Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Methods Cell Biol. Author manuscript; available in PMC 2020 October 02.

Published in final edited form as:

Methods Cell Biol. 2018 ; 148: 229–250. doi:10.1016/bs.mcb.2018.09.011.

BET-seq: Binding energy topographies revealed by microfluidics and high-throughput sequencing

Arjun K. Aditham[*,†,2], **Tyler C. Shimko**[‡,2], **Polly M. Fordyce**[*,†,‡,§,1]

[*]Department of Bioengineering, Stanford University, Stanford, CA, United States

[†]Stanford ChEM-H, Stanford University, Stanford, CA, United States

[‡]Department of Genetics, Stanford University, Stanford, CA, United States

[§]Chan Zuckerberg Biohub, San Francisco, CA, United States

Abstract

Biophysical models of transcriptional regulation rely on energetic measurements of the binding affinities between transcription factors (TFs) and target DNA binding sites. Historically, assays capable of measuring TF-DNA binding affinities have been relatively low-throughput (measuring ~10^3 sequences in parallel) and have required significant specialized equipment, limiting their use to a handful of laboratories. Recently, we developed an experimental assay and analysis pipeline that allows measurement of binding energies between a single TF and up to 10^6 DNA species in a single experiment (Binding Energy Topography by sequencing, or BET-seq) (Le et al., 2018). BET-seq employs the Mechanically Induced Trapping of Molecular Interactions (MITOMI) platform to purify DNA bound to a TF at equilibrium followed by high coverage sequencing to reveal relative differences in binding energy for each sequence. While we have previously used BET-seq to refine the binding affinity landscapes surrounding high-affinity DNA consensus target sites, we anticipate this technique will be applied in future work toward measuring a wide variety of TF-DNA landscapes. Here, we provide detailed instructions and general considerations for DNA library design, performing BET-seq assays, and analyzing the resulting data.

1 INTRODUCTION

Protein-DNA interactions play a central role in governing the transcriptional activity of cells. The strength of these interactions can be described by the change in Gibbs free energy associated with the binding process (ΔG), and the most successful efforts to model transcription factor (TF) binding at a particular genomic locus have predicted TF occupancy via thermodynamic models that consider the available TF concentration in the nucleus and the ΔG for any underlying regulatory sequence (Weirauch et al., 2013; Zhao, Granas, & Stormo, 2009; Zhao & Stormo, 2011). Reliably predicting TF occupancies genome-wide therefore requires predicted ΔG values for a wide variety of sequences including all candidate genomic binding sites. In practice, these energies are typically determined by measuring TF binding to a library of sequences containing systematic substitutions within

[1]Corresponding author: pfordyce@stanford.edu.

[2]These authors contributed equally to this work.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

them and then calculating the change in binding energy (ΔΔG) for each relative to a single reference sequence. Despite their importance, such ΔΔG measurements cannot typically be obtained using many leading characterization technologies, including high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) (Chen et al., 2016; Jolma, Kivio et al., 2010; Jolma, Yan, et al., 2013) or protein binding microarrays (PBMs) (Badis et al., 2009; Mukherjee et al., 2004). Although techniques such as electrophoretic mobility shift assays (EMSA) (Fried, 1989; Hellman & Fried, 2007) and Mechanically Induced Trapping of Molecular Interactions (MITOMI) (Fordyce et al., 2010; Maerkl & Quake, 2007; Rockel, Geertz, & Maerkl, 2012) can collect high-resolution binding energy data, this comes at the expense of throughput, as these are typically limited to 10–100 s of interactions in parallel.

To address these limitations, we recently developed a technique for the measurement of relative differences in binding energies (ΔΔGs) for up to millions of sequences in parallel (Binding Energy Topographies by sequencing, or BET-seq) (Le et al., 2018) (Fig. 1). BET-seq combines the MITOMI platform and high-throughput DNA sequencing to profile interactions between a single TF and a pooled library of DNA sequences and quantify the proportion of each species within the “bound” and “unbound” fractions. Briefly, a fluorescently-labeled transcription factor is expressed *in vitro* and introduced into a MITOMI microfluidic device that has been functionalized to purify the TF on-chip. Following TF recruitment, a DNA pool is introduced into the device and allowed to interact with surface-immobilized TFs until reaching equilibrium. At this point, pneumatic valves within the device (termed “button” valves) are actuated, thereby trapping DNA bound to the TFs and preserving transiently-bound interactions while unbound DNA is washed from the device. The button valves are then reopened, allowing collection of DNA bound to TFs for subsequent quantitation via high-throughput sequencing. As an initial demonstration of this technique’s power, we measured relative binding energies for 1,048,576 flanking sequences surrounding the consensus *E*-box binding motif for the yeast transcription factors Pho4 and Cbf1 (Le et al., 2018). We noted differential influence of non-additive dinucleotide effects between these two proteins and measured energetic specificity for positions farther from the *E*-box motif than previously reported, underscoring the importance of such flanking sequence effects on *cis*-regulatory function and evolution.

In future work, we anticipate that the BET-seq technique can be applied to a wide range of problems, from refining motifs for additional TFs to measuring affinities for smaller libraries to energetically calibrate intensities from protein binding microarray experiments (PBMs). However, successful quantitative measurement of TF-DNA affinities via BET-seq depends critically on careful upfront consideration of the desired library size, expected energetic range of TF-DNA binding interactions, and the likely distribution of these energies. In this chapter, we provide guidelines for BET-seq DNA library design, complete protocols for all steps of the BET-seq assay, and pipelines and troubleshooting tips for analysis of the resulting data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2 BET-SEQ LIBRARY DESIGN AND SYNTHESIS

For TF-DNA interactions measured at equilibrium, the relative difference in binding energies between a library of sequences can be calculated as follows:

$$\Delta\Delta G_i = -RT \ln([bound_i]/[unbound_i])$$

where ΔΔGi represents the change in Gibbs free energy of binding for a particular sequence, RT is the product of the molar gas constant and temperature (0.593 kcal/mol), and [$bound_i$] and [$unbound_i$] represent the concentrations of a particular species within the TF-bound and unbound fractions, respectively. The relative concentrations of a particular species can be measured by using deep sequencing of each fraction (Zuo & Stormo, 2014). As long as sequencing coverage is high enough that the Poisson count noise for a given species is sufficiently low, this approach allows high-resolution measurement of many energies in parallel.

The MITOMI microfluidic platform allows for rapid full-length protein purification from *in vitro* transcription and translation reactions and nucleic acid enrichment at a scale compatible with libraries of up to millions of species. This platform contains 1568 TF chambers each with a volume of about 0.6 nL. At a typical BET-seq DNA library concentration of 1 μM, the maximum amount of DNA across all chambers (assuming every TF is occupied) is 3×10^{-4} mol (roughly 1.8×10^{10} molecules). For a library of 10^6 species, a given MITOMI device would contain roughly 10,000 copies of each species, allowing high-resolution counting of both the bound and the unbound fractions.

For assays where the unbound material is in vast excess of the bound material (*e.g.*, where [DNA]≫[TF]) and the spread of expected binding energies is <4 kcal/mol, ligand depletion is not a concern. Binding energies can then be calculated for each species from the ratio of bound to input (rather than unbound) concentrations, drastically reducing the cost of running multiple related experiments. If the expected binding energy spread is too great, however, high-affinity sequences will be depleted from the unbound fraction, causing estimates of abundance derived from the input fraction to overestimate unbound concentration.

Sufficient average read depth per library species is also essential for obtaining accurate molecular counts of each species within the bound and unbound fractions. Low depth sequencing is strongly skewed by stochastic sampling noise, returning less reliable estimates of sequence concentrations. Additional detail regarding required sequencing depth required for a given library can be found in the original BET-seq publication (Le et al., 2018).

2.1 DESIGNING A BET-SEQ LIBRARY

Quantifying concentrations of each species within the bound and input fractions via Illumina sequencing places several design constraints upon the DNA library architecture. First, all species sequenced must contain the P5 and P7 sequences required to anneal molecules to the Illumina flow cell. Second, Illumina sequencing requires read primer annealing sites and library barcodes for multiplexing. These sequences can be synthesized as part of the initial library or incorporated via PCR using extended amplification primers, depending on the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

required sequencing scale. After the library is designed, single-stranded DNA libraries can be ordered in a pooled format for libraries containing degenerate bases or separate oligonucleotides for when very few sequences are being queried. In the latter case, separate syntheses must be carefully combined, as described in Section 2.2.

2.1.1 Transcription factor binding site and low affinity linker design—In theory, BET-seq-like assays can measure ΔΔGs for a wide variety of libraries, as long as (1) DNA library sequences are in sufficient excess of available TFs to allow comparison of bound material with input, and (2) DNA sequences are sequenced to a read depth that allows counting of individual species with low error. BET-seq and similar assays perform best when the total ΔΔG spread within the library is sufficiently narrow, <4 kcal/mol, ideally closer to 2 kcal/mol. Generally, the difference in binding energies between high affinity “consensus” binding and nonspecific binding is around 2–4 kcal/mol (Geertz, Shore, & Maerkl, 2012), suggesting this criterion can be satisfied in most situations. To date, however, we have only used this technique to probe effects of varying nucleotide identity around a central high-affinity site. Such experiments allow quantitative determination of subtle differences in binding energies and simplify downstream analysis due to the alignment-free nature of positional effect calculations. Therefore, we focus on this use case here.

BET-seq libraries designed to systematically assess the effects of particular nucleotides at specific positions must contain a number of fixed nucleotide positions to “anchor” the TF binding footprint, variable positions for which changes in affinity will be measured, and PCR handles for efficient incorporation of sequencing adapters. Given a position weight matrix (PWM) detailing the binding preference for a TF of interest, fixing the position, spacing, and identity of high-information content nucleotides allows systematic investigation of the effects of other nucleotides whose effects may not be visible during typical high-throughput assays (Fig. 2A). Adding a polynucleotide linker sequence lengthens the DNA strand, provides a fixed sequence context, and reduces potential edge effects. Distributing unique molecular identifiers (UMIs) throughout the linker sequence aids removal of PCR duplicates from read counts (Kivioja et al., 2012). To ensure that secondary, off-target binding sites are not generated accidentally during library design, all permutations of variable region sequences within the query library should be scanned with a motif scanning tool, such as FIMO, before synthesis (Grant, Bailey, & Noble, 2011).

Oligonucleotide libraries must be manipulated in several ways prior to and after BET-seq experiments. Prior to experiments, libraries ordered as single-stranded ultramers must be converted to double-stranded DNA. After experiments, scarce DNA (~10 fmol) eluted from the device must be amplified for sequencing, absolute quantities of bound material must be determined via qPCR, and adapters must be added for subsequent deep sequencing. In all cases, library manipulations must not accidentally incorporate a secondary high-affinity TF binding site prior to the assay and UMIs must preserve molecular identity throughout the experimental pipeline.

2.1.2 Small-scale library design—The MiSeq platform has a total output capacity of 25 million reads and performs well with low complexity libraries, making it ideal for BET-seq experiments using smaller libraries. For small-scale libraries, three fragments must be

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

synthesized to prepare the library for sequencing: the variable region (described above), two adapter incorporation primers to add indices, and the P5 and P7 sequences (Fig. 2B). The most efficient small-scale library designs use the Read 1 and Read 2 primer sites as PCR handles for adapter incorporation, minimizing the number of bases incorporated during synthesis and thereby reducing cost and the chance of synthesis errors. However, this library design requires that the TF of interest not exhibit significant affinity for the Read 1 and Read 2 sequences, which could decrease overall resolution. If a TF of interest has significant affinity for these sequences, alternative PCR handles can be used and Read 1/Read 2 can be incorporated upstream of these handles during PCR-based adapter incorporation upon collection of the bound and input fractions (similar to strategy shown in Fig. 2C).

2.1.3 Large-scale library design—Illumina also offers NextSeq, HiSeq, and NovaSeq platforms that return a significantly larger number of reads per experiment. To minimize stochastic read noise, these higher-output platforms are required for BET-seq experiments attempting to directly measure energies for a large input library. In addition, these platforms allow for greater multiplexing of samples without a directly proportional increase in cost, making them suitable for experiments with large numbers of replicates or for characterizing whole TF families in parallel.

However, these high-throughput Illumina sequencing platforms introduce additional experimental challenges for sequencing low complexity libraries (such as those containing systematic mutations within and around a known binding site). Specifically, the NextSeq and HiSeq platforms rely on the first 26 base pairs sequenced to separate clusters and determine the overall quality of the sequencing run, and low complexity within these base pairs can lead to run failure. To overcome this, BET-seq libraries must incorporate a 26 base pair randomer at the beginning of each sequence. To prevent the accidental introduction of a secondary high-affinity TF binding site within this 26 bp random region, this randomer should be incorporated within the primer used to add the P5 and Read 1 sequences (5′ to the homology region) during the post-assay PCR amplification step (Fig. 2C). The same P7 incorporation primer can be used for both large- and small-scale library preparation.

2.2 PREPARING A BET-SEQ LIBRARY

Experimental measurements of TF-DNA binding energies assess binding to double-stranded DNA (dsDNA) to mimic genomic binding *in vivo*. Therefore, single-stranded DNA (ssDNA) from DNA synthesis companies must be duplexed into dsDNA. To accomplish this, ssDNA is incubated with the Illumina Read 2 primer, which hybridizes to the 3′ end of all sequences in the library, and then extended via a single cycle of PCR with a high-fidelity polymerase (Fig. 2D). Using only a single cycle of PCR prevents the amplification of DNA sequence bias beyond that introduced during DNA synthesis. Yields from DNA synthesis are sufficient for multiple rounds of library preparation, obviating the need to amplify the DNA.

After amplification, libraries are purified using the Zymo DNA Clean and Concentrate (25 μg scale) kit, quantified via spectrophotometry, and diluted to a final working concentration of 1 μM prior to introduction into the device.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Reagents

- MM—50 μL of Q5 Hot Start High-Fidelity 2X Master Mix
- DNA—40 μL of DNA ultramer (10 μM, in Milli-Q H2O)
- Primer—5 μL of Illumina Read 2 Primer (100 μM, in Milli-Q H2O)
- Water—5 μL of DNAse-/RNase-free water
- Zymo Clean and Concentrate—25 column purification kit

Protocol (per reaction or sub-library)—Add all reagents in order into a PCR tube that has been placed on ice. Upon defrosting, keep MM, DNA, and primer on ice.

1. Add 40 μL of DNA to the tube.
2. Add 5 μL of Water to the tube and mix by pipetting.
3. Add 5 μL Primer to the tube and mix by pipetting.
4. Add 50 μL of MM to the tube. Mix by pipetting until well-mixed (no visible phase separation).

Close tubes and place into a thermocycler with the following protocol

1. Initial melt: 98 °C for 30 s (with a 4 °C/s ramp rate)
2. 1 cycle:
    a. Melt: 98 °C for 10 s (with a 0.1 °C/s ramp rate)
    b. Anneal: 30 °C for 1 min and 15 s (with a 0.1 °C/s ramp rate)
3. Final Extension: 65 °C for 5 min
4. Post-reaction temperature: 4 °C

After PCR, clean each BET-seq library using a Zymo Clean and Concentrate kit (25 μg scale) as specified in the user manual. Post-cleanup, quantify DNA concentration using a NanoDrop (or equivalent spectrophotometer), convert absorbance at 260 nm to a molar concentration using an online calculator, and then dilute to 1 μM in the elution buffer provided in the kit.

As mentioned previously, a BET-seq input DNA library can contain multiple sub-libraries. To prevent one sub-library from dominating in downstream experiments and sequencing, each sub-library must be present in the final Input in equimolar amounts. After cleanup, measure sub-library concentrations, convert to molar concentration, dilute each sub-library to 1 μM, and confirm the concentration. To form the final input library for all subsequent BET-seq experiments, combine an equal volume of each library to a final total DNA concentration of 1 μM, as depicted in Fig. 2E.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3 PERFORMING A BET-SEQ EXPERIMENT

A BET-seq experiment contains four steps (Fig. 3A and B): (1) device surface chemistry, (2) TF-DNA binding equilibration, (3) device wash, and (4) bound DNA elution. During the first step, the device's button valves labeled in Fig. 3A) are closed and BSA is introduced to passivate all device surfaces except those directly beneath the valves. These valves are then opened to allow specific patterning of surfaces beneath the button with biotinylated BSA, neutravidin, anti-GFP antibody and subsequent recruitment of mGFP-tagged (monomeric enhanced GFP) TFs. Following TF immobilization, the DNA library is introduced into the device and incubated until equilibrium is reached. After incubation, button valves are closed, thereby trapping DNA molecules bound to TFs and protecting them as nonspecifically bound TF and DNA is washed away. Finally, button valves are opened and the device is washed to elute the "bound" DNA.

Confidence in measurements of TF specificity requires controls that ensure measured binding preferences reflect true specificities and not sequence-specific background binding. To ensure that an adequate effective concentration of the TF of interest has been recruited to the device surface, TF constructs are fused to a C-terminal mGFP tag, allowing direct quantification of immobilized TFs via fluorescence microscopy. These TF-mGFP fusions are subcloned into a pTNT vector (Promega) via Golden Gate Assembly for expression in Wheat Germ Extract in vitro transcription/translation (IVTT) mix (Promega) using the protocol previously described (Le et al., 2018). We have also expressed proteins in PurExpress (New England Biolabs), with constructs cloned into the manufacturer-suggested vector. TFs should typically be expressed for ~3 h for higher protein yield and mGFP folding.

3.1 DEVICE SURFACE CHEMISTRY

Before beginning surface patterning, a MITOMI device must be fabricated, thermally bonded overnight to a epoxy-coated glass slide, and connected to the control lines of a custom-built pneumatics system (Brower, Puccinelli, et al., 2017). Auto-CAD files of MITOMI devices are available on the Fordyce Lab website (http://www.fordycelab.com/microfluidic-design-files/). We have also published extensive protocols detailing MITOMI mold and device fabrication (Brower, White, & Fordyce, 2017) and operation, including connecting the device to pneumatic systems and software control (Brower, Puccinelli, et al., 2017, https://github.com/FordyceLab)). Control lines should be pressurized to ~30 PSI to ensure complete closure of pneumatic valves during surface chemistry steps. Flow lines are typically pressurized to ~4 PSI to ensure that reagents within device channels are completely exchanged within 5–10 min. While we describe the procedure for the previously published 1500 chamber MITOMI device, experiments can employ a simpler device containing only the flow lines and button valves; with the full MITOMI device, the "neck" and "sandwich" valves should remain shut and open, respectively, throughout the experiment.

Smooth, laminar flow throughout the device requires elimination of air from device channels prior to surface patterning. To do this, pressurize all control lines within the device to fill them with water, verifying that all air bubbles are pushed out through the PDMS. Visually inspect each valve while opening and closing them to ensure the device functions properly.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Next, introduce pressurized 1 × phosphate buffered saline into all device channels for 2 min, close the device outlet valve, and wait until all air is pushed out of channels (approximately 5 min). After this step, the other reagents can be introduced onto the device.

Required reagents

- bBSA—200 μL biotinylated BSA (2 mg/mL in 140 nM citrate, pH 6.8), combined with 50 μL poly(deoxyinosinic-deoxycytidylic) acid (250 μg/mL in NGS clean-up kit provided elution buffer)
- NA—100 μL neutravidin (1 mg/mL, in 1X phosphate buffered saline)
- anti-GFP—50 μL biotin-conjugated antibody against green fluorescent protein (0.04 mg/mL in PBS)
- PBS—500 μL 1X phosphate buffered saline
- TF—100 μL IVTT TF expression solution (expressed for 3 h)

Protocol

1. For each reagent, purge any air bubbles from input lines by using the input manifold to flow reagents to “Waste” valve for 15 s. In between reagents and before proceeding, wash manifold channels by flowing PBS to “Waste” for 15 s.
2. Flow bBSA through device for 20 min.
3. Flow PBS through device for 6 min, 40 s.
4. Flow NA through device for 20 min.
5. Flow PBS through device for 6 min, 40 s.
6. Close “Button” valves. To validate that the valves have shut, examine the device underneath a stereoscope; the valves should appear rounded and engorged when shut.
7. Flow PBS through device for 5 min.
8. Flow bBSA through device for 20 min.
9. Flow PBS through device for 6 min, 40 s.
10. Flow anti-GFP antibody through device for 1 min, 20 s. This ensures that the antibody (and, subsequently, the TF) is uniformly distributed throughout the device.
11. Open button valves and confirm that the valves are open by viewing the device underneath the stereoscope. The valves should no longer appear engorged.
12. Continue flowing anti-GFP antibody through device for 11 min, 40 s.
13. Flow PBS through device for 6 min, 40 s.

At the conclusion of this step, increase both the flow and control line pressures. The higher flow line pressure facilitates faster deposition of protein onto the device and creates stringent

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

wash conditions to reduce nonspecific TF and DNA binding. The higher control line pressure further ensures the valves shut completely and slightly increases the area protected by the button valve, as DNA is soon added to the device.

After antibody deposition, introduce and immobilize mGFP-tagged TFs. To reduce the possibility of any aggregates clogging the device, centrifuge the IVTT mixture at maximum speed for 5 min to pellet any unfolded and aggregated protein. Follow these next steps.

1. For TF loading, purge air in inlet tree by flowing inlet to “Waste” valve for 15 s, and then flow PBS to “Waste” for 15 s.
2. Flow TF onto device for 20 min.
3. Flow PBS through device for 6 min, 40 s
4. Mount slide onto microscope and image the device in the GFP channel (488 nM excitation) to visualize mGFP-tagged TF. It should appear as shown in Fig. 3A.

Images from the device should be examined to confirm quality of TF deposition before proceeding. TF deposition should be consistent across the device; inconsistent deposition suggests a device flow defect or clog that reduces flow within particular channels. High GFP signal outside the button area indicates nonspecific TF adsorption that can affect the DNA binding step by depleting DNA molecules meant for TFs underneath the button valves. In all cases, a stringent wash using a protease (detailed in Section 3.3) can be used to cleave nonspecifically adsorbed TFs before introducing the DNA library. Based on a mGFP calibration curve, we estimate a normal effective concentration of immobilized TFs of ~30 nM.

3.2 INTRODUCTION OF DNA LIBRARY AND INCUBATION

Following attachment of TFs, the pooled DNA library is introduced into the device and allowed to interact with the TFs. TF-DNA interactions must reach equilibrium prior to trapping TF-bound DNA for accurate binding energy measurements. Typical time constants of on- and off-rates for TF-DNA interactions are on the order of seconds (Geertz et al., 2012; Spinner, Liu, Wang, & Schmidt, 2002), suggesting that an incubation time of 60 min should be more than sufficient to reach equilibrium. This step covers the introduction of the pooled DNA, equilibration of TF-DNA binding, pressurization of button valves to trap TF-bound DNA molecules, and removal of any unbound DNA from the device.

Note that this protocol assumes that the input DNA library can serve a proxy for the unbound fraction of DNA in subsequent analysis. If the input cannot be used to estimate concentrations within the unbound fraction, the post-equilibration wash (Step 5 of this part of the protocol) should be saved for sequencing. The fractions and overall schematic of this step are depicted in Fig. 3B.

Required reagents

- Input—40 μL library DNA substrate (1 μM, in 1X Tris-EDTA buffer)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Protocol

1. Flow Input through device until nearly all DNA has been introduced.
2. Close “Out” valve.
3. Equilibrate for 60 min.
4. Close “Button” valves and confirm buttons are shut under stereoscope.
5. Flow PBS through device for 5 min.

3.3 DEVICE WASH

A stringent wash of device channels with a relatively nonspecific protease (trypsin) can be used to cleave nonspecifically adsorbed TFs and DNA prior to elution of bound molecules. To prevent subsequent sticking of the bound DNA fraction, device surfaces are again coated with bBSA to regenerate the protein sacrificial layer.

Required reagents

- Trypsin—250 μL freshly prepared bovine trypsin (2 mg/mL, in 1X PBS)

Protocol

1. Purge air from trypsin reagent by flowing inlet to “Waste” valve for 15 s.
2. Flow trypsin through device for 30 min.
3. Flow PBS through device for 5 min.
4. Flow bBSA through device for 10 min.
5. Flow PBS through device for 10 min.

After washing, image the device once again to ensure that any background GFP signal (outside button valve areas) is significantly reduced.

3.4 BOUND DNA ELUTION

At the end of the experiment, TF-bound DNA (the “bound” fraction) must be eluted from the device. To maximize yield, TF-bound DNA molecules are physically stripped from the device via repeated button actuations under constant PBS flow. Before beginning this elution step, place a micropipette tip firmly into the outlet port of the device to collect the eluent. The pipette tip should have a capacity of at least 100 μL of liquid to prevent overflow.

Protocol

1. 300 cycles: “Button” open 2 s, “Button” shut 2 s, constant PBS flow.
2. As the DNA is eluted, monitor the pipette trip to ensure the fluid line is increasing. If there is no change in the fluid level, this may indicate leakage or loss of flow.
3. After the elution cycles finish, close “In” valve and “Out” valve.
4. Remove tip carefully and place in a clean PCR tube.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Connect micropipette and carefully expel eluent into PCR tube.

4 SEQUENCING LIBRARY PREPARATION

PCR amplification of eluted bound DNA amplifies library material prior to sequencing and provides an opportunity to add sequencing adapters, indices to de-multiplex library samples, and add random nucleotides for compatibility with higher-throughput NextSeq and HiSeq Illumina platforms (Fig. 4A and B). Although PCR amplification cycles should be limited to prevent “jackpotting” of individual sequences, the incorporation of UMIs within the initial library design allows identification of and correction for any amplification bias.

4.1 USE PCR TO ADD SEQUENCING ADAPTERS

Multiple experiments can be sequenced in a single run by adding indexed sequencing adapters to bound and input fractions that encode experiment and fraction identity. If the input fraction is common to all BET-Seq experiments being prepared, the input need only be prepared once and can be used for comparison with bound fractions in all experiments. Even when sequencing the unbound fraction, however, the input fraction must also be sequenced in order to account for library bias in downstream data analysis. Fig. 4A and B provides a schematic summarizing these steps.

Required reagents (per reaction)

- MM—35 μL of Q5 Hot Start High-Fidelity 2X Master Mix
- P5—5 μL of P5 indexed primer (10 μM, in Milli-Q H2O)
- P7—5 μL of P7 indexed primer (10 μM, in Milli-Q H2O)
- Input—25 μL of Input DNA (100 pM, in 1X Tris-EDTA buffer)
- Bound—25 μL of bound fraction DNA (variable concentration, in 1X Phosphate Buffered Saline)
- ThermoFisher GeneJET Cleanup Kit

Protocol (per reaction)—At each step, reagents should be added into PCR tubes that have been placed on ice. Upon defrosting, keep the following on ice: MM, P5, P7, and Input.

1. Add 25 μL of Bound DNA to respective PCR tube(s) and 25 μL Input DNA to its own respective PCR tube.
2. Add 5 μL of P5—note corresponding index.
3. Add 5 μL of P7—note corresponding index.
4. Add 35 μL of MM and mix by pipetting until the reaction is well-mixed (no visible phase separation).

Close tubes and place into a thermocycler with the following protocol (illustrated in Fig. 4B)

1. Initial melt: 98 °C for 30 s (with a 4 °C/s ramp rate)
2. 11–15 cycles (variable cycle number):

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a. Melt: 98 °C for 10 s (with a 4 °C/s ramp rate)

b. Anneal and Extend: 65 °C for 75 s (with a 4 °C/s ramp rate)

3. Final Extension: 65 °C for 5 min

4. Post-reaction temperature: 4 °C

After PCR, remove polymerases, unincorporated dNTPs, and unused adapters using a ThermoFisher GeneJET Cleanup Kit according to Protocol B of the manufacturer's instructions. After products are cleaned, proceed to quality control steps, detailed below.

4.2 qPCR TO QUANTIFY LIBRARY CONCENTRATION

For all BET-seq experiments, the input and bound DNA fractions must be sequenced at approximately equal depths to accurately assess the relative enrichment of each species within the bound fractions. qPCR allows reliable quantitation of DNA concentrations to create an equimolar mix of each fraction prior to sequencing. This step is especially important if multiple BET-seq experiments are combined in a single run (as in Fig. 4C) to ensure sequencing provides a sufficient number of reads for all experiments.

To quantify the library concentration, use the NEB Next Library Quant Kit for Illumina (NEB #E7630S/L) according to the manufacturer instructions and analyze data using the baseline subtracted curve fit setting. Sample traces for a successful library fall between the qPCR standards, suggesting nM sample DNA concentrations. After all sample traces are collected, calculate concentrations using the expected size of the sequencing amplicon, dilute all sub-libraries to 10 nM in the elution buffer supplied with the ThermoFisher GeneJET Cleanup Kit, and combine 10 μL of each sub-library to a final tube. Before sequencing, confirm that the final DNA concentration is approximately 10 nM via qPCR (Fig. 4C).

4.2.1 Considerations for low concentration libraries—

Cleaned DNA libraries should ideally contain at least 10 nM of DNA for sequencing, but lower yields might result. In case of low yields, repeat the PCR described in Section 4.1 with a higher number of cycles using original eluent from the BET-seq experiment. Increasing the cycle count also amplifies PCR bias; thus, surpassing 17 cycles is not advised. If increasing the PCR cycle number does not help, this might suggest BET-seq experimental failure or failure of the TF to bind DNA under the conditions of the assay.

4.3 CAPILLARY ELECTROPHORESIS QUANTIFICATION OF THE SEQUENCEABLE LIBRARY FRACTION

Checking the purity of the sequencing sample and quantifying the fraction of the total library that is sequenceable is important for ensuring success in high-throughput sequencing. Prior to sequencing, submit each library for analysis via capillary electrophoresis (*e.g.*, Bioanalyzer), which allows estimation of the concentration of sequencing library using the calculated concentration underneath the desired peak and further reveals the presence of any fragments that are not the expected length of the final library. A successful Bionanalyzer

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

trace (Agilent, High Sensitivity DNA Assay) for a cleaned sample submission with an expected peak size of 205 bp is shown in Fig. 4E.

Fragments that are smaller than the expected library size might be unadapted primers which can affect clustering. In this case, we advise starting from Section 4.1 to generate a cleaner sample. Conversely, contaminating fragments larger than the amplicon might represent library concatemers. Testing whether these fragments will affect clustering can be performed via a test PCR followed by agarose gel electrophoresis. This PCR should be performed using the same adapter addition protocol outlined in Section 4.1, except using the sequencing submission sample as template, unindexed P5 and P7 primers, and amplifying for 40 cycles. If the gel reveals that only the desired fragment can be amplified using P5 and P7, the sample is acceptable for high-throughput sequencing as long as concentration is estimated using only the peak under the desired fragment.

5 ANALYSIS

5.1 DATA QUALITY CHECK

After all samples are sequenced, several preliminary analysis steps are required to verify the quality of the returned data prior to more complex downstream energetic analysis. The first step checks that the number of total returned reads is sufficiently high. For a given sequencing kit size, the number of returned reads across all multiplexed libraries should be roughly equal to the total output of the kit less any fraction of the kit devoted to PhiX or other spike-in libraries. Lower-than-expected read counts indicate under-clustering, over-clustering to the detriment of read quality below filter specifications, or a generally low-quality library.

Assuming the sequencing run has returned a sufficient number of reads, the second step checks overall library quality. This quality check can be performed easily using a tool such as FastQC (Andrews, 2010). High quality libraries should have PHRED scores >30 with a gradual decrease as the position approaches the end of the read, as shown in Fig. 5B. Libraries with large swings in quality score or generally low quality scores likely indicate issues during the sequencing run. Additionally, skews in GC content are a known issue with sequencing-based SELEX-like assays, such as BET-seq (Orenstein & Shamir, 2014). At the time of writing, no solution has yet been implemented for such bias.

5.2 ENERGETIC SPECIFICITY ANALYSIS

5.2.1 Relative binding energy calculation—Once the library has passed initial quality controls, energetic specificity analysis can begin. For a library designed to assess the effects of nucleotides at particular positions (*e.g.*, flanking sequence analysis, Le et al., 2018), the first step is to extract positions of interest within the backbone of the library. These positions include any variable regions (such as variable positions flanking a high-affinity consensus) and any distal random nucleotides (UMIs) used for sequence deduplication. For each read in the sequencing data, the quality scores of the nucleotides at these positions and within the binding site should exceed a PHRED quality score of 30. Finally, read counts should be deduplicated using custom analysis scripts for both the TF-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

enriched and the input libraries. Following the initial quality control check for the nucleotide positions of interest, the sequences and read count-based concentration estimates can be used to calculate a relative binding energy for each species.

Because unequal read depths for the bound and input libraries can skew concentration estimates using read counts alone, each count should be normalized to total library read depth (*e.g.*, P(bound) = bound_count/sum(bound_counts)). ΔΔG values can then be calculated using the following equation to yield relative changes in binding energy in units of kcal/mol:

$$\Delta\Delta G_i = -RT \ln(P(bound_i)/P(unbound_i))$$

Here, RT indicates the product of the molar gas constant and temperature (0.593 kcal/mol). Due to potential discretization of energy values due to the nature of relying on count measurements, the mean of the ΔΔG distribution should be subtracted from all individual ΔΔG values to ensure that the distribution meets the theoretical expectation of being centered at 0. Because ΔΔG is a change in binding energy relative to a reference point, any reference point, including a defined "wild-type" sequence, can be used to define the zero point of the distribution.

5.2.2 Linear mononucleotide and dinucleotide models—Once ΔΔG values have been calculated, the next step entails checking trends within the returned data to determine whether the experiment has revealed new information about TF specificity preferences. To do this, group sequences by each nucleotide at each position and calculate the mean, standard deviation, and standard error on the mean for the ΔΔG distribution for each position-nucleotide pair. Logo representations similar to traditional position weight matrices (Schneider & Stephens, 1990; Stormo, Schneider, Gold, & Ehrenfeucht, 1982) can then be created by plotting the mean of the ΔΔG distribution for each nucleotide at each position, revealing shifts in the energetic distribution attributable to each nucleotide at each position (Fig. 5E). Such plots provide initial diagnostics for specific nucleotide/position effects, allowing observation of both preferred and disfavored nucleotides at each position and the degree to which they promote or penalize binding. If prior information regarding nucleotide specificity has been collected, these plots should recapitulate known information in most cases. Inconsistencies with expectation at this step may indicate experimental artifacts or failure. Dinucleotide representations, useful for implicit modeling of DNA shape features (Rube, Rastogi, Kribelbauer, & Bussemaker, 2018), can additionally be created by grouping on each dinucleotide pair and calculating the mean relative energetic difference.

5.2.3 High-complexity models—One advantage of a "flanking" sequence library is that variable position effects are "anchored" in place around a high-affinity binding site. This anchored position allows consideration of nucleotide effects at a particular position by simply averaging the binding energies over all species containing that position/nucleotide pair in the entire library. Attempts to use more complex libraries, however, may require more complex models designed to scan sequences for candidate binding sites and align them relative to one another. More complex energetic binding specificity models can be generated

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

using existing software tools, such as BEEML (Zhao et al., 2009) or through custom deep neural network-based analysis as in (Le et al., 2018). While neural network parameterizations, such as number and size of layers, will change based on the complexity of TF binding modalities examined, general principles exist for modeling binding energies. Generally, input to the neural network should be defined as a 4xL one-hot encoded matrix, optionally flattened, and the predicted output should be the BET-seq measured ΔΔG value. Depending on anticipated length and complexity of the binding site preferences (*e.g.*, continuous vs. gapped binding motifs), either feedforward or convolutional neural network architectures can be used. In all cases, analysts should take care to avoid overfitting of neural network-based models by monitoring a withheld validation subset of the data and halting training once loss on the validation set fails to decrease. For a comprehensive overview of the intricacies of building, training, and tuning neural network models for applications similar to those presented here, see Angermueller, Pärnamaa, Parts, and Stegle (2016).

One major advantage of neural network modeling schemes is that sequence features, normally defined as either mononucleotide or k-mers in statistical models, are learned automatically by the network, potentially creating more powerful predictive models of TF binding specificities. However, these models are far less interpretable than traditional featurized statistical models. Ultimately, analysts should decide whether prediction accuracy or model interpretation is more important and structure their analysis accordingly.

6 CONCLUSION

The advent of high-throughput, *in vitro* TF specificity characterization technologies has enabled the collection of detailed, high information content motifs for a wide variety of TFs across many species. However, the majority of these motifs lack thermodynamic meaning due to limitations of the characterization technology or the data analysis process. Here, we present a straightforward protocol for the BET-seq assay as well as generalizable principles for library design and analysis of experimental results. This assay allows for reliable collection of relative energetic specificity information surrounding a binding energy minimum and is suitable for refinement of existing binding site motifs. When combined with measurements of absolute binding energy as in (Le et al., 2018), this technique can yield specificity information on an absolute scale for up to millions of DNA species in parallel.

ACKNOWLEDGMENTS

A.K.A. and T.C.S. acknowledge support from NSF Graduate Research Fellowships. A.K.A. acknowledges support from the Stanford ChEM-H Chemistry/Biology Interface Predoctoral Training Program. P.M.F. is a Chan Zuckerberg Biohub Investigator and acknowledges Sloan Research Foundation and McCormick and Gabilan faculty fellowships. This work was supported by NIH/National Institute of General Medical Sciences Grant R00GM09984804.

REFERENCES

Andrews S (2010). FastQC: A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Angermueller C, Pärnamaa T, Parts L, & Stegle O (2016). Deep learning for computational biology. Molecular Systems Biology, 12(7), 878 10.15252/msb.20156651. [PubMed: 27474269]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science, 324(5935), 1720–1723. 10.1126/science.1162327. [PubMed: 19443739]

Brower K, Puccinelli R, Markin CJ, Shimko TC, Longwell SA, Cruz B, et al. (2017). An open-source, programmable pneumatic setup for operation and automated control of single- and multi-layer microfluidic devices. HardwareX, 3, 117–134. 10.1016/j.ohx.2017.10.001. [PubMed: 30221210]

Brower K, White AK, & Fordyce PM (2017). Multi-step variable height photolithography for Valved multilayer microfluidic devices. Journal of Visualized Experiments: JoVE, (119), e55276 10.3791/55276.

Chen D, Orenstein Y, Golodnitsky R, Pellach M, Avrahami D, Wachtel C, et al. (2016). SELMAP—SELEX affinity landscape MAPping of transcription factor binding sites using integrated microfluidics. Scientific Reports, 6(1), 33351 10.1038/srep33351. [PubMed: 27628341]

Fordyce PM, Gerber D, Tran D, Zheng J, Li H, DeRisi JL, et al. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nature Biotechnology, 28(9), 970–975. 10.1038/nbt.1675.

Fried MG (1989). Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. Electrophoresis, 10(5–6), 366–376. 10.1002/elps.1150100515. [PubMed: 2670548]

Geertz M, Shore D, & Maerkl SJ (2012). Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. Proceedings of the National Academy of Sciences of the United States of America, 109(41), 16540–16545. 10.1073/pnas.1206011109. [PubMed: 23012409]

Grant CE, Bailey TL, & Noble WS (2011). FIMO: Scanning for occurrences of a given motif. Bioinformatics (Oxford, England), 27(7), 1017–1018. 10.1093/bioinformatics/btr064.

Hellman LM, & Fried MG (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nature Protocols, 2(8), 1849–1861. 10.1038/nprot.2007.249. [PubMed: 17703195]

Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Research, 20(6), 861–873. 10.1101/gr.100552.109. [PubMed: 20378718]

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. (2013). DNA-binding specificities of human transcription factors. Cell, 152(1–2), 327–339. 10.1016/j.cell.2012.12.009. [PubMed: 23332764]

Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nature Methods, 9(1), 72–74. 10.1038/nmeth.1778.

Le DD, Shimko TC, Aditham AK, Keys AM, Longwell SA, Orenstein Y, et al. (2018). Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. Proceedings of the National Academy of Sciences of the United States of America, 115(16), E3702–E3711. 10.1073/pnas.1715888115. [PubMed: 29588420]

Maerkl SJ, & Quake SR (2007). A systems approach to measuring the binding energy landscapes of transcription factors. Science, 315(5809), 233–237. 10.1126/science.1131007. [PubMed: 17218526]

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, et al. (2004).Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genetics, 36(12), 1331–1339. 10.1038/ng1473. [PubMed: 15543148]

Orenstein Y, & Shamir R (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. Nucleic Acids Research, 42(8), e63 10.1093/nar/gku117. [PubMed: 24500199]

Rockel S, Geertz M, & Maerkl SJ (2012). MITOMI: A microfluidic platform for in vitro characterization of transcription factor-DNA interaction. Methods in Molecular Biology (Clifton, N.J.), 786, 97–114. [Chapter 6]. 10.1007/978-1-61779-292-2_6.

Rube HT, Rastogi C, Kribelbauer JF, & Bussemaker HJ (2018). A unified approach for quantifying and interpreting DNA shape readout by transcription factors. Molecular Systems Biology, 14(2), e7902 10.15252/msb.20177902. [PubMed: 29472273]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Schneider TD, & Stephens RM (1990). Sequence logos: A new way to display consensus sequences. Nucleic Acids Research, 18(20), 6097–6100. [PubMed: 2172928]

Spinner DS, Liu S, Wang S-W, & Schmidt J (2002). Interaction of the myogenic determination factor myogenin with E12 and a DNA target: Mechanism and kinetics. Journal of Molecular Biology, 317(3), 431–445. 10.1006/jmbi.2002.5440. [PubMed: 11922675]

Stormo GD, Schneider TD, Gold L, & Ehrenfeucht A (1982). Use of the "perceptron" algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Research, 10(9), 2997–3011. 10.1093/nar/10.9.2997. [PubMed: 7048259]

Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nature Biotechnology, 31(2), 126–134. 10.1038/nbt.2486.

Zhao Y, Granas D, & Stormo GD (2009). Inferring binding energies from selected binding sites. PLoS Computational Biology, 5(12), e1000590 10.1371/journal.pcbi.1000590. [PubMed: 19997485]

Zhao Y, & Stormo GD (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nature Biotechnology, 29(6), 480–483. 10.1038/nbt.1893.

Zuo Z, & Stormo GD (2014). High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. Genetics, 198(3), 1329–1343. 10.1534/genetics.114.170100. [PubMed: 25209146]

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

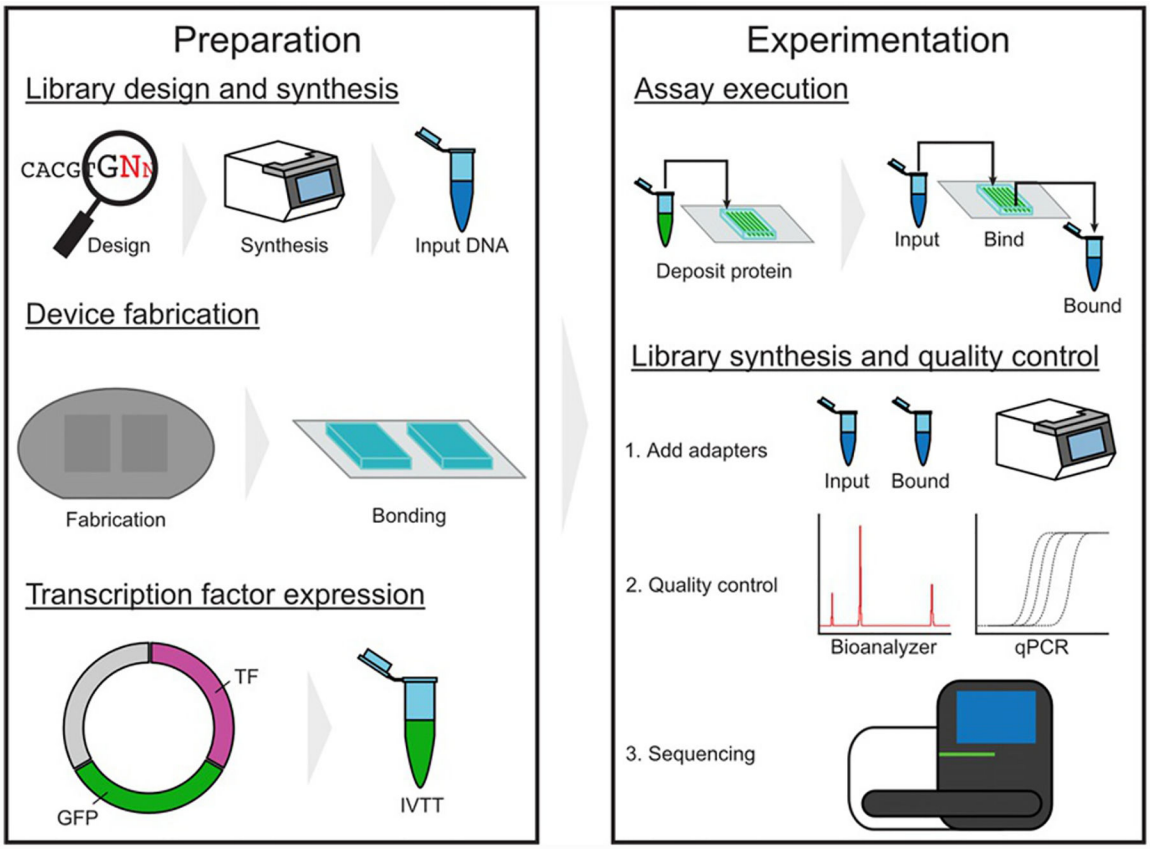

FIG. 1.

BET-seq experimental overview. Preparation steps include library design, amplification, and duplexing; fabrication of PDMS MITOMI microfluidic devices; and cloning to create a GFP fusion of the transcription factor of interest. Experimentation steps include assay execution, followed by addition of sequencing adapters and indices, assessing sample quality, and high-throughput sequencing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

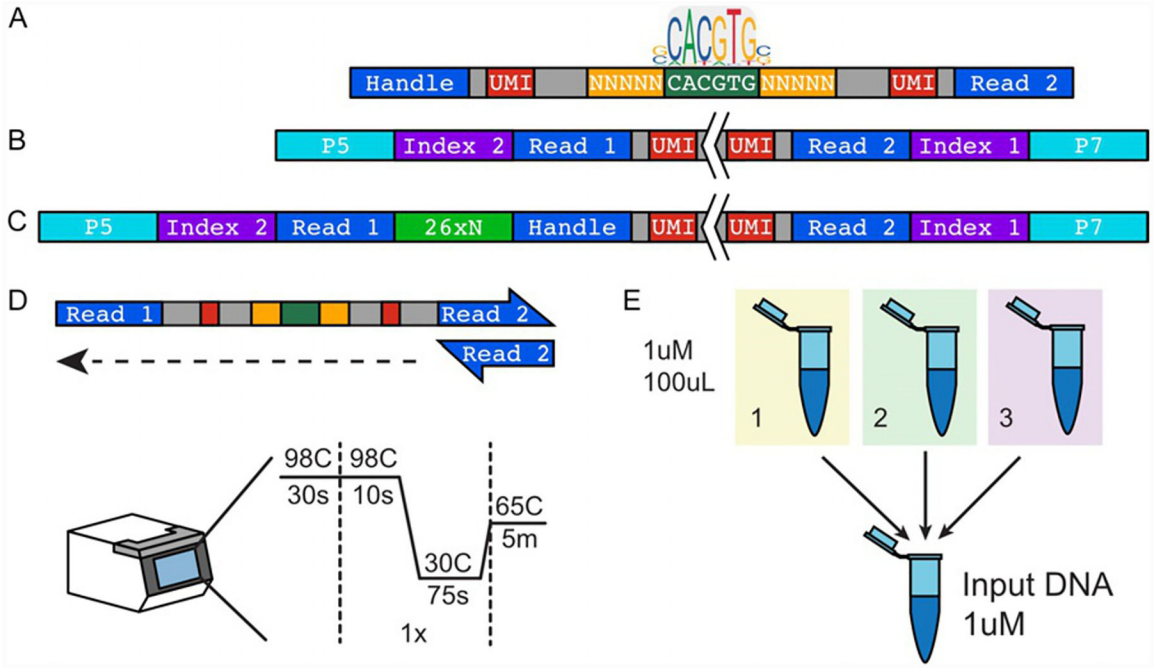

FIG. 2.

BET-seq library design. (A) Full design of BET-seq library oligo showing flanking nucleotide library surrounding known E-box binding site (CACGTG), PCR handles used for PCR-based addition of sequencing adapters (Read 1/Read 2/Handle), unique molecular identifiers (UMIs) for deduplication, and linker sequences. (B) Required oligo library design for sequencing via MiSeq. (C) Required oligo library design for sequencing via NextSeq, with 26 N randomer included to introduce library complexity. (D) Experimental schematic and thermocycler protocol for duplexing ultramer for use in BET-seq. (E) Equimolar combination of multiple sub-libraries into a final Input DNA library for BET-seq experiments.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

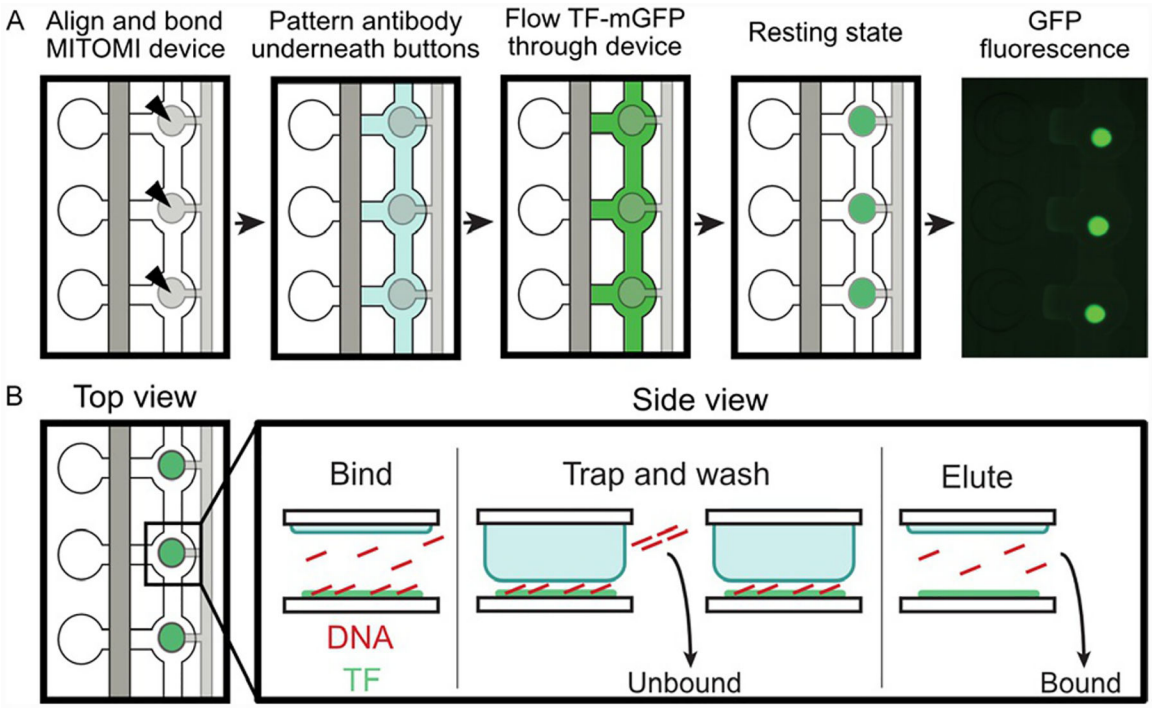

FIG. 3.

MITOMI experimental protocol and quality control. (A) Workflow for selective protein deposition underneath pneumatic "button" valves (indicated by black arrows in leftmost panel) of a MITOMI device. Example final mGFP fluorescence following protein deposition (rightmost panel). (B) Side view schematic of BET-seq assay protocol with DNA and TF of interest. Bound and unbound fractions of DNA are labeled.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

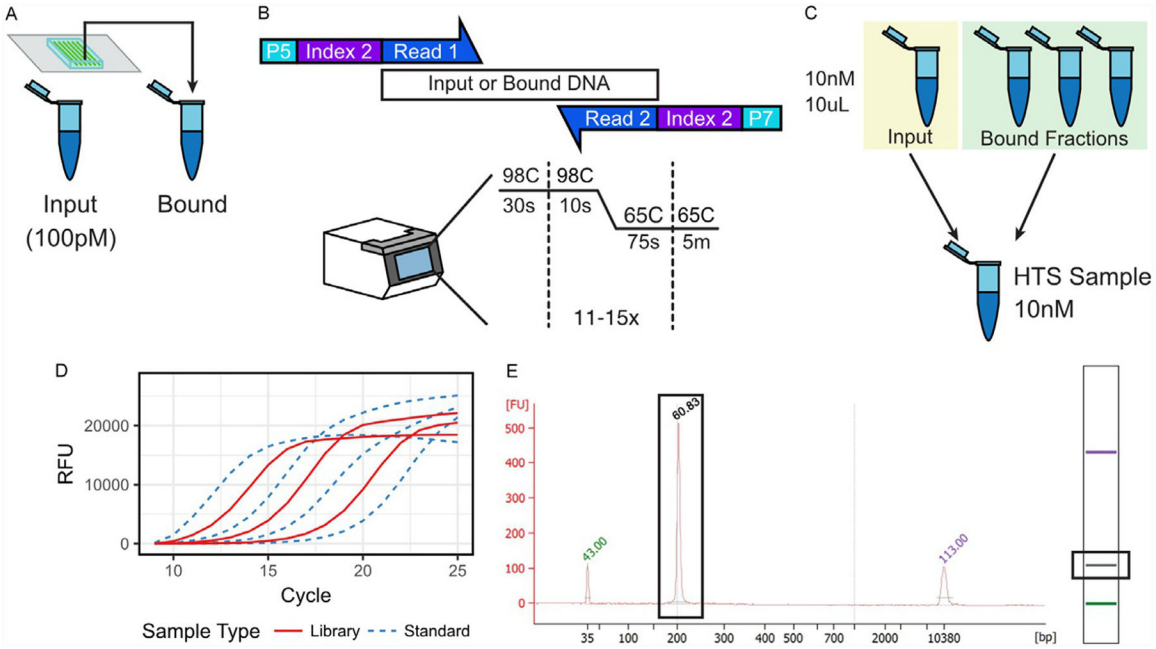

FIG. 4.

Post-experiment quality control and sequencing preparation. (A) Input DNA (at 100 pM) and bound fraction from BET-seq experiment to which Illumina sequencing and indexes adapters are added via PCR. (B) Workflow schematic depicting adapter addition and PCR protocol. (C) Equimolar combinations of indexed Input fraction and multiple uniquely indexed Bound fractions for single sequencing submission covering several BET-seq experiments. (D) Example quantification of submission sample concentration via qPCR. Curves for the quantitation standards are shown as dashed lines, curves for a dilution series of the BET-seq library of interest are shown as solid lines. (E) Example Bioanalyzer trace showing single peak at 205 bp for a clean sequencing submission. Low and high molecular weight peaks correspond to internal Bioanalyzer standards.

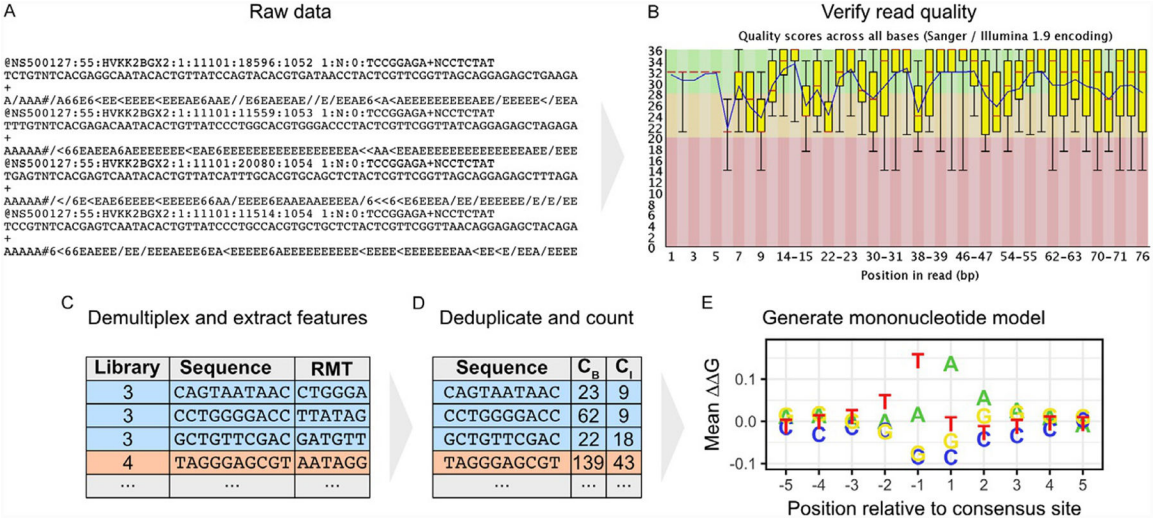

FIG. 5.

An overview of the BET-seq analysis pipeline. (A) FASTQ files are returned from sequencing-based quantitation. (B) These files are then reviewed for read quality using FastQC (an example library of usable quality is shown at top right). (C) Reads are then demultiplexed into bound and unbound read counts and important sequence features, including binding site bases and UMIs, are programmatically extracted. (D) UMIs are then used to deduplicate read counts and ΔΔG values are calculated for each species using the deduplicated counts. (E) A plot of the average mononucleotide effects at each position is shown.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript