Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity

Songjoon Baek1,2, Ido Goldstein1,2,3, and Gordon L. Hager1,3,4

1Lab of Receptor Biology and Gene Expression, The National Cancer Institute, The National institutes of Health, Bethesda, Maryland, USA, 20892

Summary

In response to activating signals, transcription factors (TFs) bind DNA and regulate gene expression. TF binding can be measured by protection of the bound sequence from DNase digestion (i.e. footprint). Here, we report that 80% of TF binding motifs do not show a measurable footprint, partly due to a variable cleavage pattern within the motif sequence. To more faithfully portray the effect of TFs on chromatin, we developed an algorithm that captures two TF-dependent effects on chromatin accessibility - footprinting and motif-flanking accessibility. The algorithm, termed Bivariate Genomic Footprinting (BaGFoot), efficiently detects TF activity. BaGFoot is robust to different accessibility assays (DNase-seq, ATAC-seq), all examined peak-calling programs, and to a variety of cut bias correction approaches. BaGFoot reliably predicts TF binding and provides valuable information regarding the TFs affecting chromatin accessibility in various biological systems and following various biological events, including in cases where an absolute footprint cannot be determined.

Graphical Abstract
Introduction

Regulation of gene expression by transcription factors (TFs) is fundamental to every aspect of cell biology. Due to their intimate involvement in virtually all biological processes and their well-documented participation in pathological conditions, TFs and their binding to DNA regulatory elements (i.e. enhancers) have been extensively studied (Voss and Hager, 2014). Enhancer regions span a few hundred bases and serve as a binding hub for various proteins - TFs, chromatin remodelers, histone and DNA modifying enzymes etc. Enhancers are also accessible to nucleases and are hypersensitive to DNase digestion (i.e. DNase hypersensitive sites - DHSs, interchangeably termed peaks or hotspots). This property has been used to map the complete enhancer repertoire of cells in a genome-wide manner via partial digestion of DNA with DNase coupled with massive parallel sequencing (DNase-seq) (Boyle et al., 2008). Many techniques that map the accessible regions of the genome by nuclease cleavage have been developed, including the widely used ATAC-seq (Buenrostro et al., 2013).

Changes in enhancer accessibility either between cell types, or following a signal within the same cell population, are indicative of changing enhancer activity. Quantifying a significant change in accessibility between two biological conditions is commonly used to isolate a subset of enhancers. The subset of differentially-accessible enhancers is used to derive TF binding motifs that are statistically enriched in the subpopulation compared to total enhancers. The enriched motif might, in some cases, be indicative of increased TF binding. Although very useful and commonly used, this approach has critical pitfalls. First, an enrichment of a motif sequence is not direct evidence for TF binding. Second, such enrichment analyses often output many motif sequences with very low p values. A p value cutoff is done randomly and cannot be determined empirically. Therefore, choosing TFs to follow with downstream experiments is challenging. Third, motif enrichment is strictly dependent on isolating a subset of enhancers based on bulk, prominent changes in
accessibility. This dramatically reduces the number of enhancers used for analyses (in most cases 95–99% of enhancers are excluded from analysis). In some cases very few enhancers show significant changes for several reasons (e.g. poor sequencing quality, few replicates or the nature of the biological signal that leads to minor, hard-to-detect changes), casting further doubts as to the validity of findings.

As detailed above, the entire sequence spanning an enhancer region is more DNase accessible compared to non-regulatory regions of the genome. However, the few nucleotides that constitute the TF motif within the enhancer are sometimes locally \textit{protected} from DNase when bound by the TF. This results in a relative lower DNase cleavage rate at the motif compared to other regions within the DHS. This TF-dependent local protection was termed a TF ‘footprint’ (Galas and Schmitz, 1978) and has long served as a marker for TF binding and activity. In genome-wide chromatin accessibility assays, it became possible to find TF footprints by determining the relative decrease of DNase cuts at motif regions (Sung et al., 2016; Vierstra and Stamatoyannopoulos, 2016). This computational procedure, termed genomic footprinting, is increasingly applied and several computational pipelines are available (Gusmao et al., 2016).

Genomic footprinting was initially regarded as an alternative way to measure TF binding which could replace more direct methods such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Neph et al., 2012). However, with the increasing application of genomic footprinting, several aspects of TF and DNase function were found to affect footprint features. First, the identification of individual footprint events (as opposed to motif-centered genome-wide footprint patterns) has proven to be very challenging (Baek and Sung, 2016). The field has mostly shifted to determining aggregate footprints at the genome-wide level, represented as decreasing DNase cleavage events at the motif site compared to its flanking region. Second, inherent DNase cut bias was shown to fundamentally affect assessment of footprinting capacity of some TFs. Notably, several TF ‘footprints’ were found to be completely unrelated to TF binding, but rather represent DNase cut bias for DNA, and were thus termed ‘signatures’ (He et al., 2014; Sung et al., 2014). Third, several TFs were shown to lack a footprint altogether despite prominent binding to DNA (Grontved et al., 2015; He et al., 2014; Sung et al., 2014; Swinstead et al., 2016).

The pitfalls in conventional motif enrichment analyses and the newly-discovered layers of complexity in genomic footprinting warrant further advances in computational methods to determine TF activity from genomic accessibility data. Here, we describe a computational approach that utilizes genome-wide accessibility data to detect changes in TF activity between two biological conditions. This approach overcomes many of the limiting factors hindering current methods, such as the need for a subset of enhancers with a significant bulk change in accessibility in motif enrichment analysis. Also, in contrast to existing footprinting methods, the method can detect changes in TF activity even in TFs where a genomic footprint pattern cannot be measured. This becomes a critical issue since we show that a lack of a measurable footprint is the rule rather than exception for most TF motifs.
Results

A motif-wide assessment of footprinting patterns reveals that most TF motifs do not measurably protect from DNase cutting

The utility of genomic footprinting relies on a local decrease in nuclease/transposase accessibility due to TF binding. Computationally, this decrease is evaluated genome-wide by aggregating enzymatic cleavage events (i.e. ‘cut counts’) at and around a motif. Such a motif-centered aggregate plot has two distinct features; a decreased cut count at the motif compared to the flanking regions (‘footprint’) and a zigzag pattern of variable cut frequency between nucleotides within the motif sequence (‘signature’, Fig. 1A, B, bottom). The signature pattern derives primarily from sequence-dependent enzymatic cut bias and is also observed when cutting naked DNA (Fig. 1B, bottom) (He et al., 2014; Sung et al., 2014). Genomic footprinting patterns of many TFs in various cell types have been reported in recent years. In some reports a TF was found to leave no footprint in motif sequences bound by the factor, although a signature was evident (Grontved et al., 2015; He et al., 2014; Sung et al., 2014; Swinstead et al., 2016). We wished to explore if the lack of a footprint is a sporadic phenomenon occurring in a few cases or a more general feature of TFs. To achieve this genomic perspective, we generated a computational workflow for quantifying the genome-wide footprint depths of all known TF motifs. First, TF motifs were collected from three databases (JASPAR, TRANSFAC and UniPROBE) and motif occurrences in the mouse genome were scanned using FIMO (Grant et al., 2011). DHS sites were determined in a high quality DNase-seq experiment performed on mouse liver (Goldstein et al., 2017). Motifs outside DHSs were excluded and the DNase cut count in each motif occurrence was calculated and normalized to total reads in the sample. To determine the baseline, cut counts were also calculated and normalized in the motif-flanking region (±200 bp from motif center). Then, for each motif, cut counts of all genomic motif occurrences were aggregated and plotted (Fig. 1C).

Inherent DNase cut bias can profoundly affect a footprint’s signature. This bias was previously corrected with data from DNase-digested naked DNA (He et al., 2014; Lazarovici et al., 2013; Stergachis et al., 2014; Sung et al., 2014). However, this method of bias correction is specific for DNase-seq and is not applicable to other chromatin accessibility assays that will present their own bias (e.g. ATAC-seq). Also, technical variability between experiments and alternate commercial nuclease sources might present a different degree of bias. To overcome these issues, we employed a dataset-intrinsic bias correction methodology. We counted DNase cuts genome-wide (including cuts outside DHSs) and calculated the cut frequency of all possible nucleotide hexamer combinations normalized to their frequency of occurrence in the genome as previously described (Lazarovici et al., 2013) (Table S1). Thus, we compiled a dataset-intrinsic estimation of the cut bias of each nucleotide in our dataset. We then bias-corrected the cuts within DHSs by calculating the log ratio of observed cut counts (aggregated from all motif occurrences within DHSs) divided by the expected cut counts (deduced from hexamer cut bias). In many cases, the cut bias correction smoothed the zigzag patterns of motif signatures, allowing for an easier assessment of footprinting depth (Figs. 1B, S1). Finally, we determined the genome-wide
footprint depth (FPD) of each motif as the depression of the average log ratio value within the motif compared to the flanking accessibility baseline (Fig. 1C).

We determined FPDs in mouse liver for all known motifs following digestion with three different nucleases and sorted them by FPD value (Figs. 2A, S2A, Table S2). Surprisingly, we found a continuum of FPD values ranging from a negative value (i.e. local protection from cutting at motifs) to a positive value. To validate that the positive FPD values do not stem from our dataset-intrinsic bias correction, we also determined FPD values following bias correction based on three naked DNA controls (Lazarovici et al., 2013; Yardimci et al., 2014). In addition, we performed the dataset-intrinsic cut bias only on cuts within DHSs to dismiss the possibility that cuts outside DHSs are skewing our results. The pattern of continuous FPD values ranging from a negative to a positive value was observed regardless of the correction approach (Fig. S2B). Moreover, the identity of motifs with a positive or a negative FPD value was preserved across correction approaches, suggesting motifs do not randomly obtain FPD values due to different bias estimates. Rather, specific motifs are prone to present a positive FPD value regardless of bias correction approach (Table S2).

A positive FPD value could be interpreted as increased DNase digestion at the motif compared to its flanking region. However, inspection of uncorrected footprints presenting a positive FPD value revealed a signature pattern with considerable cut count variability between nucleotides comprising the motifs. In most cases, some nucleotides’ cut count values were below the flanking region baseline while others were above it, a phenomenon we termed an ‘incoherent signature’ (Figs. 2B, S2C). This suggests that there is not a uniform increase in cut count across the motif as the corrected plot and the FPD value might indicate. Rather, the positive FPD value is a result of averaging intra-signature variability. If this is true, positive FPD values may not represent a true biological event but stem from incoherent signature patterns. To explore this option, we filtered FPDs according to a stringent criterion. Motifs with a mixed pattern (whereby prior to correction some nucleotides have a positive value and others have a negative value; i.e. incoherent signatures were excluded from further analysis. Motifs where all nucleotides showed a consistent value (i.e. all showed either a positive or a negative value) were included and plotted (Fig. 2B). As expected with such a stringent criterion, 90% of motifs were discounted. Remarkably, the vast majority of motifs that met this cutoff showed a negative FPD (n = 61, 9% of total motifs) with only 8 showing a positive FPD value (1% of total motifs). Importantly, while all positive FPDs were very low (<0.1), negative FPD values had a wide spectrum of values with the lowest value reaching −1.4 (Fig. 2C). This is in stark contrast to unfiltered FPDs whereby negative and positive values were found at a similar rate (56% negative and 44% positive).

This observation suggests that a positive FPD value is an artifact stemming from the variability in DNase cutting between individual nucleotides comprising the motif (i.e. variability within motif signature leading to an incoherent signature). This leads to FPD values slightly below or above the baseline and reflects a feature of the DNA sequence rather than a feature of TF activity. Thus, when estimating FPD, one in fact conflates two factors determining the final value - the actual protection from DNase cutting together with intra-signature cut variability. Indeed, when examining the total group of FPD values, we found
that they fit a mix of two populations, each normally distributed. One population includes 80% of motifs and averages around zero FPD (average FPD = 0.016) with low variance, while the second population consists of 20% of motifs, is much more variable and averages at an FPD of −0.42 (Fig. 2D). Of note, this is not a liver-specific phenomenon and was observed in every one of the eight datasets we examined (Table S3). The parameters of these two populations are consistent with a model wherein most motifs average around FPD of zero due to signature incoherence while a fifth of motifs lead to a measurable footprint, averaging at a negative FPD indicating protection from cleavage. Thus, FPD values of most motifs in all examined datasets are so profoundly affected by intra-signature variability that we cannot estimate whether they leave a footprint or not.

The lack of a footprint in many motifs could be partially explained by the lack of expression of the TFs binding these motifs in the examined tissue (Gusmao et al., 2016; Stergachis et al., 2014). However, this cannot account for many footprint-lacking motifs in the datasets we examined (Table S3) as many footprint-lacking motifs are bound by TFs that are expressed in the given cell type. For example, some motifs lacking a footprint in the liver dataset bind TFs that are expressed in liver and play pivotal roles in liver biology by binding chromatin. Examples include the glucocorticoid receptor (GR), FoxA1/2, FoxO1, PPARA, STAT3, STAT5 and RORA (Fig. S2C).

Taken together, these findings reveal that measuring absolute FPD is dramatically affected by intra-signature variability and that 80% of motifs do not show a measurable footprint. Within this group of motifs one cannot determine whether this is due to lack of TF expression, lack of protection from DNase cutting or whether such protection exists but is masked by intra-signature cut variability.

**Bivariate genomic footprinting**

The observation that most motifs do not present an easily-measured footprint pattern and that absolute FPD calculation is confounded by intrinsic intra-signature cut variability seriously impedes conclusions drawn by genomic footprinting. This led us to design a computational method that would more faithfully capture a TF’s effect on chromatin accessibility in a manner resistant to these drawbacks. In some cases, alterations in TF expression (Gusmao et al., 2016) or activity (Goldstein et al., 2017) can lead to changes in FPD. In addition to this direct effect, TF binding events also lead to major changes in the accessibility of the chromatin environment around their binding sites by recruitment of chromatin remodeling complexes, histone modifying enzymes, hetero-dimerization and assisted loading of other TFs (Voss and Hager, 2014). Indeed, several studies have shown that TF activation leads to increases in the accessibility of the motif-flanking regions (He et al., 2012; Swinstead et al., 2016) and this property has been used in predicting TF activity (Sherwood et al., 2014).

We set out to generate a framework that will utilize changes in FPD and flanking accessibility (FA, set as ±200 bp from motif center) to predict changes in TF activity following a biological signal (e.g. untreated vs. treated cells, tumor vs. normal etc.). We designed an algorithm termed Bivariate Genomic Footprinting (BaGFoot) which enables a genomic, motif-wide comparison of FPD and FA. For every experimental condition, each
motif was assigned a FPD value and a FA value. Then, to compare the difference between two conditions we subtracted the FPD and FA of condition A from condition B yielding ΔFPD and ΔFA values (Fig. 3A, B). The Δ values for all motifs are presented in a scatter plot, ΔFA in the X axis and −ΔFPD in the Y axis. Because most TFs are not expected to change in activity and affect accessibility following a specific signal, most Δ values would center around the origin. Conversely, condition-dependent changes in TF activity may be reflected in deviation from the origin. For example, if the activity of a TF is increased in condition B leading to an increase in FPD and FA (Fig. 3A), both Δ values would increase and the motif bound by the TF would be found in the first quadrant (Fig. 3C).

To statistically evaluate the extent of the change in FA and FPD we present the data in a ‘bag plot’, a visual way to present statistics of bivariate data (Rousseeuw et al., 1999). This presentation is analogous to the well-known ‘box and whiskers plot,’ which shows the main statistical characteristics of univariate data. In a bag plot, the ‘bag’ area encompasses 50% or less of data points (analogous to the box in a box plot). The ‘fence’ area usually encompasses 97–100% of data points (analogous to the whiskers in a box plot). Any data point outside the fence is defined as an outlier; similarly to a box plot, this analysis does not dictate a minimal percentage of outliers (Fig. 3C). Because most TF motifs are expected to localize around the origin, the population median would be proximal to the origin and motifs with a significant deviation from the median would be characterized as outliers (Fig. 3C). Statistical significant change in FA/FPD was evaluated by chi squared distribution.

To examine the efficiency of BaGFoot in predicting TF activity, we first analyzed DNase-seq data from mouse liver following fasting, a metabolic stress known to affect many TFs (Goldstein et al., 2017; Goldstein and Hager, 2015). Comparing the fed to the fasted states identified four outlier groups of motifs deviating from the fence area. A group consisting of CREB-related motifs increased in both FPD and FA. A group of CEBP-related motifs increased mainly in FPD. The GR motif increased only in FA and a group of STAT motifs decreased in both FPD and FA. All outlier motifs represent statistically significant changes following fasting, while the motifs within the fence are not significant. (Fig. 3D). To further assess outlier significance, we applied a two-sample t-test comparing the outlier motifs in subsamples of data (90% of reads compared to the original read count, 10 subsamples in each condition). All outlier motifs were statistically significantly called in both the FA and FPD parameters ($P < 0.01$).

Remarkably, the four outlier groups of motifs bind TFs that are well known mediators of the hepatic response to fasting (Goldstein and Hager, 2015). In line with the predictions made by BaGFoot, the binding of CREB, GR and CEBPB were all significantly increased following fasting as measured by ChIP-seq in the same experimental conditions (Goldstein et al., 2017).

**BaGFoot is robust to different peak calling methods, bias correction approaches and to low-read datasets**

DHS sites in this study were identified by the DNase2Hotspot program. To examine if BaGFoot can provide a reproducible output regardless of the peak calling algorithm, we also used MACS2 or F-seq to determine DHSs, and executed BaGFoot on the two peak lists. The
number of called peaks varied considerably between the three algorithms (DNase2Hotspots: 78,821, MACS: 115,760, F-seq: 324,606). Nonetheless, the BaGFoot output from MACS-called or F-seq-called peaks (Fig. S3A, B) was extremely similar to the output from DNase2Hotspots-called peaks (Fig. 3D). Even when BaGFoot was applied to the entire genome with no peak calling step, the output closely resembled that for the DHS-exclusive BaGFoot analysis (Fig. S3C). Of note, BaGFoot analysis on the entire genome with no focus on called peaks is computationally intense and is not recommended for most datasets in which peak calling is possible.

Because bias correction profoundly affects absolute FPD values (Fig. 2), we examined whether it also affects BaGFoot output. We performed BaGFoot on naked DNA-corrected data whereby the cut bias pattern of each DNA hexamer is determined by DNase-digested naked DNA (Table S1) (Lazarovici et al., 2013; Yardimci et al., 2014). The plot following the various naked DNA corrections (Fig. S4A–C) was essentially identical to our dataset-intrinsic correction approach described above (Fig. 3D). Additionally, the BaGFoot output was unaffected when we corrected for bias based only on cuts within DHSs (Fig. S4D), rather than throughout the genome (Fig. 3D); suggesting the dataset-intrinsic correction can be done on either population. Furthermore, the output was not affected even when the cleavage rates were randomly shuffled (Fig. S4E). The resilience of BaGFoot to correction approach prompted us to examine the output following two mock corrections – one where the DNase cleavage rate was assumed to be uniform throughout hexamers, and a second where the hexamer frequency was assumed to be uniformly distributed in the genome (Table S1). Even with these simulated corrections, BaGFoot performed similarly (Fig. S4F–G).

Thus, we conclude that BaGFoot is robust to different corrections and even performs similarly when no correction is done (equal cleavage rate, Fig. S4G). The resistance of BaGFoot to assay bias stems from the fact that BaGFoot estimates the difference between two conditions. By definition, bias stemming from enzymatic activity should not change between similarly-processed samples. Because we subtract one value from another the cut bias that is identical between conditions is negated and the delta value represent only relevant changes.

A major drawback in extrapolating TF activity from genome-wide accessibility data is the need to determine a subset of enhancers that show a bulk change in accessibility between conditions in a statistically-significant manner, then analyze that sub-population for motif enrichment. Because most DHSs do not change dramatically in their accessibility, they are excluded from analysis and the subset of included DHSs represents a small fraction of total enhancers. Furthermore, this subset is sometimes challenging to determine in a statistically-significant manner due to few replicates or poor quality of data. We postulated that BaGFoot would be able to find relevant TFs even in DHSs that do not show a statistically-significant difference in accessibility between conditions. We employed the fed/fasted dataset and discounted all DHSs with a bulk change in accessibility following fasting (≥2 fold, adjusted p value ≤0.05). Performing BaGFoot only on unchanged DHSs (Fig. 4A) provided an almost identical output with the same outlier motifs as those found on total DHSs (Fig. 3D).

Many footprinting algorithms require deeply-sequenced DNase-seq data to detect and measure absolute footprints (Baek and Sung, 2016). Because BaGFoot relies on relative...
rather than absolute measurements, we surmised it might be robust even in low-read datasets. To test this hypothesis, we randomly subsampled reads from the fed/fasted dataset and artificially generated datasets with a lower number of reads (90%, 80% ... 10% of reads compared to original dataset). Each subsampling percentage tier was randomly selected 10 times (90%, n = 10 etc.) to a total of 90 different datasets. Then, we performed BaGFoot on these datasets. ΔFPD and ΔFA values were reliably determined with very low variance in as low as 20% of reads for six examined motifs (Fig. 4B, S5). Remarkably, the vast majority of outliers identified in the original dataset were also identified as outliers in all subsampled datasets, suggesting that BaGFoot is able to detect changes in TF activity even in low-read datasets (Fig. 4C, Table S4).

Taken together, BaGFoot efficiently detects outlier motifs regardless of peak calling algorithm or bias correction method. BaGFoot also reliably isolates outlier motifs in datasets with varying read depths.

**BaGFoot detects relevant motifs following a wide range of biological signals and in various cell types**

The response to fasting is a systemic one, mediated by various hormonal and metabolic stimuli. We wished to determine if BaGFoot can retrieve TFs relevant to several biological events. First, we examined BaGFoot’s performance following a TF-specific activating signal. To that end we analyzed DNase-seq data obtained following short-term treatment of cells with a GR agonist – dexamethasone (John et al., 2011). Comparing dexamethasone-treated cells to their untreated counterparts revealed increases in the accessibility around Fox and Sox motifs, suggesting that GR facilitates chromatin opening around these motifs and thereby promoting both Fox and Sox TF activity (Fig. 5A). Both Sox and Fox motifs do not leave a footprint (Fig. S6A) and therefore show only an increase in FA along the X axis. This observation highlights the advantage of BaGFoot compared to a univariate examination of footprinting capacity that would be uninformative in this case due to lack of footprints.

Interestingly, a role for GR in enhancing FoxA1 binding was recently described (Swinstead et al., 2016) and is in line with our finding that activating GR leads to increasing accessibility around Fox motifs. A similar relationship between Sox and GR is undescribed and warrants further investigation. The two other motifs detected, Pax4 and Zfp105, are very prevalent in the DHSs with over 70,000 occurrences, suggesting this motif cannot provide reliable information about TF binding because most of these motifs are unbound.

Next, we examined the sensitivity of our algorithm in detecting lineage-determining factors in early developmental stages. Hemangiblasts are the early common precursors for hematopoietic and endothelial cells; thus the differentiation of mesodermal cells to hemangiblasts is thought to constitute the first step towards hematopoiesis (Cao and Yao, 2011). We compared DNase-seq of hemangiblasts to their parental mesodermal cells using BaGFoot (Goode et al., 2016). Increases in accessibility around the GATA motif were found in hemangiblasts when compared to their parental mesodermal cells (Fig. 5B). Supporting the validity of our findings are observations indicating GATA1 is a pivotal erythroid-determining TF during hematopoiesis with documented roles in hemangiblasts (Cao and Yao, 2011; Yokomizo et al., 2007).
The accessibility data described above was based on DNase hypersensitivity. The ability to recover footprints from other chromatin accessibility assays has been questioned (Vierstra and Stamatoyannopoulos, 2016). We performed BaGFoot analysis on a recent dataset utilizing ATAC-seq to compare accessibility between a group of primary small cell lung cancer (SCLC) tissues and a group consisting of SCLC liver metastases (Denny et al., 2016). The major motif showing increased accessibility and a deeper footprint was NF1 (Fig. 5C). This is in agreement with the discovered role of NFIB in regulating metastatic development (Denny et al., 2016). Intriguingly, the motif for the hepatic TF HNF6 was also called as an outlier, raising the possibility that the liver microenvironment affected metastatic cells to express liver TFs. However, while the NF1 motif was still called an outlier when correcting the data based on naked DNA-digested ATAC-seq, the HNF6 motif was not (Fig S6B).

Taken together, the data suggests that BaGFoot is efficient in predicting relevant TFs in various settings including systemic metabolic stress, early developmental stages, specific activation of a TF and in tumor-to-metastases transitions.

**Candidate motifs discovered by BaGFoot are bound by the predicted TFs**

The outlier motifs detected in liver following fasting (Fig. 3D) directly correspond to increased binding of three TFs following fasting (Goldstein et al., 2017). To further validate changes in TF binding behaviors corresponding to outlier motifs, we processed multiple datasets with the BaGFoot algorithm.

First, we analyzed data in which mesodermal cells were in-vitro differentiated into macrophages (Goode et al., 2016). All the outlier motifs in the BaGFoot bag plot comparing macrophages to their mesodermal origin cells (Fig. 6A) have been reported to bind TFs with critical roles in macrophage activity and lineage determination (Tussiwand and Gautier, 2015). Attesting to the validity of BaGFoot’s observations, ChIP-seq data revealed increased binding of CEBPB (bound by an identified outlier motif) in macrophages compared to mesodermal cells (Fig. 6B). The purpose of BaGFoot is to suggest putative motifs for further exploration using only accessibility data. Therefore, in the lack of ChIP-seq data, BaGFoot considers both bound and unbound motifs. With the availability of CEBPB ChIP-seq in this dataset, we divided CEBP motifs to two groups - bound motifs and unbound motifs. While unbound motifs were close to the origin, bound motifs were extreme outliers (extending further from the value for total motifs, Fig. 6C). This shows that the changes in FPD/FA are due to TF binding and not motif sequence.

To further explore the efficiency of BaGFoot in detecting biologically-relevant TFs, we examined a DNase-seq dataset from T lymphoblast cells (Bevington et al., 2016). Upon encountering an antigen and activation of T cell receptor (TCR) signaling, naive T cells become lymphoblasts and begin rapidly proliferating to eventually become effector cells. Lymphoblasts respond more efficiently to a second TCR activation by PMA/ionomycin (PMA/I) (Bevington et al., 2016). Comparing activated vs. resting lymphoblast cells using BaGFoot, we found known TCR-activated TFs such as AP-1 and NFAT (Fig. 6D). In line with BaGFoot’s prediction, the AP-1 TF was activated in this setting as measured by increased binding (Fig. 6E). The motifs bound by AP-1 were responsible to the increases in FA and FPD while unbound motifs had no effect (Fig. 6F).
The above-described TFs were also detected in the original motif analysis performed on a subset of DHSs that increased in accessibility upon PMA/I treatment (Bevington et al., 2016). However, BaGFoot also detected the motifs bound by BATF and IRF4, two TFs that play important roles in T cell biology (Man and Kallies, 2015) (Fig. 6D). These were not detected in conventional motif analysis, demonstrating that in some cases, BaGFoot can identify critical TFs that were missed in existing methods.

Interestingly, some motifs appeared in the second quadrant showing an increase in FPD but a decrease in FA. This accessibility pattern would be consistent with a repressor that decreases its surrounding accessibility as it binds to its motif. Indeed, two motifs found in the second quadrant were reported to bind repressors: EGR (Feng et al., 2015) and Sp4 (Kwon et al., 1999). As for the motifs shown in Fig. 5A, three frequently occurring motifs were detected as outliers in this dataset (Pax4, Zfp281 and Zfp740), but we cannot deduce TF activity from this observation due to the abnormally high number of motif occurrence in DHSs (>50,000).

We also examined a dataset where changes in chromatin accessibility were measured during early adipogenesis (Siersbaek et al., 2011). The CEBP motif was called as an outlier (Fig. S6C), in line with the role of CEBPB in the early establishment of a chromatin landscape facilitating adipogenesis (Siersbaek et al., 2011). Indeed, CEBPB binding increased following treatment and the bound motifs were responsible for the shift in the FA/FPD parameters (Fig. S6D, E).

Finally, we processed ATAC-seq data from mouse embryonic cells (mESCs) where the level of Oct4, a critical stemness TF, was knocked-down (King and Klose, 2017). The most distant outlier motif showing a decrease in FPD and FA was the Oct4 motif (Pou5f1, Fig. 6G). Interestingly, the motifs for Sox2 and NANOG also showed the same trend in line with findings showing Oct4-dependent binding of these TFs (King and Klose, 2017). Examining ChIP-seq data for the three TFs revealed that as predicted by BaGFoot, binding of all three was reduced following Oct4 downregulation (Fig. 6H). The binding of Oct4, Sox2 and NANOG was shown to be regulated by the chromatin remodeler BRG1 (King and Klose, 2017). Accordingly, BaGFoot analysis showed a reduction in both FPD and FA for all three motifs in cells where BRG1 level was knocked-down (Fig. S6F).

Taken together, the results from eight different datasets collectively show that the BaGFoot algorithm efficiently predicts TF activity following short- and long-term activating signals in various settings (differentiation, metabolic stress, immunogenic response, cancer-related chromatin alterations etc.). BaGFoot can detect motifs that elude conventional motif enrichment analysis as well as motifs that show no measurable footprint. Importantly, the BaGFoot algorithm operates well on ATAC-seq data, indicating its utility for this increasingly utilized methodology. Finally, BaGFoot’s predictions of TF activity were confirmed by ChIP-seq in five different datasets.
Discussion

TF footprinting approaches have evolved over the last few years, leading to advances in our understanding of TF biology. By the same token, inference of TF activity from footprinting was realized to be more complex than initially imagined, as footprint patterns were discovered to be affected by several determinants. Here we show that most TFs do not leave a measurable footprint on DNA, either due to biological properties of the TF or due to intra-signature cut variability.

To negate signature variability, to identify footprint-lacking TFs and to maximize information obtained from accessibility assays, we devised a computational workflow incorporating two properties of TFs that are indirectly measured by accessibility information. Upon binding, TFs increase their FA by recruiting other factors to the enhancer region and, in some cases, also leave a footprint. BaGFoot quantifies the difference in these two parameters between two experimental conditions. A signal that would activate a certain TF, leading to increased binding, has the potential to lead to increased FA and FPD.

A critical feature of BaGFoot is an unbiased output that does not require prior knowledge on TFs activated in the examined pathway. This is especially advantageous in circumstances where the TFs relevant to the examined biological condition are either unknown, or are too many to assess individually. This stands in stark contrast to ChIP-seq experiments that necessitate both knowledge of the target TF as well as TF-specific antibodies that are often difficult to obtain. Thus, when using BaGFoot, one accessibility experiment can direct our attention to relevant TFs in an unbiased manner. Indeed, a nascent, proof-of-principle, version of the algorithm was included in a recently-published paper. There, the algorithm was able to narrow down a list of 20 potential TFs involved in the regulation of fasting to just 3 TFs (Goldstein et al., 2017). Additionally, because BaGFoot relies on FA as well as on FPD, it is able to detect TF activity of footprint-lacking TFs. This represents a major advance compared to methods relying strictly on footprinting capacity.

Commonly, detection of bulk changes in enhancer accessibility followed by motif enrichment is used to predict TF binding. We found that BaGFoot is superior to conventional motif enrichment analysis for several reasons: (i) in BaGFoot there is no need to analyze a subpopulation of differentially accessible enhancers. (ii) BaGFoot can detect relevant TFs omitted by motif analysis. (iii) BaGFoot measures properties directly related to TF activity and is therefore more indicative of actual TF binding.

In addition to motif enrichment, other approaches were recently employed to discern TF activity from accessibility data. As found with BaGFoot, increases in motif-flanking accessibility were correlated to increases in TF activity during differentiation (Sherwood et al., 2014). Additionally, increases in TF gene expression (Gusmao et al., 2016) and binding (Pique-Regi et al., 2011) were correlated to footprinting capacity. The strength of BaGFoot is in combining these two TF properties in a systematic and robust way to a single, statistically-evaluated output portraying the overall effect TFs have on chromatin accessibility.

Cell Rep. Author manuscript; available in PMC 2017 July 27.
Although presenting clear advantages, several aspects of BaGFoot may, in some cases, limit conclusions drawn by it. First, in contrast to ChIP that directly measures TF binding, BaGFoot only suggests TF binding by proxy. Second, as motifs are sometimes bound by a family of TFs, isolating the involved TF from the family might be challenging. In some cases, this is easily resolved if gene expression data is available for the relevant cell type and unexpressed TFs could be excluded. Third, a minority of TFs bind within inaccessible regions (Thurman et al., 2012). Because BaGFoot is reliant on quantifying enough cleavage events, it is restricted to accessible regions and would therefore not be able to detect these TFs. Lastly, BaGFoot considers all motifs within DHS sites, some of which are unbound by TFs, potentially leading to under-representation of a TF’s effect on chromatin (i.e. some actual outliers would not be called because most motifs are unbound and ‘dilute’ the signal). Thus, TF-specific downstream experiments may be needed to validate BaGFoot’s output. Nonetheless, in all five datasets where this issue was tested, outliers detected by BaGFoot were associated with differential binding of the relevant TF. In the other three datasets, the outlier motifs were previously shown to bind TFs under the same conditions.

In summary, BaGFoot efficiently detects TF activity following various biological events. BaGFoot is robust to different peak-calling algorithms, low-read datasets, and to different genomic accessibility methods. BaGFoot can isolate footprint-lacking TFs. Lastly, the algorithm functions seamlessly even in datasets where differentially-accessible enhancers are not observed. Therefore, BaGFoot presents a major improvement in the retrieval of differential TF activity from genomic accessibility experiments, and provides valuable information regarding the TFs affecting chromatin accessibility in biological pathways.

**Experimental Procedures**

**BaGFoot**

In the BaGFoot scatter plot, each point represents a motif, the X axis values represent the change in the average normalized tag counts between two experimental conditions of DNase data of the ±200 bp region at each motif centers (FA). The Y axis values represent the change in FPD. To visualize the location, spread, skewness and to pinpoint outliers among motifs, we employed the bag plot approach (Rousseeuw et al., 1999). We observed that the population median (the point with the highest possible Tukey depth) locates close to the origin. This implies that our normalization works effectively in measuring both FA and FPD. The inner polygon (“bag”) contains at most 50% of the TFs. The outer polygon “fence” is formed by inflating the bag geometrically by a default factor of 3.

**Publicly available data used in the study**

All data used in this study were obtained from published datasets cited in the main text and deposited in Gene Expression Omnibus (GEO). Accession numbers for naked DNA data: GSE18927, GSE32970 and GSE92674. Accession numbers for chromatin digestions/ChIP-seq: GSE72087, GSE39982, GSE67465, GSE81258, GSE69101, GSE26189, GSE87822 and GSE27826 (see also Table S3).
Code Availability

Under GNU General Public License, v 3.0, all software implementations used in this paper are available at https://sourceforge.net/projects/bagfootr/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Prof. Hongshik Ahn, Stony Brook University for comments that greatly helped in improving the algorithm and Erin E. Swinstead for help in manuscript editing. This work was supported by the Intramural Research Program of the National Institutes of Health, the National Cancer Institute and the Center for Cancer Research.

References

Baek S, Sung MH. Genome-Scale Analysis of Cell-Specific Regulatory Codes Using Nuclear Enzymes. Methods Mol Biol. 2016; 1418:225–240. [PubMed: 27008018]

Bevington SL, Cauchy P, Piper J, Bertrand E, Lalli N, Jarvis RC, Gilding LN, Ott S, Bonifer C, Cockerill PN. Inducible chromatin priming is associated with the establishment of immunological memory in T cells. EMBO J. 2016

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132:311–322. [PubMed: 18243105]

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; 10:1213–1218. [PubMed: 24097267]

Cao N, Yao ZX. The hemangioblast: from concept to authentication. Anatomical record (Hoboken, NJ: 2007). 2011; 294:580–588.

Denny SK, Yang D, Chuang CH, Brady JJ, Lim JS, Gruner BM, Chiou SH, Schep AN, Baral J, Hamard C, et al. Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. Cell. 2016; 166:328–342. [PubMed: 27374332]

Feng Y, Desjardins CA, Cooper O, Kontor A, Nocco SE, Naya FJ. EGR1 Functions as a Potent Repressor of MEF2 Transcriptional Activity. PLoS One. 2015; 10:e0127641. [PubMed: 26011708]

Galas DJ, Schmitz A. DNase footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978; 5:3157–3170. [PubMed: 212715]

Goldstein I, Baek S, Presman DM, Pakkinaho V, Swinstead EE, Hager GL. Transcription factor assisted loading and enhancer dynamics dictate the hepatic fasting response. Genome Res. 2017; 27:427–439. [PubMed: 28031249]

Goldstein I, Hager GL. Transcriptional and chromatin regulation during fasting - The genomic era. Trends Endocrinol Metab. 2015; 26:699–710. [PubMed: 26520657]

Goode DK, Obier N, Vijayakaskar MS, Lie ALM, Lilly AJ, Hannah R, Lichtinger M, Batta K, Florkowska M, Patel R, et al. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev Cell. 2016; 36:572–587. [PubMed: 26923725]

Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27:1017–1018. [PubMed: 21330290]

Grontved L, Waterfall JJ, Kim DW, Baek S, Sung MH, Zhao L, Park JW, Nielsen R, Walker RL, Zhu YJ, et al. Transcriptional activation by the thyroid hormone receptor through ligand dependent receptor recruitment and chromatin remodeling. Nat Commun. 2015; 6:7048. [PubMed: 25916672]

Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat Methods. 2016; 13:303–309. [PubMed: 26901649]
He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res. 2012; 22:1015–1025. [PubMed: 22508765]

He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Methods. 2014; 11:73–78. [PubMed: 24317252]

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet. 2011; 43:264–268. [PubMed: 21258342]

King HW, Klose RJ. The pioneer factor OCT4 requires the chromatin remodeler BRG1 to support gene regulatory element function in mouse embryonic stem cells. Elife. 2017:6.

Kwon HS, Kim MS, Edenberg HJ, Hur MW. Sp3 and Sp4 can repress transcription by competing with Sp1 for the core cis-elements on the human ADH5/FDH minimal promoter. J Biol Chem. 1999; 274:20–28. [PubMed: 9867805]

Lazarovici A, Zhou T, Shafer A, Dantas-Machado AC, Riley TR, Sandstrom R, Sabo PJ, Lu Y, Rohs R, Stamatoyannopoulos JA, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci US A. 2013; 110:6376–6381.

Man K, Kallies A. Synchronizing transcriptional control of T cell metabolism and function. Nature reviews Immunology. 2015; 15:574–584.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012; 489:83–90. [PubMed: 22955618]

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011; 21:447–455. [PubMed: 21106904]

Rousseeuw PJ, Ruts I, Tukey JW. The Bagplot: A Bivariate Boxplot. The American Statistician. 1999; 53:382–387.

Sherwood RI, Hashimoto T, O’Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014; 32:171–178. [PubMed: 24441470]

Siersbaek R, Nielsen R, John S, Sung MH, Baek S, Loft A, Hager GL, Mandrup S. Extensive chromatin remodelling and establishment of transcription factor ‘hotspots’ during early adipogenesis. EMBO J. 2011; 30:1459–1472. [PubMed: 21427703]

Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature. 2014; 515:365–370. [PubMed: 25409825]

Sung MH, Baek S, Hager GL. Genome-wide footprinting: ready for prime time? Nat Methods. 2016; 13:222–228. [PubMed: 26914206]

Sung MH, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol Cell. 2014; 56:275–285. [PubMed: 25242143]

Swinstead EE, Miranda TB, Paakinaho V, Baek S, Goldstein I, Hawkins M, Karpova TS, Ball DA, Mazza D, Lavis LD, et al. Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. Cell. 2016; 165:593–605. [PubMed: 27062924]

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. [PubMed: 22955617]

Tussiwand R, Gautier EL. Transcriptional Regulation of Mononuclear Phagocyte Development. Front Immunol. 2015; 6:533. [PubMed: 26539196]

Vierstra J, Stamatoyannopoulos JA. Genomic footprinting. Nat Methods. 2016; 13:213–221. [PubMed: 26914205]

Voss TC, Hager GL. Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet. 2014; 15:69–81. [PubMed: 24342920]
Yardimci GG, Frank CL, Crawford GE, Ohler U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res. 2014; 42:11865–11878. [PubMed: 25294828]

Yokomizo T, Takahashi S, Mochizuki N, Kuroha T, Ema M, Wakamatsu A, Shimizu R, Ohneda O, Osato M, Okada H, et al. Characterization of GATA-1(+) hemangioblastic cells in the mouse embryo. Embo j. 2007; 26:184–196. [PubMed: 17159898]
Figure 1. Motif-wide genomic footprinting workflow
(a) TFs bind DNA regulatory elements that are hypersensitive to DNase digestion. Upon binding, TFs relatively protect their binding sites from DNase digestion, leading to a footprint detected by reduced cleavage events at the motif. A genome-wide footprint pattern is visible in an aggregation plot that includes all motif occurrences. DNase cutting is strictly affected by DNA sequence, leading to a ‘signature’ at the motif sequence with a typical zigzag pattern.

(b) A heat map and aggregate plot of cut counts in mouse liver at all CEBP motif occurrences (n = 12,318). Uncorrected data (‘observed cuts’) are shown in the left-hand side. Values corrected for cut bias (‘expected cuts’) are shown on the right-hand side. The
correlation between decreased cleavage events at the motif and TF binding is shown by ChIP-seq data of CEBPB on the far right obtained from (Goldstein et al., 2017).

(c) Genomic footprinting workflow: (1) DHS sites are determined. (2) Motif occurrences of all known motifs within the genome assembly are scanned. (3) For each known motif, all motif occurrences within DHSs are aggregated and a cut count aggregate plot is generated by counting cleavage events (normalized to total reads in the sample). (4) Cut bias is calculated for every hexamer in the genome by measuring cleavage events throughout the dataset. (5) Cut bias is corrected for every motif by a log ratio of observed cuts divided by expected cuts. (6) FPD is determined as the distance between the motif-flanking baseline value and the average value within the motif.

See also Fig. S1
Figure 2. Most TF motifs do not show a measurable footprint due to incoherent signatures
(a) Log ratio-corrected FPD values in mouse liver for all motifs are shown. FPD span a wide spectrum of values, with a prominent fraction of motifs showing a positive value.
(b) To filter out incoherent signatures, a cutoff was determined whereby any signature crossing the flanking baseline is excluded from further analysis.
(c) Log ratio-corrected FPD values in mouse liver for motifs meeting the criteria described in (b) are shown. After filtering, very few positive values remain and the population consists mainly of motifs with a negative FPD value.
(d) A probability density plot reveals a mixture of two normal distributions within FPD values ($\mu =$ mean, $p =$ mixture weight).
See also Fig. S2
Figure 3. BaGFoot - quantifying the changes in footprint depth and flanking accessibility between two experimental conditions

(a) The effects on chromatin accessibility mediated by TF activation can be measured via changes in accessibility within the motif and in the sequence flanking it.

(b) Changes in FPD and FA following a biological signal are measured for every motif genome-wide and delta values are given.

(c) Delta values for all motifs are plotted in a bag plot. The population median is marked in light orange. The bag area (dark blue) is the region where 50% of the population is located. The fence area (light blue) is the region where most (typically 97–99%) motifs are located. Motifs outside the fence and with a p value < 0.05 are determined as statistically-significant outliers, namely motifs with a significant condition-dependent change in FA and/or FPD.

(d) Bag plot depicting ΔFA and ΔFPD in mouse liver following fasting. Statistically-significant outlier motifs are marked in colored squares.

See also Fig. S3, S4
Figure 4. BaGFoot efficiently detects TF activity in datasets with low reads and with weak changes in enhancer hypersensitivity
(a) Bag plot depicting ΔFA and ΔFPD in mouse liver following fasting. All DHSs with a significant bulk change between conditions (fold ≥2, adjusted p value ≤0.05) were excluded, leading to only very subtle changes in outlier motifs detected.
(b) Box plots depicting ΔFA and ΔFPD values of selected motifs in randomly-generated datasets. The number of BaGFoot input reads was reduced to the indicated percentage (% reads from the original mouse liver dataset, each percentage tier was randomly generated 10 times).
(c) Bag plot depicting all outliers called in the entirety of subsampled datasets (n = 90). Each motif was determined an outlier in X% of datasets. Circle color and size represent binned ranks. For example, an outlier determined as such in 100% of datasets is marked in a large blue circle.
See also Fig. S5
Figure 5. BaGFoot detects TF activity in a variety of biological systems and following various biological transitions

Bag plots depicting ΔFA and ΔFPD in: (a) dexamethasone-treated mammary adenocarcinoma cells compared to untreated cells. (b) Hemangioblasts compared to their cells of origin - mesodermal cells. (c) a group of tumors, consisting of metastatic tumors ('hyper') compared to a primary tumor group ('hypo').

See also Fig. S6
Figure 6. BaGFoot reliably predicts changes in TF behavior between conditions

Bag plot in macrophages compared to their cells of origin revealed CEBP as an outlier motif (a). That motif was increasingly bound by CEBPB in macrophages as measured by ChIP-seq obtained from (Bevington et al., 2016) (b). Bound CEBP motifs were responsible for the increase in FA and FPD whereas unbound motifs had no effect (c).

Bag plot in PMA/I treated lymphoblasts compared to untreated lymphoblasts revealed AP-1 as an outlier motif (d). The AP-1 motif was increasingly bound by junB in treated cells as measured by ChIP-seq obtained from (Goode et al., 2016) (e). Bound AP-1 motifs were responsible for the increase in FA and FPD whereas unbound motifs had no effect (f).

Bag plot in mESCs following Oct4 downregulation finds the Oct4, Sox2 and NANOG motifs as outliers (g). The proteins binding these motifs are more weakly bound at Oct4 enhancers as measured by ChIP-seq obtained from (King and Klose, 2017) (h).

See also Fig. S6