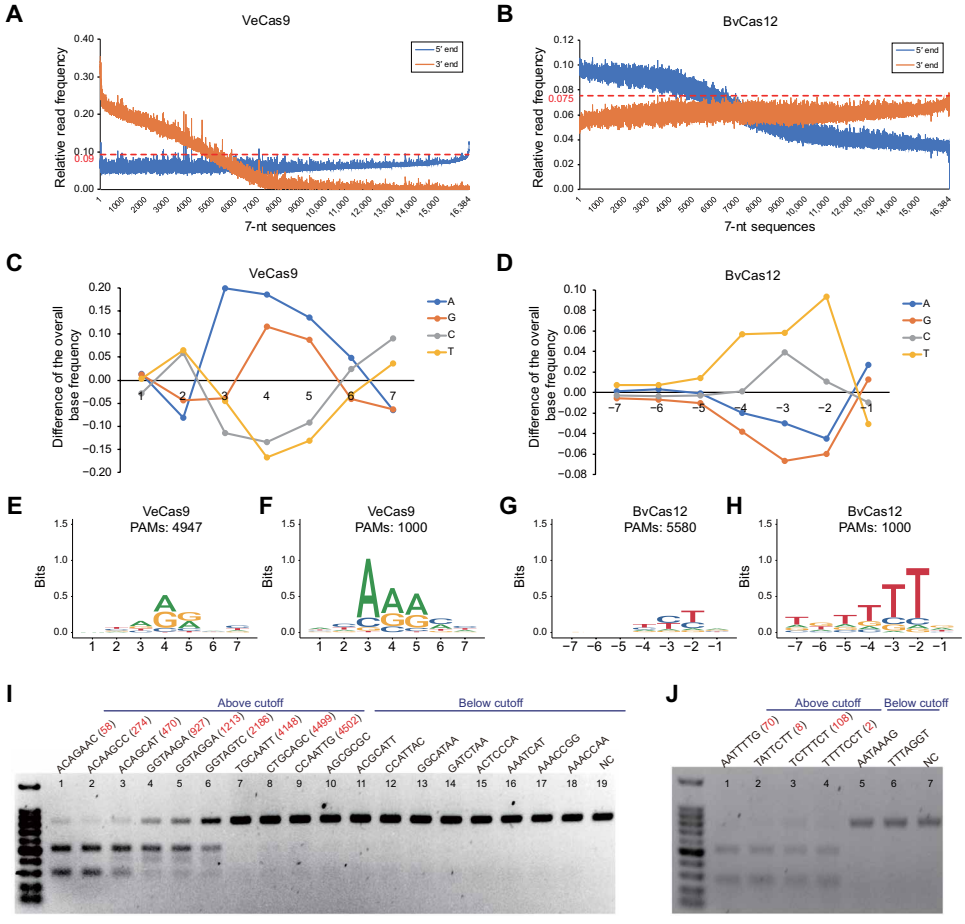

Fig. 4. PAM identification in novel CRISPR-Cas systems using DocMF. (**A**) The relative read frequency at both the 5′ end and the 3′ end for VeCas9. (**B**) The relative read frequency at both the 5′ end and the 3′ end for BvCas12. Consensus PAM sequence by frequency plot with all detected 7-nt sequences for VeCas9 (**C**) and BvCas12 (**D**). PAM sequence by sequence logo for VeCas9 generated by all detected 7-nt sequences (**E**) and by the top 1000 7-nt sequences from FET analysis (**F**). PAM sequence by sequence logo for BvCas12 generated by all detected 7-nt sequences (**G**) and by the top 1000 7-nt sequences from FET analysis (**H**). (**I**) In vitro validation of VeCas9 PAM sequences. Nine 7-nt sequences each above/below the cutoff were selected. The FET ranking numbers are shown in red. NC, negative control. (**J**) In vitro validation of BvCas12 PAM sequences. Five 7-nt sequences above the cutoff and two 7-nt sequences below the cutoff were selected.

compared with one with a larger *P* value. After ranking the PAMs in order, we examined the consensus PAM sequences for the top 1000 sequences (Fig. 4, F and H). A slightly distinct PAM consensus sequence, 5′-NNARR-3′ for VeCas9 and 5′-TTTN-3 for BvCas12, was observed under these stringent selection criteria, which correlated better with the frequency plotting results (Fig. 4, C and D). To further validate the FET prediction, an in vitro nuclease assay was performed with randomly selected PAM sequences. PCR products containing individual PAMs and a common protospacer were incubated with either Cas9 or Cas12/Cpf1 proteins at 37°C for 1 hour. Reactions with 50 ng of input were run on TAE gels, and the remaining input quantity was used to calculate cleavage efficiency. As demonstrated in Fig. 4 (G and H), the PAMs with higher FET ranking numbers (in red) had less input remaining, indicating better cutting efficiency. The least ranked PAM gave minimal cutting, the product of which was almost not visible on agarose gels whose sensitivity is several orders of magnitudes lower than NGS. The consistency suggested that we could use our FET prediction to select the most active PAMs for in vivo gene editing.

A PAM wheel was also used to comprehensively understand the PAM sequences and their base dependence. The PAM wheel, derived from interactive Krona plots, captures individual PAM sequence and their relative activity, including the ones with low enrichment (*20*). It can be also expanded at any sector of the wheel to better view a subset of sequences and study the function of those PAMs. Figure 5 (A and B) depicts the respective PAM wheels for VeCas9 and BvCas12. For VeCas9, there is a strong base dependence between position 3 and 4. If position 3 had a base R (A or G), position 4 tended to have R (>80%; fig. S6) and a small but notable level of C (>0%). If position 3 is Y (T or C), position 4 favored R only (~99%). T is the least favored base at position 4 or 5, which agrees with the in vitro cutting results from Fig. 5C (lanes 9 and 10). The gel also demonstrated that NYARRMY (the consensus PAM based on the most dominant consensus sequence; lane 1 in Fig. 4C), NNARRNN (lanes 2 to 4), and ACAAGCC (58th ranked sequence as positive control; lane 11) were cut more efficiently than NNCRRNN (lane 5), NNTRRNN (lane 6), or NNGRRNN (lane 7), which explains why A comprised 47% at position 3, while the other three bases were each between 16 and 19% (Fig. 5A and fig. S6). For the BvCas12 PAM wheel shown in Fig. 5B, we found that position −4 was random when both positions −2 and −3 were Y (T/C). So PAM YYN generated more cutting products than YRN in Fig. 5D (lanes 1 and 2). However, position −4 tended to be T if one of the −2 or −3 positions was not Y. As a result, we observed slightly more cutting with T than V at position −4, when position −2 and −3 are either RY or YR (Fig. 5D; lanes 7 to 10). R at position −2 also dictated that position −3 will be Y (100%; fig. S6). As shown in Fig. 5D, the BvCas12 system demonstrated clear cutting on 5′-TTTN-3′ or TYTN (Fig. 5D). Our data suggested that both VeCas9 and BvCas12 had a set of relaxed PAM sequences that were comprehensively captured by DocMF.

DocMF can accurately identify protein binding sites

Protein-DNA interactions have been characterized in many high-throughput platforms including microarrays, HT-SELEX, and CHAMP (*4*, *11*). We modified the DocMF workflow mentioned above to detect protein-DNA binding motifs. The steps remained unchanged until the natural complementary strand was resynthesized to form 50-bp dsDNA and end labeled with fluorescent dyes. After binding the protein of interest to its dsDNA targets and washing off any excess, we acquired the first images to record signal intensity. After this first imaging, an on-chip incubation with MDA reaction buffer, dNTPs, and a polymerase with strong strand displacement was performed

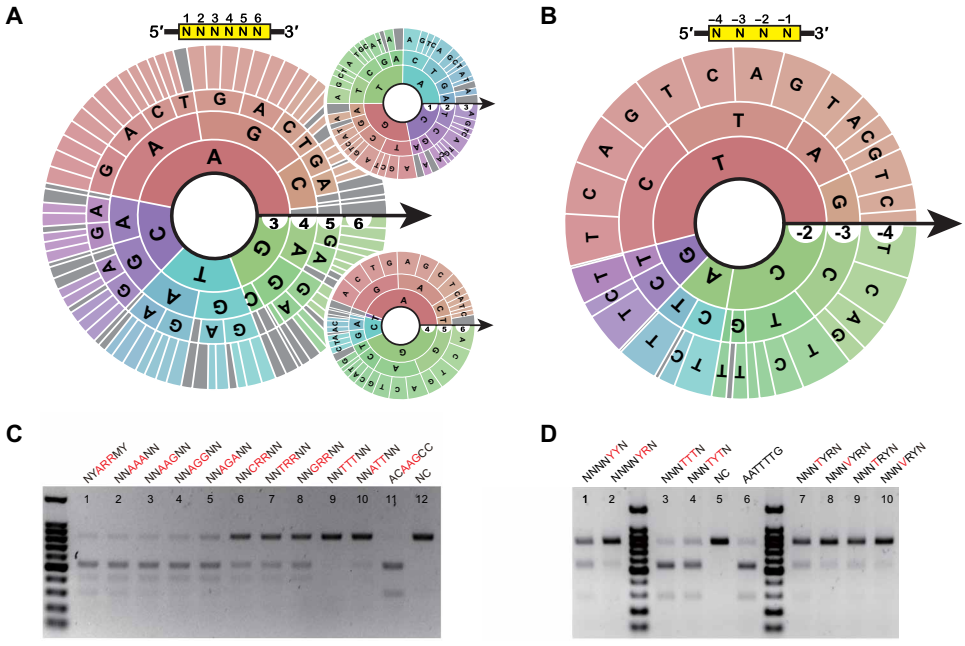

Fig. 5. PAM wheel results. (**A**) PAM wheel for VeCas9. The upper yellow box gives an indication about each position of the PAM sequence, and the arrow illustrates the orientation of each base. The area of a sector of the ring for one base at one particular position represents its frequency at this position. (**B**) PAM wheel for BvCas12. (**C**) In vitro validation of VeCas9 PAM wheel results. NYARRMY is the consensus sequence based on positional frequency, while ACAAGCC is 58th FET ranked sequence included as a positive control. (**D**) In vitro validation of BvCas12 PAM wheel results. NNNTTTN or NNNTYTN is the consensus sequence based on positional frequency, while AATTTTG is 70th FET ranked sequence included as positive control. NC (negative control): positive PAMs incubated with corresponding protein but without any sgRNA.

at 30°C for 30 min to synthesize a second complementary strand using the ssDNB as template (fig. S1). Consequently, the original fluorescent strand would be replaced and displaced from DNBs, leading to signal drop when there was no protein binding to prevent MDA. To test this idea, we used a well-studied protein, dCas9, and removed its endonuclease activity through point mutations in its endonuclease domains HNH and RucV (*17*). The point mutations D10A and H840A changed two important residues for endonuclease activity, which ultimately results in its deactivation. Although dCas9 lacks endonuclease activity, it remains capable of binding to its gRNA and the DNA strand that is being targeted because the binding is mediated through its REC1, BH, and PI domains (*30*). Moreover, dCas9 has previously been shown not to bind to its target sequence when there is no PAM (NGG) present (*31*). Unlike the studies above, the reads with fluorescence fold change below the threshold were considered positive, indicating the DNBs that dCas9 could recognize and bind. In addition, we included a negative control lane without dCas9 incubation in the same process, since BGISEQ-500 has two lanes on a single chip. Approximately 95% of a total of 253M reads from the negative control lane lost half of the signal intensity (data not shown). We chose a signal change at 0.5 as threshold (image 2/image 1) and retrieved 14,371,289 of 335,497,075 and 15,647,574 of 337,529,837 reads from experimental and control lanes, respectively. We introduced a reliable relative binding strength concept to evaluate the binding strength for each 7-nt sequence. Similarly, the data at the 5′ end fitting a normal distribution were regarded as background noise, and the three sigma rule was adopted to define the cutoff at 0.135. After deducting the noises, we observed the NGG sequence was essential for dCas9's binding (Fig. 6), consistent with previous findings. This suggests that with the modified DocMF, workflows can be harnessed as a general tool for identifying DNA binding motifs.

DISCUSSION

Similar to CHAMP (*11*), DocMF uses NGS chips to decipher protein-DNA interactions in a high-throughput manner. However, the two systems differ in many aspects. First, CHAMP needs to tag proteins with epitopes and uses fluorescent antibodies against the epitope to label the proteins on chip. In contrast, DocMF directly incubates proteins with dye-labeled target DNAs, enabling a simpler protocol and a cleaner result without the concern of nonspecific noise from surface immunostaining. Second, CHAMP uses a random clustering chip and therefore needs a fluorescent alignment marker for cluster localization information. In contrast, DocMF is performed on chips with patterned arrays followed by a straightforward sequential imaging workflow to provide protein-DNA binding information. This approach significantly simplifies and expedites the on-chip biochemistry and the downstream bioinformatic analysis. Third, like many other high-throughput technologies such as PBMs and SELEX (*4*), CHAMP can only examine protein-DNA binding intensity. CHAMP researchers include ATP inhibitors to prevent Cas3 from digesting DNA clusters when assaying for Cas3 recruitment to the DNA-bound cascade complex (*11*). However, DocMF is designed to explore various types of protein-DNA interactions, including both binding and cutting.

In this study, one important application of DocMF was to quickly and accurately identify the PAMs of any CRISPR-Cas system (Figs. 3 and 4). The CRISPR-Cas system was initially identified in the prokaryotic immune system and was quickly adopted as a reliable genome engineering tool (*25*, *32*). The PAM sequence is adjacent to the target site and is essential for Cas endonuclease specificity. Different Cas proteins bind to a variety of PAM sequences and exhibit different off-target rates of cleavage (*24*). To increase the number of potential genome editing sites, we are in urgent need of new Cas proteins that recognize PAM sequences beyond the commonly used 5′-NGG-3′ site for SpCas9 (*17*). DocMF could be a very useful tool for characterizing the novel Cas protein PAM sequences with the following advantages: (i) a universal system for different CRISPR-Cas systems, with the same DNB pools containing a common 25-nt protospacer flanked by two 15-bp randomized sequences (5′ and 3′ regions), as demonstrated in this study for SpCas9, VeCas9, and BvCas12 proteins; (ii) high sensitivity, offering an average of 20,000× to 30,000× per unique sequence for a 7-bp PAM (400 to 500 M reads for a total of 16,384 sequences) from a single BGISEQ-500 lane, compared to approximately 20× coverage from the *E. coli* completion assay; (iii) inclusion of an internal negative control, i.e., the 15-mer that is not bound by Cas proteins, to better define true positives; and (iv) high accuracy with validated cutting efficiency by in vitro assays. Using the DocMF system, we found that VeCas9, a new ortholog of Cas9 found in *Veillonella* sp., recognizes 5′-NNNRR-3′, especially 5′-NNARR-3′. The diverse PAM sequences of VeCas9 could potentially be advantageous in gene editing, especially when no suitable SpCas9 PAM is present. In addition, with a size of 1064 amino acid residues, VeCas9 can be easily packaged into adeno-associated virus (AAV), assisting better AAV delivery. The same DocMF system was applied to understand a Cas12/Cpf1 ortholog, BvCas12. The PAMs of BvCas12 are T-rich sequences, but they are on the 5′ side of the common 25-nt target sequence. Cas12a (previously named as Cpf1) has recently emerged as another powerful tool for gene editing with features distinct from Cas9, such as a requirement for T-rich 5′ PAMs,

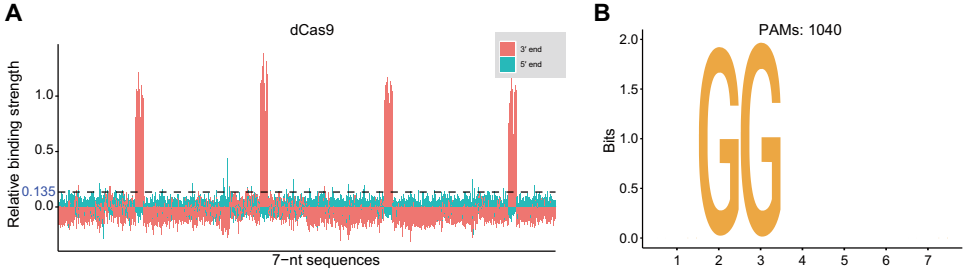

Fig. 6. DocMF used to identify protein binding site of dCas9. (**A**) The relative binding strength at both the 5′ end and the 3′ end for dCas9. The *X* axis is all combinations of 7-nt sequences and is automatically sorted by letter using Excel. (**B**) Sequence logo for dCas9 was generated by all detected 7-nt sequences based on those with the highest relative binding strength.

a single gRNA, and the production of a staggered DNA double-stranded break (*32*). Our characterization of VeCas9 and BvCas12 helps to expand the existing CRISPR toolbox and provide more candidates for genome engineering. Moreover, diverse PAM sequences with a full spectrum of cutting efficiency can be obtained from DocMF. This sensitive sorting from our FET analysis can not only help researchers to identify the strongest cutting sites but also predict potential off targets for their in vivo experiments.

Next, we performed a proof-of-concept study by assaying protein-DNA binding affinity using DocMF. dCas9, like many TFs, binds to DNA in a sequence-specific manner. In the modified DocMF workflow, an enzymatic reaction called MDA is added between the two imaging steps. MDA displaces the dye-labeled strand and therefore causes signal loss. However, if there is any dCas9 associated with the fluorescent strand, MDA mediated by phi29 polymerase will stop at the protein-DNA binding sites, leading to no or minimal signal change. In this experiment, we also ran a negative control experiment without dCas9 incubation on a separate lane. After removing the false-positive sequences, we found that dCas9 exclusively bound to a motif of NGG. Previous studies suggest that the noncanonical binding is generally not kinetically and thermodynamically favored, but cleavage-dependent conformational change can lower the energy barrier and subsequently make noncanonical cleavage more thermodynamically favored (*33*–*36*). That could explain why we only observed NAG/NGA PAM sequences in SpCas9, but not in dCas9.

In the nucleus, TFs and their cofactors normally form multiprotein complexes and bind to chromatin through DNA binding motifs to mediate gene expression. The ease of adding other proteins through the fluidic system of the NGS sequencers also provides the potential to use DocMF to examine DNA interactions with protein complex. The only caveat of this protocol is that if the protein-DNA binding affinity is too weak to block MDA, we must first cross-link the protein and DNA using a simple formaldehyde fixation before MDA. The dissociation constant (K_d) between dCas9 and the entire DNA substrate is estimated to be 2.70 nM (*37*). Therefore, TF-DNA interactions with KDs in the nanomolar and picomolar range can be directly assayed with the DocMF protocols provided in this study.

In summary, DocMF, to our knowledge, is the first high-throughput platform that can characterize motifs for both DNA binding and cleaving proteins. DocMF offers high levels of accuracy and sensitivity in motif identification. In addition to the restriction site identification, PAM characterization, and DNA binding motif examination demonstrated in this study, the utilities of our DocMF platform can also be extended to predict other aspects of protein-DNA interactions, such as on-chip identification of off-target sites for any CRISPR-Cas system, single-stranded DNA cleavage sites, or binding motifs of protein complex–like TFs. In addition, it is feasible to use DocMF to identify structure-dependent interactions, if the DNB can preserve the secondary structure of DNA targets. This workflow can also be easily adopted by Illumina's patterned flow cell, making it accessible to non-BGISEQ users. We believe that DocMF can provide valuable information for researchers from distinct communities.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/31/eabb3350/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. W. Yang, Nucleases: Diversity of structure, function and mechanism. *Q. Rev. Biophys.* **44**, 1–93 (2011).
2. L. M. Hellman, M. G. Fried, Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.* **2**, 1849–1861 (2007).
3. M. L. Bulyk, E. Gentalen, D. J. Lockhart, G. M. Church, Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.* **17**, 573–577 (1999).
4. N. Ogawa, M. D. Biggin, High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. *Methods Mol. Biol.* **786**, 51–63 (2012).
5. S. Wang, G. M. K. Poon, W. D. Wilson, Quantitative investigation of protein-nucleic acid interactions by biosensor surface plasmon resonance. *Methods Mol. Biol.* **1334**, 313–332 (2015).
6. N. Skivesen, A. Têtu, M. Kristensen, J. Kjems, L. H. Frandsen, P. I. Borel, Photonic-crystal waveguide biosensor. *Opt. Express* **15**, 3169–3176 (2007).
7. C. H. Leung, D. S. H. Chan, H. Z. He, Z. Cheng, H. Yang, D. L. Ma, Luminescent detection of DNA-binding proteins. *Nucleic Acids Res.* **40**, 941–955 (2012).
8. B. Dey, S. Thukral, S. Krishnan, M. Chakrobarty, S. Gupta, C. Manghani, V. Rani, DNA-protein interactions: Methods for detection and analysis. *Mol. Cell. Biochem.* **365**, 279–299 (2012).
9. D. Wanke, K. Harter, Analysis of plant regulatory DNA sequences by the yeast-one-hybrid assay. *Methods Mol. Biol.* **479**, 291–309 (2009).
10. J. D. Buenrostro, C. L. Araya, L. M. Chircus, C. J. Layton, H. Y. Chang, M. P. Snyder, W. J. Greenleaf, Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
11. C. Jung, J. A. Hawkins, S. K. Jones, Y. Xiao, J. R. Rybarski, K. E. Dillard, J. Hussmann, F. A. Saifuddin, C. A. Savran, A. D. Ellington, A. Ke, W. H. Press, I. J. Finkelstein, Massively parallel biophysical analysis of CRISPR-Cas complexes on next generation sequencing chips. *Cell* **170**, 35–47.e13 (2017).
12. Y. Xu, Z. Lin, C. Tang, Y. Tang, Y. Cai, H. Zhong, X. Wang, W. Zhang, C. Xu, J. Wang, J. Wang, H. Yang, L. Yang, Q. Gao, A new massively parallel nanoball sequencing platform for whole exome research. *BMC Bioinformatics* **20**, 153 (2019).
13. K. N. Natarajan, Z. Miao, M. Jiang, X. Huang, H. Zhou, J. Xie, C. Wang, S. Qin, Z. Zhao, L. Wu, N. Yang, B. Li, Y. Hou, S. Liu, S. A. Teichmann, Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biol.* **20**, 70 (2019).
14. T. Fehlmann, S. Reinheimer, C. Geng, X. Su, S. Drmanac, A. Alexeev, C. Zhang, C. Backes, N. Ludwig, M. Hart, D. An, Z. Zhu, C. Xu, A. Chen, M. Ni, J. Liu, Y. Li, M. Poulter, Y. Li, C. Stähler, R. Drmanac, X. Xu, E. Meese, A. Keller, cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenetics* **8**, 123 (2016).
15. S. Liu, S. Huang, F. Chen, L. Zhao, Y. Yuan, S. S. Francis, L. Fang, Z. Li, L. Lin, R. Liu, Y. Zhang, H. Xu, S. Li, Y. Zhou, R. W. Davies, Q. Liu, R. G. Walters, K. Lin, J. Ju, T. Korneliussen, M. A. Yang, Q. Fu, J. Wang, L. Zhou, A. Krogh, H. Zhang, W. Wang, Z. Chen, Z. Cai, Y. Yin, H. Yang, M. Mao, J. Shendure, J. Wang, A. Albrechtsen, X. Jin, R. Nielsen, X. Xu, Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. *Cell* **175**, 347–359.e14 (2018).
16. R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Huser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church, C. A. Reid, Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
17. M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier, A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
18. L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, W. A. Lim, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
19. O. Wagih, Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
20. R. T. Leenay, K. R. Maksimchuk, R. A. Slotkowski, R. N. Agrawal, A. A. Gomaa, A. E. Briner, R. Barrangou, C. L. Beisel, Identifying and visualizing functional PAM diversity across CRISPR-Cas systems. *Mol. Cell* **62**, 137–147 (2016).
21. A. Pingoud, A. Jeltsch, Recognition and cleavage of DNA by type-II restriction endonucleases. *Eur. J. Biochem.* **246**, 1–22 (1997).
22. P. Horvath, D. A. Romero, A.-C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, R. Barrangou, Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1401–1412 (2008).

23. H. Deveau, R. Barrangou, J. E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, S. Moineau, Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
24. S. Shmakov, O. O. Abudayyeh, K. S. Makarova, Y. I. Wolf, J. S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang, E. V. Koonin, Discovery and functional characterization of diverse class 2 CRISPR-cas systems. *Mol. Cell* **60**, 385–397 (2015).
25. W. Jiang, D. Bikard, D. Cox, F. Zhang, L. A. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).
26. F. Pukelsheim, The three sigma rule. *Am. Stat.* **48**, 88–91 (1994).
27. X. Meng, X. Hu, Q. Liu, X. Song, C. Gao, J. Li, K. Wang, Robust genome editing of CRISPR-Cas9 at NAG PAMs in rice. *Sci. China Life Sci.* **61**, 122–125 (2018).
28. Y. Zhang, X. Ge, F. Yang, L. Zhang, J. Zheng, X. Tan, Z.-B. Jin, J. Qu, F. Gu, Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. *Sci. Rep.* **4**, 5405 (2015).
29. B. P. Kleinstiver, M. S. Prew, S. Q. Tsai, V. V. Topkar, N. T. Nguyen, Z. Zheng, A. P. W. Gonzales, Z. Li, R. T. Peterson, J.-R. J. Yeh, M. J. Aryee, J. K. Joung, Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
30. C. Anders, O. Niewoehner, A. Duerst, M. Jinek, Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
31. D. L. Jones, P. Leroy, C. Unoson, D. Fange, V. Ćurić, M. J. Lawson, J. Elf, Kinetics of dCas9 target search in *Escherichia coli*. *Science* **357**, 1420–1424 (2017).
32. F. Safari, K. Zare, M. Negahdaripour, M. Barekati-Mowahed, Y. Ghasemi, CRISPR Cpf1 proteins: Structure, function and implications for genome editing. *Cell Biosci.* **9**, 36 (2019).
33. E. A. Boyle, J. O. L. Andreasson, L. M. Chircus, S. H. Sternberg, M. J. Wu, C. K. Guegler, J. A. Doudna, W. J. Greenleaf, High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5461–5466 (2017).
34. H. O'Geen, A. S. Yu, D. J. Segal, How specific is CRISPR/Cas9 really? *Curr. Opin. Chem. Biol.* **29**, 72–78 (2015).
35. S. H. Sternberg, B. Lafrance, M. Kaplan, J. A. Doudna, Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110–113 (2015).
36. F. Jiang, D. W. Taylor, J. S. Chen, J. E. Kornfeld, K. Zhou, A. J. Thompson, E. Nogales, J. A. Doudna, Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867–871 (2016).
37. E. A. Josephs, D. D. Kocak, C. J. Fitzgibbon, J. McMenemy, C. A. Gersbach, P. E. Marszalek, Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic Acids Res.* **43**, 8924–8941 (2015).

Acknowledgments: We thank the China National GeneBank and several scientists from BGI Research, including L. Xiao, Y. Zou, J. Li, Y. Li, and Z. Ji, for the friendly support and constructive suggestions. **Funding:** This work was supported by the Shenzhen Municipal Government of China Peacock Plan (no. KQTD20150330171505311), the Guangdong Provincial Key Laboratory of Genome Read and Write (no. 2017B030301011), Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics (DRC-SZ[2016]884), Shenzhen Institute of Synthetic Biology (ZTXM20190013), Shenzhen Institutes of Advanced Technology, and the China National GeneBank. **Author contributions:** Zhuokun L., X.W., X.D., Q.W., W.O., S.Z., B.H., J.Z. designed and/or conducted the wet-lab experiments; D.X., D.Z., D.W., Zhou L., J.G. designed and/or conducted bioinformatic analysis; Zhuokun L., X.W., D.Z., D.W., X.D., Y.J. prepared the tables, figures, and/or drafted the manuscript; Y.G., A.C., Y.S., Y.D., W.Z., Y.J. coordinated all the groups and edited the manuscript; X.X., C.X., Y.J. supervised the project and provided funding. All authors discussed the results and commented on the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The data that support the findings of this study have also been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession code CNP0000723.

Submitted 20 February 2020
Accepted 17 June 2020
Published 31 July 2020
10.1126/sciadv.abb3350

Citation: Z. Li, X. Wang, D. Xu, D. Zhang, D. Wang, X. Dai, Q. Wang, Z. Li, Y. Gu, W. Ouyang, S. Zhao, B. Huang, J. Gong, J. Zhao, A. Chen, Y. Shen, Y. Dong, W. Zhang, X. Xu, C. Xu, Y. Jiang, DNB-based on-chip motif finding: A high-throughput method to profile different types of protein-DNA interactions. *Sci. Adv.* **6**, eabb3350 (2020).