Universal correction of enzymatic sequence bias

André L. Martins[1], Ninad M. Walavalkar[1], Warren D. Anderson[1,2], Chongzhi Zang[1,2], Michael J. Guertin[1,2,*]

[1]Biochemistry and Molecular Genetics Department, University of Virginia
[2]Center for Public Health Genomics, University of Virginia
*correspondence

André Martins: alm253@cornell.edu
Ninad Walavalkar: ninad@virginia.edu
Warren Anderson: wa3j@virginia.edu
Chongzhi Zang: zang@virginia.edu
Michael Guertin: guertin@virginia.edu

Abstract

Coupling molecular biology to high throughput sequencing has revolutionized the study of biology. Molecular genomics techniques are continually refined to provide higher resolution mapping of nucleic acid interactions and structure. Sequence preferences of enzymes can interfere with the accurate interpretation of these data. We developed *seqOutBias* to characterize enzymatic sequence bias from experimental data and scale individual sequence reads to correct intrinsic enzymatic sequence biases. *SeqOutBias* efficiently corrects DNase-seq, TACh-seq, ATAC-seq, MNase-seq, and PRO-seq data. Lastly, we show that *seqOutBias* correction facilitates identification of true molecular signatures resulting from transcription factors and RNA polymerase interacting with DNA.

Keywords

transcription, chromatin, DNase-seq, PRO-seq, ATAC-seq, MNase-seq, enzyme preference, bias correction

Background

The field of molecular genomics emerged as classical molecular biology techniques were coupled to high throughput sequencing technology to provide unprecedented genome-wide measurements of molecular features. Molecular genomics assays, such as DNase-seq [1,2], ChIP-exo [3], and PRO-seq [4,5], are converging on single-nucleotide resolution measurements. The enzymes that are routinely used in molecular biology and cloning have inherent and often uncharacterized sequence preferences. These preferences manifest more prominently as the resolution of genomic assays increase. Therefore, we developed *seqOutBias* (https://github.com/guertinlab/seqOutBias) to characterize and correct enzymatic biases that can obscure proper interpretation of molecular genomics data.

Enzymatic hypersensitivity assays, such as DNase-seq [1,2], TACh-seq [6], and ATAC-seq [7], have the potential to measure transcription factor (TF) binding sites genome-wide in a single experiment. These assays strictly measure enzymatic (DNase, Tn5 transposase, Benzonase, or Cyanase) accessibility to DNA, and not a specific biological event, so these data can be challenging to deconvolve. Standard algorithms scan for footprints, which are depletions of signal in larger regions of hypersensitivity [8–12]. Many transcription factors, however, do not exhibit composite footprints if enzymatic cut frequency is averaged at all ChIP-seq validated binding sites with strong consensus motifs [10–12]. Moreover, the inability to detect a footprint at any individual TF binding site results in high false negative rates for footprinting algorithms [13]. Accurate footprinting is also confounded by the artifactual molecular signatures that result from enzymatic sequence preference. DNase footprinting algorithms can incorporate DNase cut preference data to abrogate this bias [12]. However, no existing tools specialize in correcting intrinsic sequence bias for a diverse set of enzymes and experimental methodologies. Herein, we find that correcting for enzymatic sequence bias highlights true molecular signatures that result from TF/DNA interactions. Despite the limitations of enzymatic hypersensitivity footprinting and sequence bias signatures, hypersensitive regions reveal a near-comprehensive set of functional regulatory regions in the genome [14]. Therefore, we present *seqOutBias*, which calculates sequence bias from an aligned BAM file and corrects individual reads accordingly. While this software does not directly infer transcription factor binding, correction of sequence bias provides a more accurate measurement of three key features of enzymatic hypersensitivity data: 1) raw peak height; 2) footprint depth; and 3) true molecular signatures. These measurements, taken together with DNA sequence, can be used to develop algorithms that infer TF binding genome-wide. Moreover, footprint depth and the presence of true molecular signatures are unique to each TF and these features should be characterized for each TF using corrected data in order to optimize these algorithms.

Enzymatic sequence biases are most well-characterized for DNase-seq experiments [10–12], but nearly all molecular genomics experiments employ enzymatic treatments and these enzymes also have intrinsic biases. Herein, we show that DNase, Cyanase, Benzonase, MNase, Tn5 transposase, and T4 RNA ligase all exhibit sequence preferences that are effectively corrected with *seqOutBias*. We also characterize enzymatic bias that results from T4 DNA Polymerase, T4 Polynucleotide Kinase, and Klenow Fragment (3'→5' exo-) treatment of DNA in preparation of high throughput sequencing libraries. Lastly, we show that correction of enzymatic sequence bias highlights true molecular signatures, such as sharp peaks of hypersensitivity and footprints, that result from protein/DNA interactions.

Results

The computational workflow of seqOutBias

Enzymes that are commonly used in molecular biology have nucleic acid preferences for their substrates and the sequence at or near the active site of the enzyme typically dictates enzymatic preference. The *seqOutBias* software aims to correct sequence biases by scaling the aligned read counts by the ratio of genome-wide observed read counts to the expected sequence counts for each k-mer. In *seqOutBias*, the k-mer sequence recognized by the enzyme to confer specificity, is characterized by three parameters: k-mer size and a pair of offsets for the plus and minus strands (Figure 1A). These parameters enable flexibility and *seqOutBias* works with enzymes that have a variety of recognition site lengths. For situations where some base positions surrounding the first sequenced nucleotide do not contribute to site recognition, it is possible to specify the *k-mer mask* parameter (Figure 1A). Positions that should be ignored are represented by an *X* and informative positions by an *N*. This parameter provides an alternative way to specify the position intervening between the first base sequenced and the base directly upstream by inserting a *C* in the mask string. For example, a possible 8-mer that spans 16 bp could be represented as *NNXXNNXXCXXNNXXNN*; likewise, *NNNCNNN*, would represent a recognition site with *kmer-size* = 6, *plus-offset* = 3 and *minus-offset* = 3 (Figure 1A).

In the implementation of *seqOutBias*, our algorithm first calculates the expected sequence detection frequency for each k-mer by determining the positions in the reference genome that are uniquely mappable for a given sequence read length. Due to the large size of genomic datasets, *seqOutBias* reads compressed FASTA files. *SeqOutBias* invokes GenomeTools' *tallymer* to compute mappability at each position in the genome [15,16]. The *tallymer* subcommand computes the mappability information, parses the reference sequence to compute k-mer indexes, and creates a mappability file for a given read length (Figure 1B). This process consists of three parts: 1) creating a suffix tree; 2) creating a genome index; and 3) creating the mappability file. These processes are the most computationally intensive steps and *seqOutBias* will recognize the existence of intermediate files in the directory to avoid unnecessary recomputation. For instance, if the *seqOutBias tallymer* step is executed for different read lengths, but using the same FASTA file, then the first suffix-tree portion is re-used across invocations.

The next step, *seqOutBias seqtable*, creates an intermediate table that combines mappability information, read length, and plus/minus offsets (Figure 1A) to map k-mer indexes to the aligned read positions (Figure 1B). This intermediate file reduces the amount of computation needed when processing aligned read files and provides an intermediate file that decouples the reference sequence processing from the remaining steps. The resultant TBL file is an input for the *seqOutBias tabulate* subcommand, this subcommand produces a k-mer count table based on the TBL sequence information and the optional sorted BAM file. Counts correspond to the entire genome by default, but counts can be constrained to specific regions by supplying a BED file with the *regions* option. When no BAM file is supplied, the output will have four columns: k-mer index, k-mer string, plus strand count, and minus strand count. If a BAM file is supplied, the output will have two additional columns with the plus and minus strand counts of observed aligned reads.

The final subcommand, *seqOutBias scale*, produces the corrected aligned read pile-ups, both as BED and bigWig files. This command provides flexibility in the output, with options to shift the minus strand reads to align with the plus strand reads (Figure 1A). This option is used when enzymatic cleavage of individual sites can result in a single base shift depending on whether the nicking event was detected by sequencing the upstream or downstream DNA (red nucleotides in Figure 1A). The *tail-edge* option outputs the 3´ end of the reads; this option is used primarily for analysis of PRO-seq data [4,5]. Therefore, *seqOutBias* reads compressed files (FASTA, mappability information, and sorted BAM files), reuses intermediate results, and allows for flexibility in specifying sequence features for data correction.

Correction of individual DNase-seq reads

DNase-seq measures the accessibility of the phosphodiester backbone of DNA at single-nucleotide resolution [1,2,9,17]. Composite DNase-seq profiles centered on sequence motifs of TF binding sites accentuate molecular features that inform on TF binding properties. DNase footprints are defined as depletions of sensitivity within large regions of hypersensitivity; footprints align with TF recognition sites and result from TF interactions with DNA [18,19]. High throughput DNase-seq experiments described a cleavage pattern at the footprint that was interpreted as a measure of TF/DNA interactions [9]; however, subsequent work attributed these artifactual signatures to differential substrate specificity of DNase conferred by the presence of the TF motif [10–12]. As a result, some footprint detection programs now incorporate sequence biases into their algorithms [12,20]. *SeqOutBias* provides the option to correct enzymatic sequence bias prior to footprint detection and the output files can be used with existing footprinting algorithms that do not incorporate a correction step.

We scaled individual reads based on the preference of DNase using *seqOutBias* and a 6-mer correction factor (Figure 2). Figure 2 illustrates that DNase prefers to nick the sequence *CCTTGC* and the read associated with this window was reduced to an intensity of 0.15. DNase disfavors nicking of *GGGGAA*, thus the read associated with this hexamer was scaled to an intensity of 5.6. DNase sequence preferences are most apparent in composite profiles of DNase cut frequency surrounding TF motifs. We tested the efficacy of 6-mer correction on DNase-digested naked DNA [21]; corrected profiles of naked DNA digestion should not exhibit footprints or molecular signatures that result from protein/DNA interactions. We observe that sharp peaks and troughs are smoothed in the corrected composite profiles for ELF1, GATA3, and MAX motifs (Figure 3).

True signatures that result from TF/DNA interactions are not smoothed by *seqOutBias*. For instance, ChIP-seq validated CCCTC-binding factor (CTCF) binding sites exhibit strong footprints and composite profiles at CTCF motifs highlight a sharp signature upstream of CTCF binding [22]. This CTCF signature is unaffected after correcting for DNase intrinsic sequence preference [12]. We plotted DNase-seq profiles at GATA3 and MAX binding sites to determine whether true molecular signatures are apparent after intrinsic bias correction (Figure 4). We observe a clear composite footprint at MAX binding sites in chromatin, as expected, this footprint is not present in the naked DNA digestion (Figure 4A). The MAX footprint is obscured by sharp peaks of hypersensitivity (sequence artifact signatures) in the raw uncorrected traces (Figure 4A). We observe a sharp DNase signature upstream of GATA3 binding sites, which is present only in the chromatin digested samples (Figure 4B). We conclude that this molecular signature is a result of GATA3/DNA interactions, because this peak is neither smoothed following *seqOutBias* correction nor present in the naked DNase digested sample. Note that GATA3 does not have an appreciable composite footprint, but TF inference algorithms may use TF-specific signatures, as we observe for GATA3, to inform on TF occupancy and binding intensity. Therefore, correction of intrinsic DNase sequence bias highlights true molecular features: footprints and sharp hypersensitivity peaks. We propose that these features can be systematically characterized for all TF and used as informative priors when inferring TF binding profiles genome-wide from enzymatic hypersensitivity data.

Correction of TACh-seq, MNase-seq, ATAC-seq, and PRO-seq data

We characterized and corrected the biases of Benzonase and Cyanase using Tissue Accessible Chromatin (TACh-seq) data [6]. TACh-seq is a variant of traditional enzymatic hypersensitivity assays whereby frozen tissue samples are treated with either Benzonase or Cyanase endonuclease. Benzonase is an endonuclease cloned from *Serratia marcescens* that functions as a dimer and Cyanase is a non-*Serratia* monomeric enzyme. These enzymes are more highly active under high salt and high detergent conditions, so these enzymes are more suited for digestion of solid tissue sample, which requires harsh dissociation treatments. We corrected TACh-seq data generated from frozen mouse liver tissue [6]. Composite profiles from CEBP-beta, FOXA2, and CTCF binding sites in mouse liver indicate that an eight base pair mask centered on the nick site is sufficient to correct both Cyanase and Benzonase biases (Figure S1 and Figure S2). Next, we applied *seqOutBias* correction to MNase-seq data generated from MCF-7 cells [23]. An eight base pair mask abrogates the intrinsic sequence bias of MNase-seq data (Figure S3).

ATAC-seq is unique among enzymatic accessibility assays because each transposition event inserts two adapters into the chromatin. Each Tn5 molecule can be pre-loaded with any combination of the paired-end 1 and paired-end 2 adapter. Reads that align to the plus and minus strand are processed separately because the Tn5 recognition site is distinct for plus and minus reads. We applied *seqOutBias* correction to published ATAC-seq data from GM12878 cells [7]. We generated and analyzed naked DNA libraries using the ATAC-seq work flow to measure Tn5 specificity in the absence of chromatin (GEO accession: GSE92674). We optimized the k-mer mask for ATAC-seq data; *N* positions of *NXXNNCNNXNNN* are the most influencial for Tn5 recognition of plus strand reads and *NNNXNNCNNXXN* is the optimal mask for minus strand reads. The sharp ATAC-seq spikes at the site of TF binding are reduced in the corrected data (Figure S4 and Figure S5). The complex nature of Tn5 recognition and dual loading of adapters, taken together with the incomplete smoothing of ATAC composite profiles, suggests that a simple spaced k-mer correction may not be sufficient to fully correct Tn5 bias.

PRO-seq couples terminating nuclear run-on assays with high throughput sequencing to quantify engaged RNA polymerase molecules genome-wide at nucleotide resolution [4]. Sequence composition of transcripts may affect run on efficiency, therefore, the sequence immediately downstream of RNA polymerase may influence detection of RNA molecules. The sequence upstream of RNA polymerase could affect ligation efficiency because T4 RNA ligase treatment may exhibit sequence preference. We used *seqOutBias* to scale published PRO-seq data from K562 cells [24]. We specifically used annotated transcripts to calculate expected k-mer frequency, as opposed to genomic k-mer frequency, because the vast majority of transcription occurs within gene annotations [25]. We found that a k-mer mask that spans the last three bases of the ligated RNA molecule and the three bases downstream from RNA polymerase is sufficient to correct the PRO-seq data (Figure 5 and Figure 6). RNA polymerase density decreases at the polypyrimidine tract upstream of the 3´splice site and we observe a sharp peak at position -3 from the 5´ end of exon. This sharp peak is absent using *seqOutBias*-corrected reads (Figure 5), therefore we propose that this peaks results from either inefficient adenine incorporation during the nuclear run-on or a preference for cytosine or uracil during either the run-on or ligation reaction.

Genome-wide binding data for CTCF is available for K562, GM12878, mouse liver, and MCF-7 cells. Upon correcting for enzymatic sequence bias, the sharp signature artifacts at CTCF motifs are abrogated in each molecular genomics dataset we tested (Figure 6). The naked DNA profiles for ATAC-seq and DNase-seq are not restricted to CTCF-bound sites, all genomic CTCF motifs are included in these composites (Figure 6). In the chromatin TACh, DNase, and ATAC experiments, we observed protection resulting in a footprint and a sharp peak upstream of the CTCF motif. Taken together, we show that *seqOutBias* effectively corrects enzymatic sequence bias resulting from a diverse set of molecular genomics experiments.

Enzymatic DNA end repair and ligation bias

We found that the bases upstream and downstream of a DNase nick site are not equally likely to be detected by sequencing (red nucleotides in Figure 1A). In Figure 1A, for *GATGTC* we would expect the ratio of reads that begin with *GACCAGATGACA* (plus strand) and *ATCATATCCCGT* (minus strand) to be approximately equal to one if this site was nicked repeatedly and there was no enzymatic end repair and ligation bias. We performed this analysis for all instances of each *NNNGAC*-mer in the genome (Figure 7A) and all 4096 pairwise combinations of 3-mers (Figure 7B). Reverse palindromic 6-mers are balanced (Figure

7B), but for most 3-mer combinations we identified a preference for which 3-mer is detected by sequencing, we term this “detection bias.”

Preparing digested DNA for Illumina high throughput sequencing requires several enzymatic treatments. T4 DNA Polymerase treatment removes 3´ overhangs and fills in 3´ recessed (5´ overhang) ends. T4 Polynucleotide kinase phosphorylates the 5´ end and Klenow Fragment (3´ to 5´ exo-) adds a single 3´ adenine overhang. We hypothesized that the overhanging sequences dictate the detection bias, because the detection bias is distinct for Benzonase and DNase (Figure 7C top panel). Although four nick events are necessary to sequence a DNA molecule, enzymatic hypersensitivity assays only detect one nick on each end of the molecule and it is impossible to determine the precise location of the other nicks. By assuming that two enzymes with similar nick specificity (Figure S6) will have comparable distribution of sequence overhangs, we can test the hypothesis that the overhang sequences contribute to post-nicking enzymatic treatment biases. We compared this post-nicking bias using DNase-seq data from two different labs and two different organisms (Figure S6). We also compared the detection bias of Cyanase and Benzonase, which have similar sequence preferences (Figure S6). Indeed, digestions with enzymes that have similar nick preferences, which results in comparable distributions of overhanging sequences, have highly correlated detection biases (Figure 7C bottom two panels and Figure S7). Importantly, *seqOutBias* calculates the ratio of genomic k-mers and experimentally observed k-mers to scale individual reads and this calculation inherently corrects for the convolution of biases resulting from multiple enzymatic steps.

Discussion

We previously described the challenge of interpreting single-nucleotide resolution DNase-seq data [10,11]. Subsequently, groups have developed algorithms that consider this bias for DNase-seq footprinting detection [12,20]. However, this is the first report of stand-alone software that specializes in correcting sequence bias for a diverse set of molecular genomics datasets. *SeqOutBias* is a command line tool and designed for a UNIX environment, making the software compatible for seamless integration into existing high throughput sequencing analysis pipelines. *SeqOutBias* is conceptually and mathematically simple, effectively counting k-mer occurrences and scaling data accordingly. This calculation sufficiently corrects biases associated with many different assays. However, we anticipate that subsequent software may incorporate more complex calculations and models into data correction. For instance, RNA hairpins may affect the efficiency of ligating adapter to RNA using T4 RNA ligase. Due to the complexity of secondary and post-secondary RNA structure predictions [26], we suspect that more sophisticated models are necessary for correction of datasets such as PRO-seq.

Enzymatic hypersensitivity assays have the potential to identify regulatory elements genome-wide and infer TF binding intensity at each regulatory element. Four features of enzymatic hypersensitivity assays can aid in TF binding inference: 1) the presence of a TF’s recognition motif; 2) the raw enzyme cleavage frequency in the region surrounding the motif; 3) a depletion in sensitivity at the motif (footprint); and 4) the presence of TF-mediated molecular signatures (sharp peaks and valleys) that surround the motif. Correction of enzymatic sequence bias provides a more accurate measurement of all these features except sequence composition. Correction of intrinsic experimental biases will prove important as the field continues to refine experiments and algorithms to more accurately infer TF binding intensity genome-wide from enzymatic hypersensitivity data.

Conclusion

We and others have previously shown that enzymatic sequence preferences can be misinterpreted as biologically important phenomena [10–12]. Sequence bias correction is an important step in analyzing high resolution molecular genomics data and we introduce *seqOutBias* as flexible and novel software that efficiently characterizes biases and appropriately scales individual sequence reads.

Methods

Installation and analyses

The user guide and install instructions are available through GitHub: https://guertinlab.github.io/seqOutBias/seqOutBias_user_guide.pdf.

The analyses presented herein are reproduced in full with rationale in the accompanying *seqOutBias* vignette on GitHub: https://guertinlab.github.io/seqOutBias/seqOutBias_vignette.pdf.

Deproteinized DNA ATAC-seq

The naked DNA ATAC-seq library was prepared as previously described [7] with several modifications: 1) we used purified genomic DNA, as opposed to crude nuclei isolations; 2) we omitted IGEPAL CA-630 from all buffers; and 3) we performed PCR cleanup using AMPure XP beads to select DNA <600 bp. The naked DNA ATAC-seq data were deposited in the Gene Expression Omnibus (GEO) database, with accession number GSE92674.

Figure Legends

Figure 1. *SeqOutBias* overview and parameter definitions. A) An enzymatic cleavage event that results in a blunt end can be detected by sequencing the upstream or downstream DNA (red bases). The hexamer sequence centered (red block) on the nick sites (dotted vertical lines) confers specificity; this parameter is referred to as the k-mer. The plus-offset and minus-offset parameters specify the nick site relative to the first position and last position of the k-mer. As opposed to specifying the immediate upstream base for the minus strand, we shift the base position by +1 to match the position of the immediate upstream base from the plus aligned read. B) This panel illustrates the high-level overview of the inputs, intermediate files, and output of the *seqOutBias* program and the computation steps that the program performs. The *tallymer* step indexes the reference sequence (FASTA) and computes mappability for the given read length. The *seqTable* step parses the reference sequence (FASTA) together with the mappability information to compute the k-mer that corresponds to each possible read alignment position. The *tabulate* step tallies the k-mer counts across the selected regions (or the full genome), as well as the k-mers corresponding to observed aligned reads (if a BAM file is supplied). Lastly, *scale* computes the genome-wide aligned read pile-ups, scaling sequence reads by the expected/observed cut frequency.

Figure 2. *SeqOutBias* scales individual sequence reads. The bottom track shows six nick positions from DNase-seq data; each position was found once in the data. The top track reports corrected read intensities, which scale inversely with DNase sequence preference.

Figure 3. DNase nick bias is corrected in a naked DNA DNase experiment. Each composite profile illustrates the average cut frequency at each position between nucleotides. The blue trace is the raw data and the black trace is the corrected data; the opaque boundaries represents the 75% confidence interval. A seqLogo representation for each TF's binding site is shown at the top of each plot and vertical dashed lines show the boundaries of sequence information content.

Figure 4. True molecular signatures resulting from TF/DNA interactions are visible in corrected composite profiles. A) A true footprint is highlighted in corrected composite profile (right panel) of DNase cleavage at ChIP-seq confirmed MAX binding sites compared to raw frequency counts (left panel). The black trace is DNase-digested chromatin and the green trace is DNase-digested naked DNA. As expected, the composite footprint is not detected in the naked DNA composite. B) A true molecular signature is highlighted in the corrected composite profile (right panel) of GATA3 binding sites. The signature is exclusively detected in the chromatin digested experiment and may result from GATA3/DNA interaction.

Figure 5. *SeqOutBias* corrects sequence bias at CTCF binding sites associated with DNase, Tn5 Transposase (ATAC),Benzonase (TACh), Cyanase (TACh), MNase, and T4 RNA ligase (PRO). Upon correcting for enzymatic sequence bias, the artifactual spikes at the CTCF binding site are abrogated in each molecular genomics dataset we tested. However, in cases of CTCF binding to chromatin, we observe protection that results in a footprint; note that MNase is not expected to result in a composite footprint. We observe the previously characterized sharp peak upstream of the CTCF motif and this molecular signature is likely caused by CTCF-mediated enhancement of cleavage activity.

Figure 6. *SeqOutBias* corrects sequence bias associated with the 3´ splice site recognition motif. Upon correcting for enzymatic sequence bias, the artifactual signature at the 3´ splice site is abrogated. The first base of the exon spans position 0-1 on the x-axis and the sequence bias peak is at position -3.

Figure 7. Detection biases are highly correlated between enzymes with similar cut specificities, suggesting that ssDNA overhangs drive enzymatic specificity. A) For all sequence-detected DNase-nicked 6-mers that end in *GAC* we compare the ratio of sequence reads that start with *GAC* to the oppositely oriented 3-mer. This bias results from enzymatic end repair and ligation sequence preference during the library preparation. B) The relative bias of all 3-mers sequenced (the ratio of x-axis 3-mer to y-axis 3-mer). C) This figure plots the values from panel B. The post-nick sequence preferences are highly correlated between DNase-seq experiments and between Benzonase and Cyanase experiments, but not between DNase and Benzonase.

Figure S1. *SeqOutBias* corrects cyanase endonuclease bias. Each composite profile illustrates the average cut frequency at each position between nucleotides. The blue trace is the raw data and the black trace is the 8-mer corrected data.

Figure S2. *SeqOutBias* corrects benzonase endonuclease bias. The composite profiles for FOXA2, CTCF, and CEBP-beta binding sites illustrate the average cut frequency at each position between nucleotides. The blue trace is the raw data and the black trace is the 8-mer corrected data.

Figure S3. *SeqOutBias* corrects MNase sequence bias. The composite profiles for MAX, GATA3, and ELF1 indicate that sequence correction abrogates the sharp peaks in the traces. The blue trace is the raw data and the black trace is the 8-mer corrected data.

Figure S4. Tn5 insertion bias is corrected in a ATAC-seq experiment from GM12878 cells. The composite profiles for SP1, EBF1, and REST indicate that sequence correction dampens the sharp peaks in the traces.

Figure S5. Tn5 insertion bias is corrected in a ATAC-seq experiment from naked DNA. We generated ATAC-seq data with naked DNA and we find that the composite profiles for TFs exhibit dampened sharp peaks in the corrected traces.

Figure S6. Enzymatic nick biases are correlated between DNase-seq experiments and correlated between Cyanase and Benzonase digestion experiments. These scatter plot show that the enzymatic nick biases, as measured by the *seqOutBias* scale factor, are correlated between DNase experiments and correlated between Cyanase and Benzonase.

Figure S7. Post-nick enzymatic processing biases of DNase are correlated between experiments and the post nick biases of Cyanase and Benzonase are similar. The relative bias of all 3-mers sequenced (the ratio of x-axis 3-mer to y-axis 3-mer) for four separate experiments.

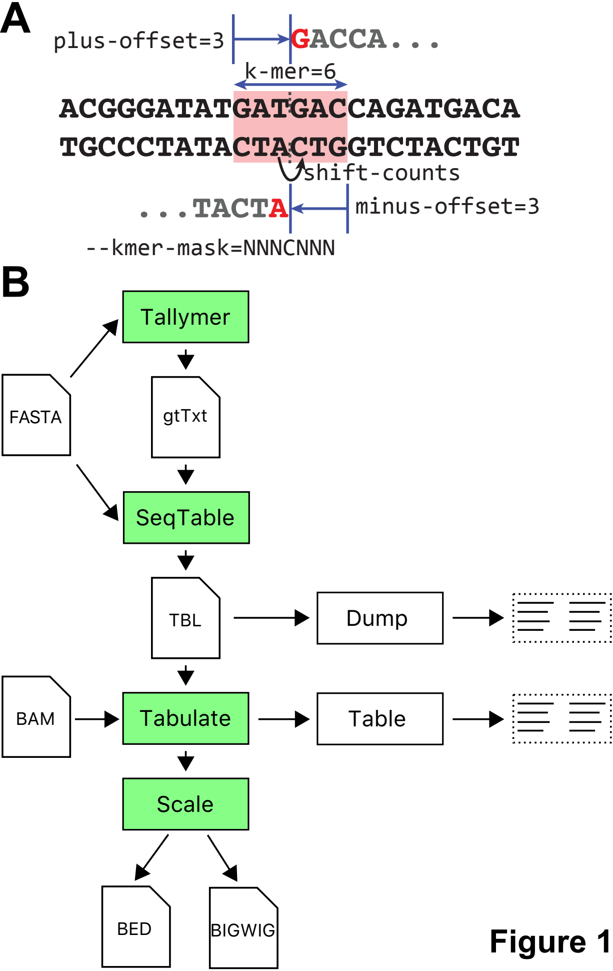

A
plus-offset=3
GACCA...
k-mer=6
ACGGGATATGATGACCAGATGACA
TGCCCTATACTACTGGTCTACTGT
shift-counts
...TACTA
minus-offset=3
--kmer-mask=NNNCNNN
B
Tallymer
FASTA
gtTxt
SeqTable
TBL
Dump
BAM
Tabulate
Table
Scale
BED
BIGWIG

Figure 1

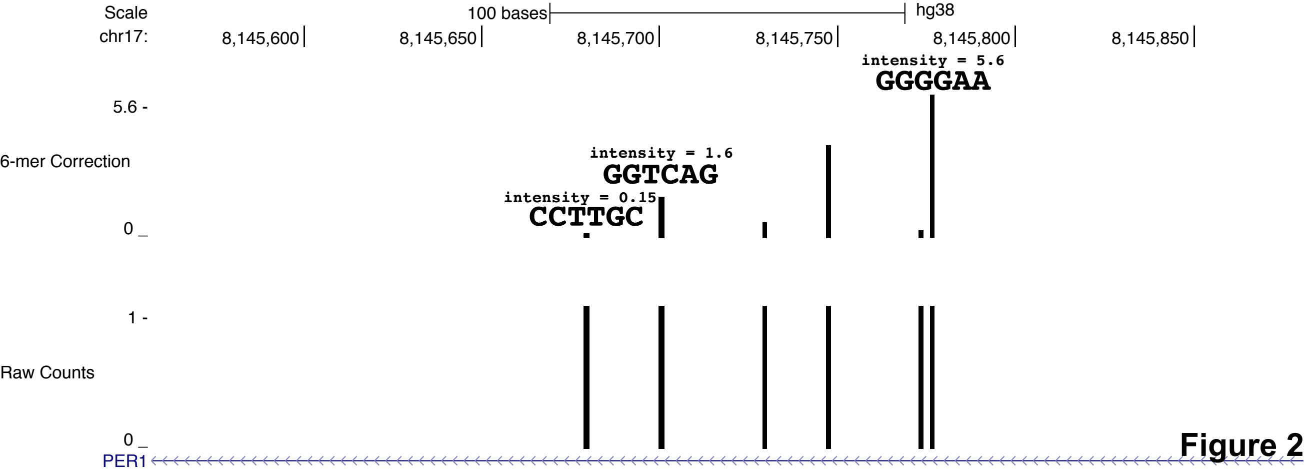

Scale
100 bases
hg38
chr17:
8,145,600
8,145,650
8,145,700
8,145,750
8,145,800
8,145,850
intensity = 5.6
GGGGAA
5.6 -
6-mer Correction
intensity = 1.6
GGTCAG
intensity = 0.15
CCTTGC
0 _
1 -
Raw Counts
0 _
PER1

Figure 2

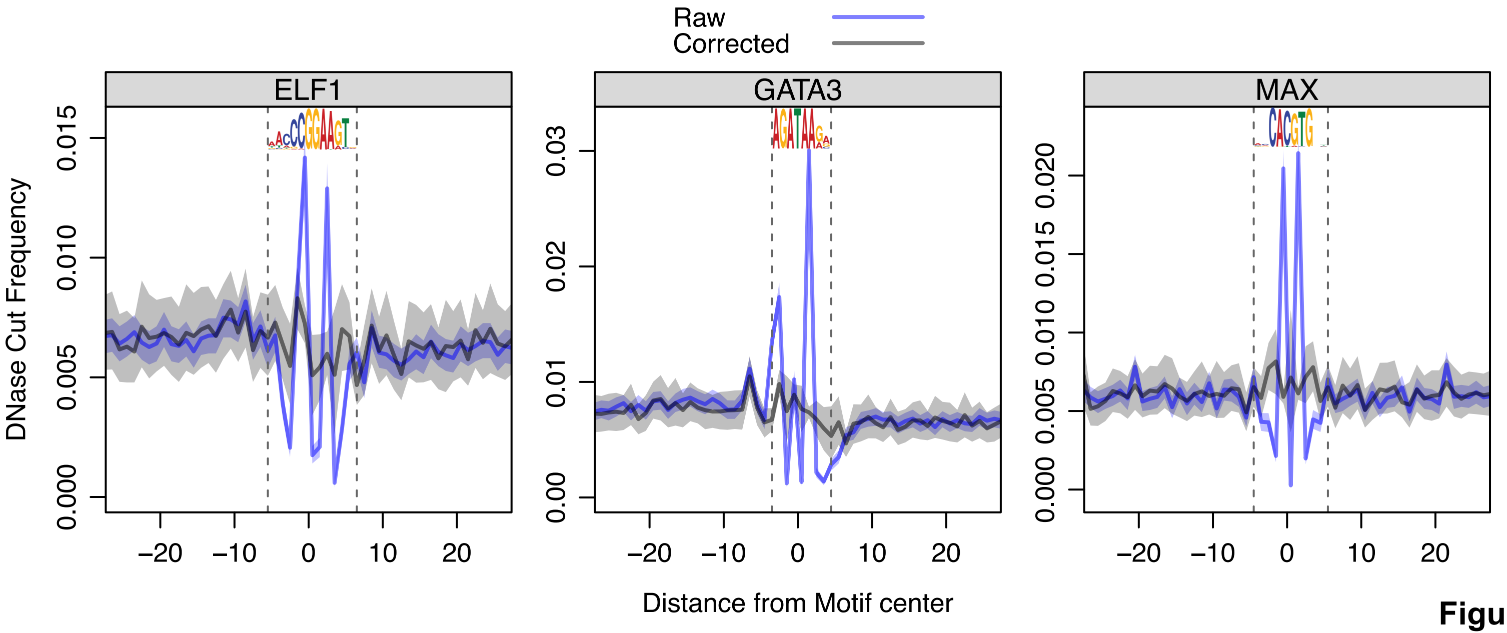

Raw
Corrected
ELF1
GATA3
MAX
DNase Cut Frequency
0.000
0.005
0.010
0.015
0.00
0.01
0.02
0.03
0.000
0.005
0.010
0.015
0.020
−20
−10
0
10
20
Distance from Motif center

Figure 3

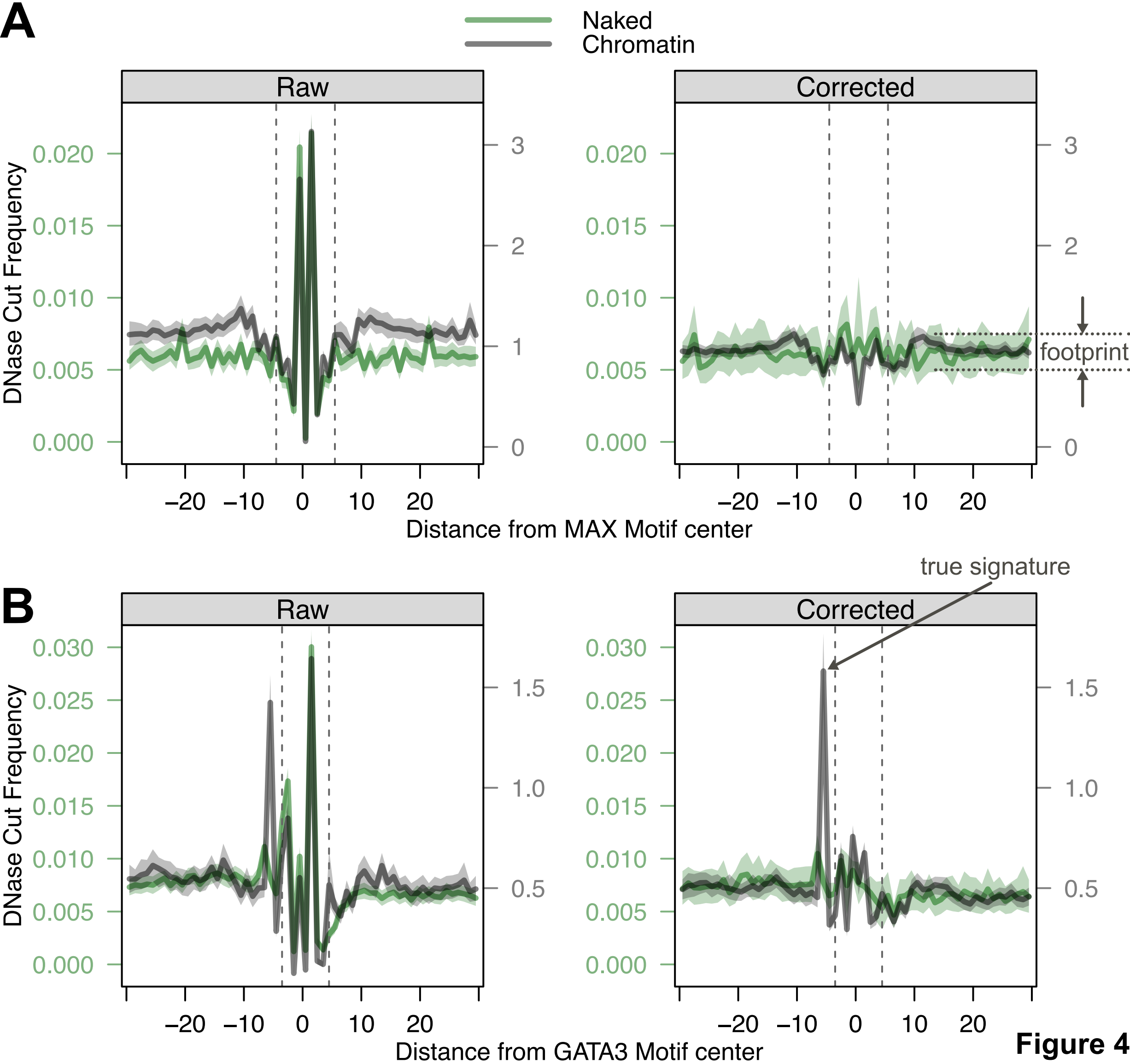
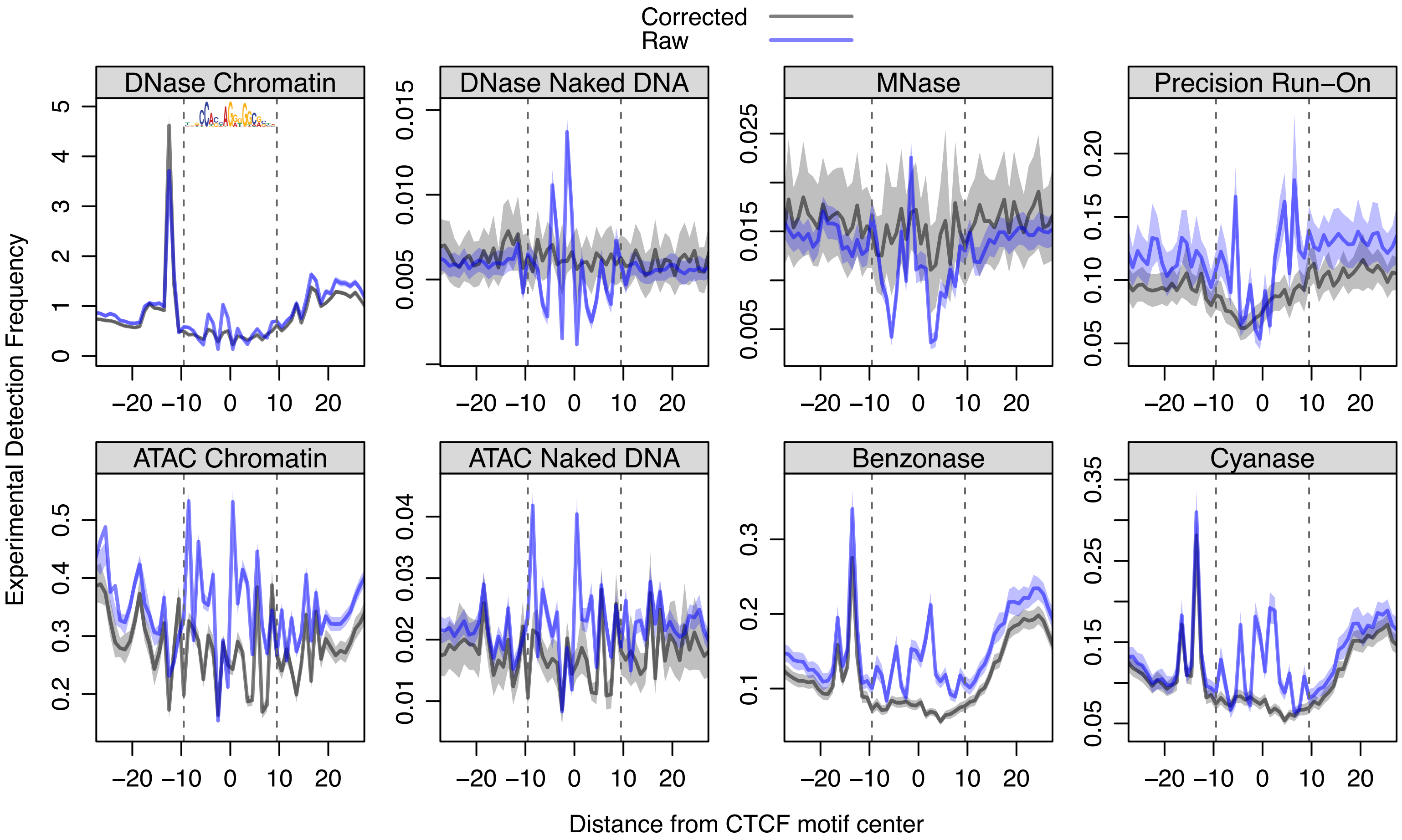

A
Naked
Chromatin
Raw
Corrected
DNase Cut Frequency
0.020
0.015
0.010
0.005
0.000
3
2
1
0
footprint
−20 −10 0 10 20
Distance from MAX Motif center
B
true signature
0.030
0.025
0.020
0.015
0.010
0.005
0.000
1.5
1.0
0.5
Distance from GATA3 Motif center

Figure 4



Corrected
Raw
DNase Chromatin
DNase Naked DNA
MNase
Precision Run−On
ATAC Chromatin
ATAC Naked DNA
Benzonase
Cyanase
Experimental Detection Frequency
Distance from CTCF motif center
−20
−10
0
10
20
0
1
2
3
4
5
0.005
0.010
0.015
0.025
0.05
0.10
0.15
0.20
0.2
0.3
0.4
0.5
0.01
0.02
0.03
0.04
0.1
0.2
0.3
0.05
0.15
0.25
0.35

Figure 5

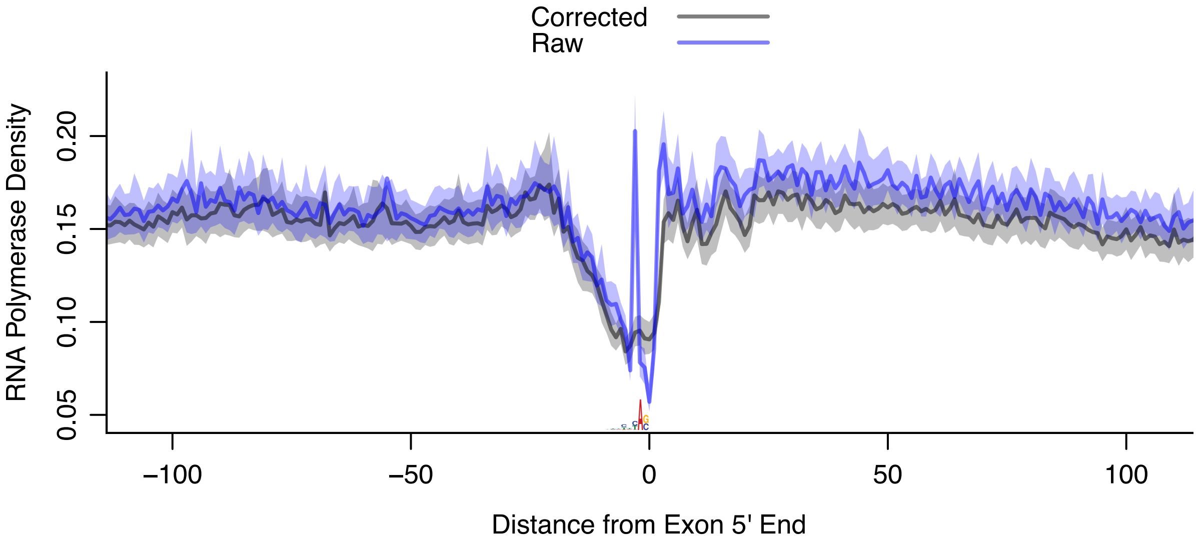

Corrected
Raw
RNA Polymerase Density
0.05
0.10
0.15
0.20
−100
−50
0
50
100
Distance from Exon 5' End

Figure 6

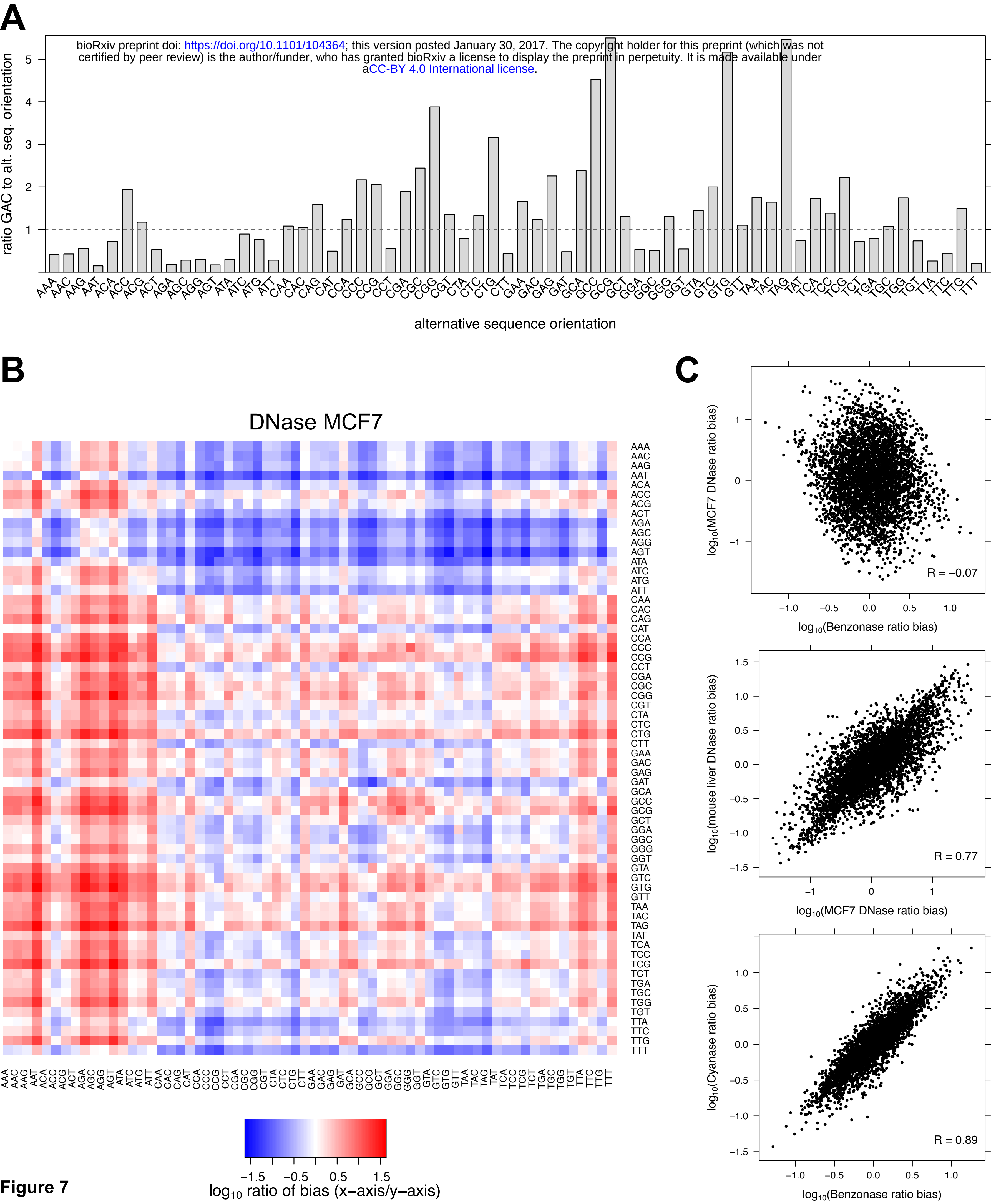

A
ratio GAC to alt. seq. orientation
1
2
3
4
5
AAA AAC AAG AAT ACA ACC ACG ACT AGA AGC AGG AGT ATA ATC ATG ATT CAA CAC CAG CAT CCA CCC CCG CCT CGA CGC CGG CGT CTA CTC CTG CTT GAA GAC GAG GAT GCA GCC GCG GCT GGA GGC GGG GGT GTA GTC GTG GTT TAA TAC TAG TAT TCA TCC TCG TCT TGA TGC TGG TGT TTA TTC TTG TTT
alternative sequence orientation
B
DNase MCF7
AAA AAC AAG AAT ACA ACC ACG ACT AGA AGC AGG AGT ATA ATC ATG ATT CAA CAC CAG CAT CCA CCC CCG CCT CGA CGC CGG CGT CTA CTC CTG CTT GAA GAC GAG GAT GCA GCC GCG GCT GGA GGC GGG GGT GTA GTC GTG GTT TAA TAC TAG TAT TCA TCC TCG TCT TGA TGC TGG TGT TTA TTC TTG TTT
−1.5
−0.5
0.5
1.5
log$_{10}$ ratio of bias (x−axis/y−axis)
C
log$_{10}$(MCF7 DNase ratio bias)
log$_{10}$(Benzonase ratio bias)
R = −0.07
log$_{10}$(mouse liver DNase ratio bias)
log$_{10}$(MCF7 DNase ratio bias)
R = 0.77
log$_{10}$(Cyanase ratio bias)
log$_{10}$(Benzonase ratio bias)
R = 0.89

Figure 7

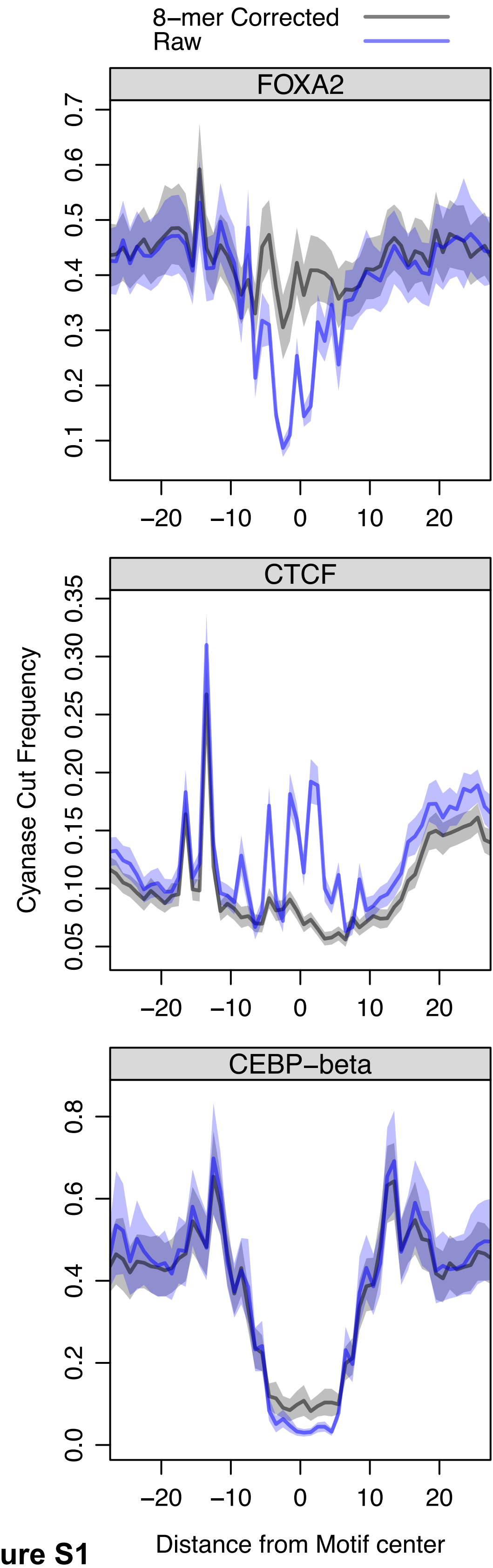

8−mer Corrected
Raw
FOXA2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−20
−10
0
10
20
CTCF
Cyanase Cut Frequency
0.05
0.10
0.15
0.20
0.25
0.30
0.35
−20
−10
0
10
20
CEBP−beta
0.0
0.2
0.4
0.6
0.8
−20
−10
0
10
20
Distance from Motif center

Figure S1

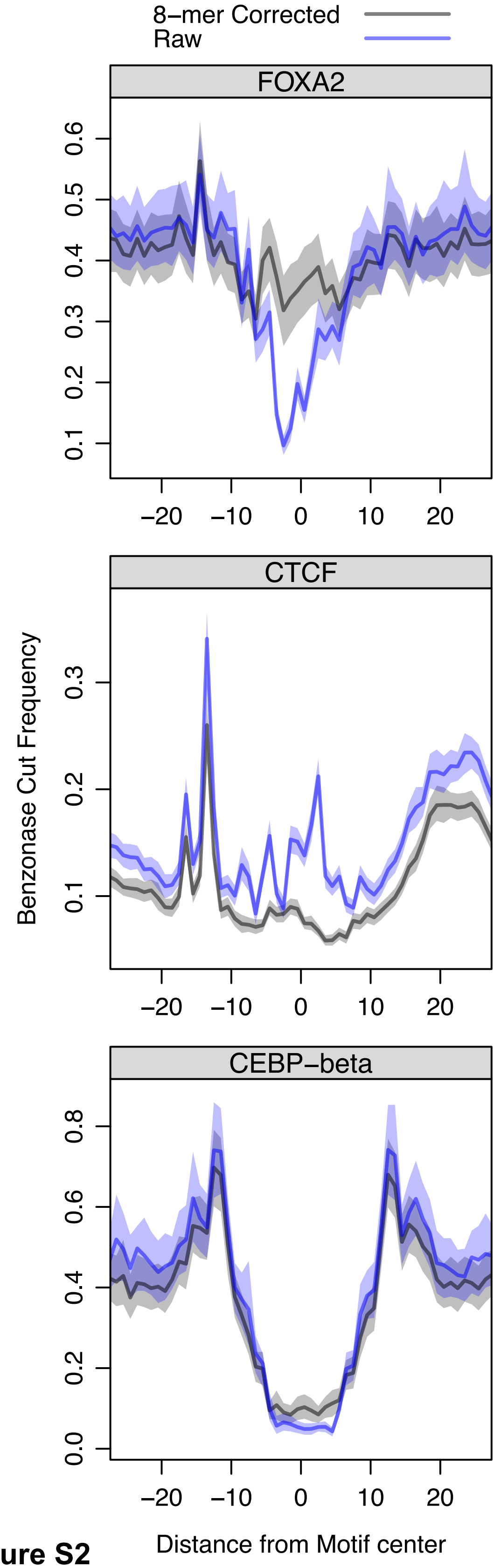

8−mer Corrected
Raw
FOXA2
0.6
0.5
0.4
0.3
0.2
0.1
−20
−10
0
10
20
CTCF
Benzonase Cut Frequency
0.3
0.2
0.1
−20
−10
0
10
20
CEBP−beta
0.8
0.6
0.4
0.2
0.0
−20
−10
0
10
20
Distance from Motif center

Figure S2

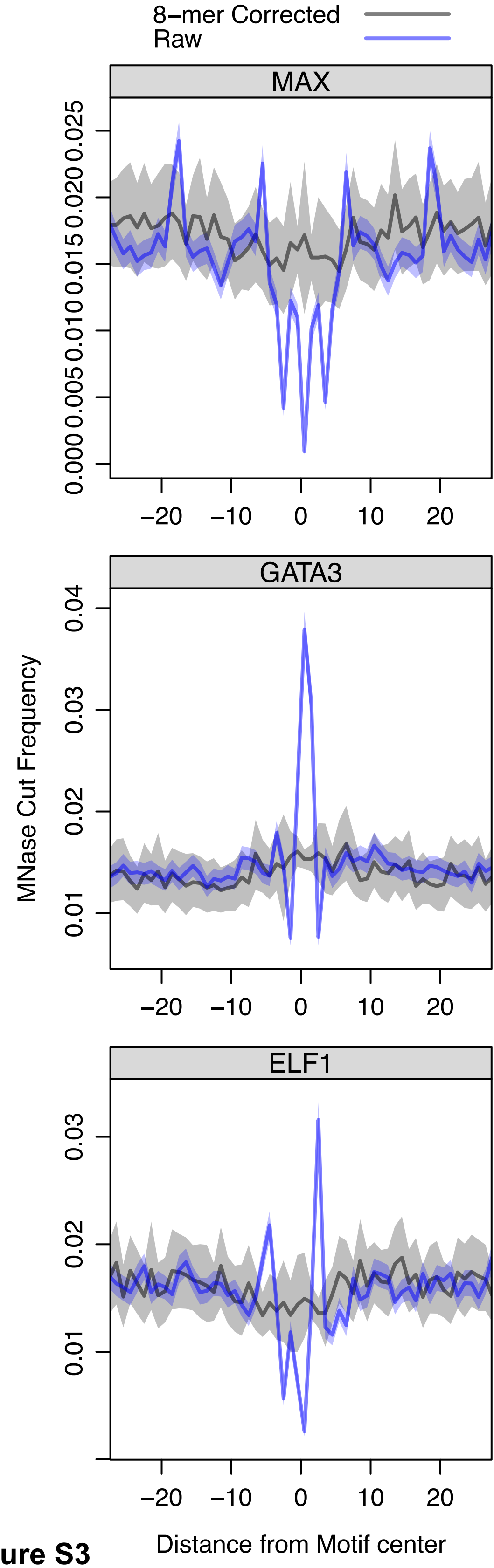

8−mer Corrected
Raw
MAX
0.000
0.005
0.010
0.015
0.020
0.025
GATA3
0.01
0.02
0.03
0.04
ELF1
0.01
0.02
0.03
MNase Cut Frequency
−20
−10
0
10
20
Distance from Motif center

Figure S3

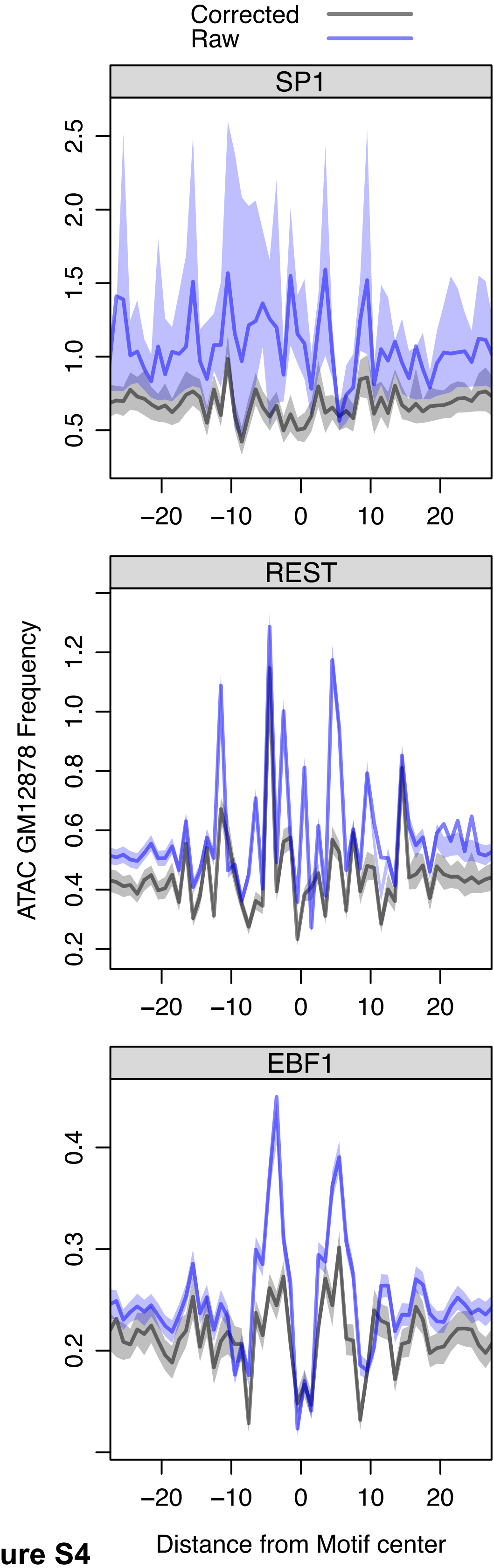

Corrected
Raw
SP1
0.5
1.0
1.5
2.0
2.5
REST
0.2
0.4
0.6
0.8
1.0
1.2
EBF1
0.2
0.3
0.4
ATAC GM12878 Frequency
–20
–10
0
10
20
Distance from Motif center

Figure S4

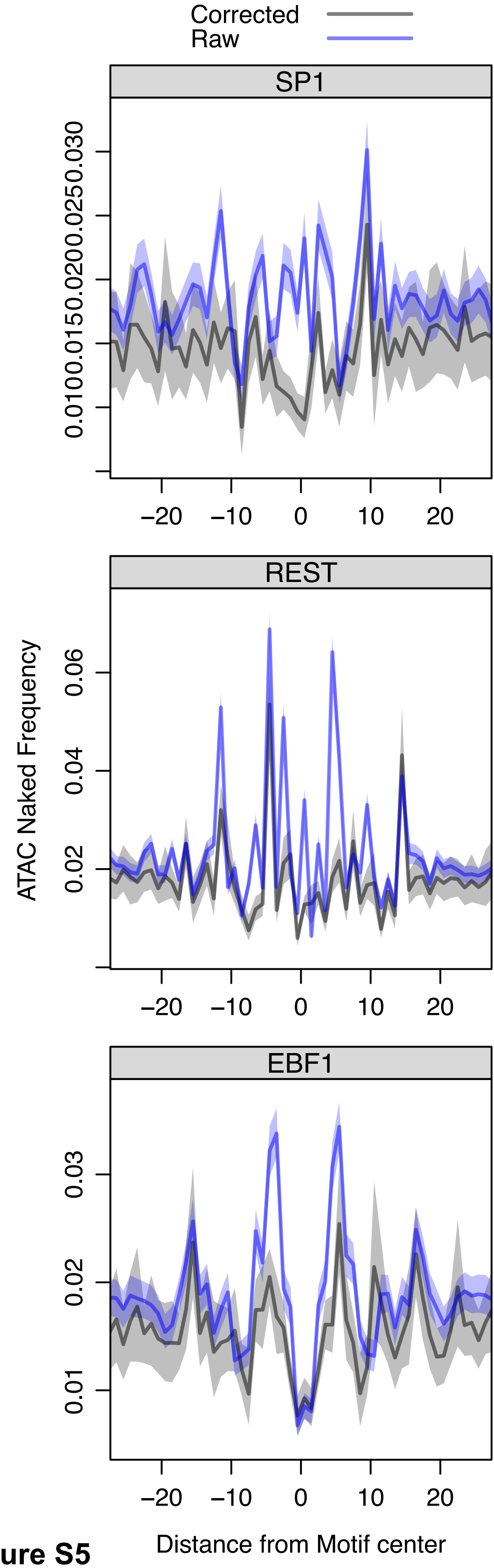

Corrected
Raw
SP1
0.010
0.015
0.020
0.025
0.030
–20
–10
0
10
20
REST
ATAC Naked Frequency
0.02
0.04
0.06
–20
–10
0
10
20
EBF1
0.01
0.02
0.03
–20
–10
0
10
20
Distance from Motif center

Figure S5

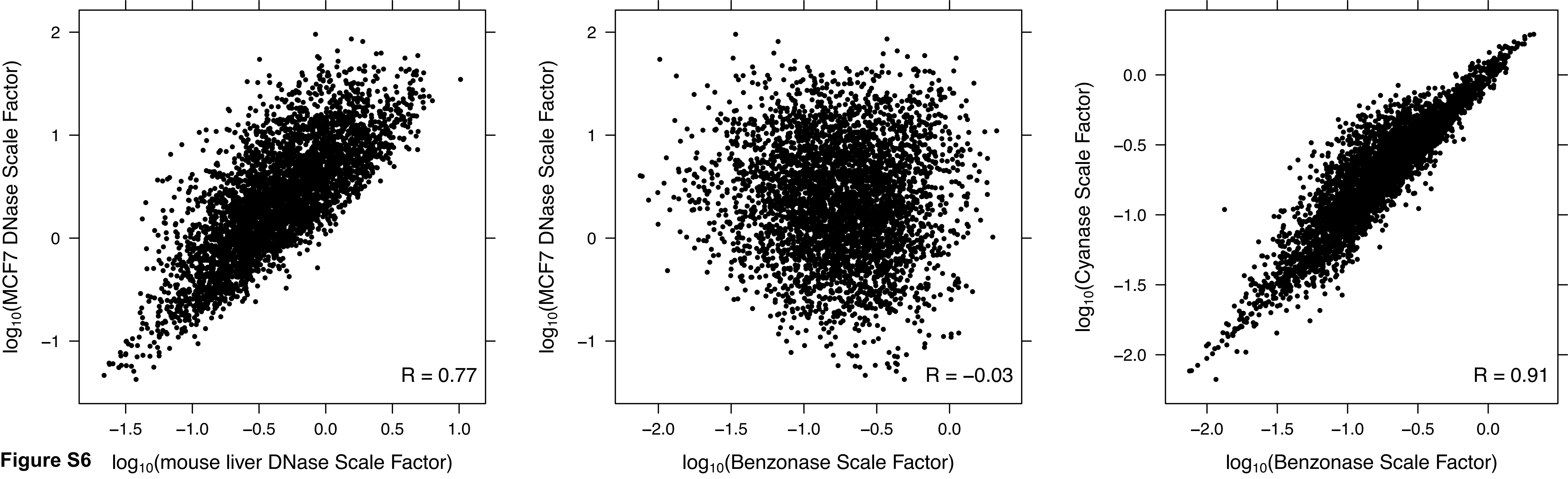

log10(MCF7 DNase Scale Factor)
2
1
0
−1
R = 0.77
−1.5
−1.0
−0.5
0.0
0.5
1.0
log10(mouse liver DNase Scale Factor)
log10(MCF7 DNase Scale Factor)
2
1
0
−1
R = −0.03
−2.0
−1.5
−1.0
−0.5
0.0
log10(Benzonase Scale Factor)
log10(Cyanase Scale Factor)
0.0
−0.5
−1.0
−1.5
−2.0
R = 0.91
−2.0
−1.5
−1.0
−0.5
0.0
log10(Benzonase Scale Factor)

Figure S6

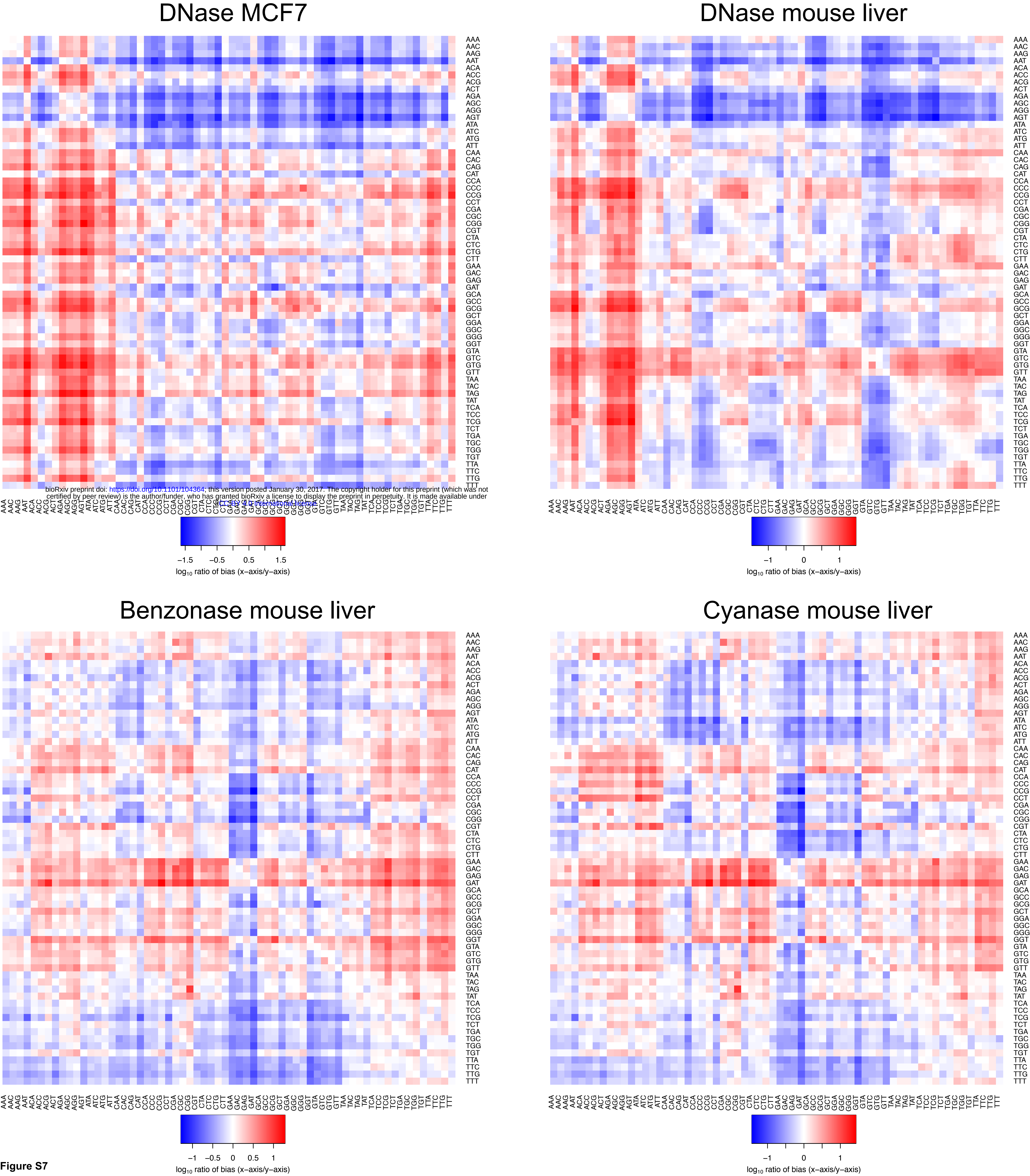

DNase MCF7
DNase mouse liver
Benzonase mouse liver
Cyanase mouse liver
AAA AAC AAG AAT ACA ACC ACG ACT AGA AGC AGG AGT ATA ATC ATG ATT CAA CAC CAG CAT CCA CCC CCG CCT CGA CGC CGG CGT CTA CTC CTG CTT GAA GAC GAG GAT GCA GCC GCG GCT GGA GGC GGG GGT GTA GTC GTG GTT TAA TAC TAG TAT TCA TCC TCG TCT TGA TGC TGG TGT TTA TTC TTG TTT
−1.5 −0.5 0.5 1.5
−1 0 1
−1 −0.5 0 0.5 1
−1 0 1
log$_{10}$ ratio of bias (x−axis/y−axis)

Figure S7

1. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods. Nature Publishing Group; 2009;6:283–9.

2. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132:311–22.

3. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011;147:1408–19.

4. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013;339:950–3.

5. Duarte FM, Fuda NJ, Mahat DB, Core LJ, Guertin MJ, Lis JT. Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. Genes Dev. 2016;30:1731–46.

6. Grøntved L, Bandle R, John S, Baek S, Chung H-J, Liu Y, et al. Rapid genome-scale mapping of chromatin accessibility in tissue. Epigenetics Chromatin. 2012;5:10.

7. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods. 2013;10:1213–8.

8. Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell. 1979;16:797–806.

9. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489:83–90.

10. Sung M-H, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol. Cell. 2014;56:275–85.

11. He HH, Meyer CA, Hu SS, Chen M-W, Zang C, Liu Y, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat. Methods. 2014;11:73–8.

12. Yardımcı GG, Frank CL, Crawford GE, Ohler U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res. 2014;42:11865–78.

13. Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat. Methods. 2016;13:303–9.

14. Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell. 2013;154:888–903.

15. Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008;9:517.

16. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans. Comput. Biol. Bioinform. 2013;10:645–56.

17. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

18. Wu C, Wong YC, Elgin SC. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell. 1979;16:807–14.

19. Wu C. The 5′ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature. Nature Publishing Group; 1980;286:854–60.

20. Kähärä J, Lähdesmäki H. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. Bioinformatics. 2015;31:2852–9.

21. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc. Natl. Acad. Sci. U. S. A. 2013;110:6376–81.

22. Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21:456–64.

23. Guertin MJ, Zhang X, Coonrod SA, Hager GL. Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. Mol. Endocrinol. 2014;28:1522–33.

24. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. 2014;46:1311–20.

25. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–8.

26. Homan PJ, Tandon A, Rice GM, Ding F, Dokholyan NV, Weeks KM. RNA tertiary structure analysis by 2’-hydroxyl molecular interference. Biochemistry. 2014;53:6825–33.