© 2016 Nature America, Inc. All rights reserved.

Genomic footprinting

Jeff Vierstra[1,2] & John A Stamatoyannopoulos[1–3]

The advent of DNA footprinting with DNase I more than 35 years ago enabled the systematic analysis of protein-DNA interactions, and the technique has been instrumental in the decoding of *cis*-regulatory elements and the identification and characterization of transcription factors and other DNA-binding proteins. The ability to analyze millions of individual genomic cleavage events via massively parallel sequencing has enabled *in vivo* DNase I footprinting on a genomic scale, offering the potential for global analysis of transcription factor occupancy in a single experiment. Genomic footprinting has opened unique vistas on the organization, function and evolution of regulatory DNA; however, the technology is still nascent. Here we discuss both prospects and challenges of genomic footprinting, as well as considerations for its application to complex genomes.

Combinatorial binding of transcription factors (TFs) within regulatory DNA forms the basis for gene regulation in all organisms[1]. Precise determination of the location and dynamics of TF occupancy *in vivo* is vital to a mechanistic understanding of genome regulation and the interpretation of genetic variation. Soon after the discoveries of Southern blotting and DNA sequencing in the mid-1970s, Galas and Schmitz[2] reported that subjecting a protein-DNA complex to a nonspecific nuclease *in vitro* gave rise to stretches of contiguous nucleotides that were protected from nucleolytic attack, which they termed protein 'footprints'. Genomic footprinting extends the classical *in vitro* assay to nuclei[3–5] and whole cells[6,7], whereby the *in vivo* binding of TFs to their cognate recognition sequences occludes small segments (~6–20 nt) of nucleotides from the otherwise dense cleavage activity that characterizes active regulatory DNA (~150–300 nt). The genomic footprinting paradigm provided a powerful experimental approach that was rapidly and widely exploited to unveil the organization and function of regulatory DNA.

Recent advances in modern sequencing technologies have enabled *in vivo* DNA footprinting at the genome scale (digital genomic footprinting (DGF)), creating the potential for simultaneous global characterization of TF occupancy in a single experiment[8]. The application of DGF to diverse organisms from yeast to humans[8–12] has yielded important insights into the structure, function and evolution of TF occupancy patterns across different cell types, differentiation states and environmental conditions. DGF has also been powerfully combined with databases containing defined TF recognition sequences to enable construction and analysis of direct TF regulatory network dynamics[11–13].

Here we review the biophysical basis of genomic footprinting; highlight key experimental, analytical and interpretive considerations with a focus on the human genome; and discuss challenges and prospects for future development.

History of DNA footprinting

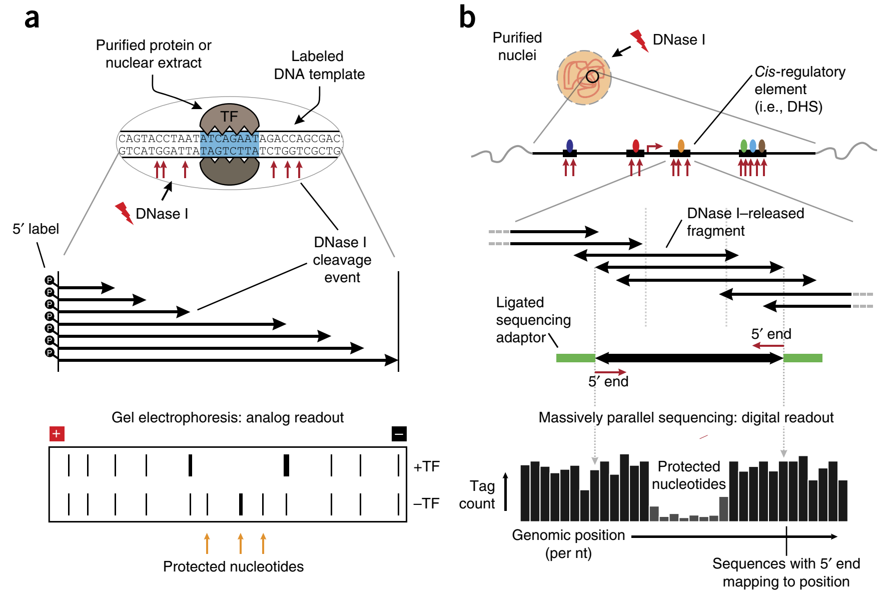
DNA footprinting assays are conceptually simple, comprising two basic steps: (1) exposure of cells or nuclei to a protein or chemical agent capable of cleaving or modifying DNA that exhibits altered reactivity when a DNA-binding protein is engaged, and (2) visualization of the resulting cleavage or modification patterns (together with a relevant control) at single-nucleotide resolution (**Fig. 1a**). The key insight of Galas and Schmitz's classical DNA footprinting assay was that the set of nested DNA fragments released by DNase I cleavage of protein-complexed DNA could be quantitatively visualized using the same direct single-end labeling strategy used in previously developed sequencing methods[14] (**Fig. 1a**).

A vital characteristic of all *in vivo* footprinting reagents is their ability to introduce dense single- or double-stranded DNA cleavages into chromatinized DNA templates. A wide variety of both enzymatic and chemical probes have been used as *in vivo* footprinting

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Altius Institute for Biomedical Sciences, Seattle, Washington, USA. [3]Division of Oncology, Department of Medicine, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to J.V. (jeffrv@uw.edu) or J.A.S. (jstam@uw.edu).

RECEIVED 14 SEPTEMBER 2015; ACCEPTED 13 JANUARY 2016; PUBLISHED ONLINE 25 FEBRUARY 2016; DOI:10.1038/NMETH.3768

Figure 1 | Principles of a DNase I footprinting experiment. (**a**) The classical DNase I footprinting technique was performed *in vitro* and combined purified protein or nuclear extract with a radiolabeled DNA probe. A limited DNase I digestion resulted in a series of nested fragments that were resolved using gel electrophoresis. (**b**) Digital genomic footprinting combines exposure of nuclei to DNase I, purification of small DNase I–released fragments, and massively parallel sequencing of fragment ends (DNase I cleavage sites) to generate a digital readout of per-nucleotide cleavages genome-wide.


a
Purified protein or nuclear extract
Labeled DNA template
TF
CAGTACCTAATATCAGAATAGACCAGCGAC
GTCATGGATTATAGTCTTATCTGGTCGCTG
DNase I
5′ label
DNase I cleavage event
P
Gel electrophoresis: analog readout
+
–
+TF
–TF
Protected nucleotides
b
Purified nuclei
DNase I
Cis-regulatory element (i.e., DHS)
DNase I–released fragment
Ligated sequencing adaptor
5′ end
5′ end
Massively parallel sequencing: digital readout
Tag count
Protected nucleotides
Genomic position (per nt)
Sequences with 5′ end mapping to position

reagents (reviewed in refs. 15–17), including DNase I, copper phenanthroline, dimethyl sulfate, iron(II)-EDTA (hydroxyl radical catalysis) and micrococcal nuclease. DNase I has long been the footprinting reagent of choice for probing DNA-protein interactions because of its ease of use, small size, rapid nuclear penetration, robust cleavage activity, consistency of performance in defined buffer conditions and extraordinary selectivity for non-nucleosomal templates[15], which greatly eclipses its sequence and/or structural preferences[18–21]. It has recently been suggested that transposases such as the hyperactive Tn5 variant used in the ATAC-seq (assay for transposase-accessible chromatin using sequencing) DNA-accessibility assay might have utility as a footprinting reagents[22]. However, this utility remains undetermined, as generalized delineation of individual TF footprints from transposase insertions has not yet been demonstrated. Unlike conventional footprinting reagents, Tn5 is characterized by slow (nearly zero) kinetic turnover and binds its reaction product more tightly than its intermediate substrate, thus potentially inducing its own footprint around its preferred insertion-site sequences[23].

Applied initially to define basic DNA-interaction characteristics of the lac[2] and lambda[24] repressors, DNase I footprinting was widely and rapidly adopted. Since its introduction, footprinting has had critical enabling roles in unveiling the organization and function of regulatory DNA by facilitating the identification of *cis*-regulatory elements and TF consensus sequences[25], the cloning of sequence-specific eukaryotic TFs[26] and the dissection of cooperative TF binding[5]. The advent of direct DNA sequencing from genomic DNA in the mid-1980s[27] enabled the extension of footprinting to specific sequence elements in an *in vivo* context[3–7], albeit with severely limited sensitivity[28,29].

Mechanism of DNase I footprinting

The propensity for DNase I to cleave phosphate bonds in accessible DNA derives from its inherent structural and enzymatic features. At 35 kDa, DNase I is comparable in size to typical TFs and engages only ~5.5 bp (ref. 30), with site selection potentiated by fluctuations in minor groove width and angle[21,30,31]. The enzymatic activity of DNase I is coordinated by the divalent metal cations Ca^{2+}, Mg^{2+} and Mn^{2+} (ref. 32), the availability and relative concentrations of which dramatically alter both reaction kinetics and the propensity for single-stranded nicking versus double-stranded DNA cleavage[32,33]. Because of the aforementioned features, DNase I rapidly became the probe of choice for studies of DNA and chromatin structure[34,35].

By engaging DNA, TFs both protect individual phosphate bonds and alter DNA shape (i.e., groove parameters), such that each bond becomes more or less available for cleavage relative to its unbound state. Classically, DNA footprints have been defined as short stretches of nucleotides that display relative protection from cleavage after protein engagement (**Fig. 1**). However, because DNase I is highly sensitive to the width of the minor DNA groove (which varies with the stacking order of DNA bases[21]), the signature of DNA engagement by some TFs may include potentiation of cleavage between specific bases[36]. Notably, both the cleavage depletion and the potentiation seen in classical footprinting assays have been highly reproducible under similar reaction conditions[15].

Biophysics of DNase I footprinting

Protein-DNA interactions are classically described in terms of affinity and occupancy, where “affinity” refers to the intermolecular force between a protein and a DNA ligand and “occupancy” describes the equilibrium proportion of a DNA template population bound by a protein (**Supplementary Box 1**). In the simplest of cases, affinity and occupancy can be used interchangeably; however, *in vivo* their relationship is complex and depends on many additional parameters such as intracellular protein concentrations, allostery, and binding cooperativity and/or competition[37].

DNA footprints fundamentally reflect the relative occupancy of a protein on its cognate DNA substrate[2,38]. Intuitively, DNase I footprinting reflects the outcome of a competition between a DNA-binding protein and DNase I for access to DNA (**Supplementary Box 1**). For a given TF, detection of a footprint reflects the ratio of that TF’s affinity for a given binding site (and its propensity to bind nonspecifically, which can be modeled using kinetic parameters) versus the relative intrinsic propensity of DNase I to cleave at specific sequences or structures. A critical feature of footprinting is that the affinity of sequence-specific TFs for their cognate binding sites (e.g., ~10^{-9} M for Hox factors[39] or ~10^{-10} M for CTCF[40]) is markedly greater than the affinity

© 2016 Nature America, Inc. All rights reserved.

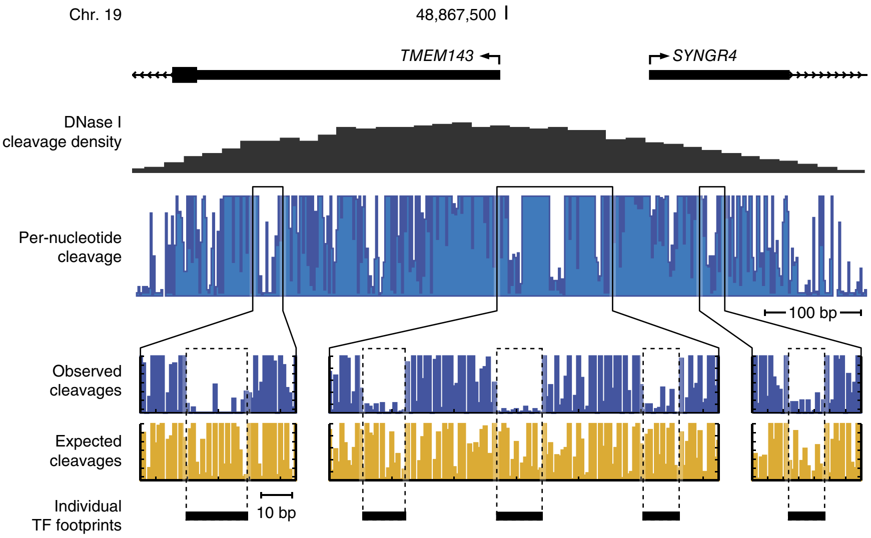


Figure 2 | Resolving *cis*-regulatory architecture at nucleotide resolution in individual regulatory regions. Digital genomic footprints within the promoter of *TMEM143* and *SYNGR4* in regulatory T cells. Dashed boxes highlight individual TF footprints that are marked by decreased cleavage rates (blue) compared with those expected (yellow) considering the intrinsic sequence preference of DNase I.

of DNase I for the same sequences (~10^{-5} M (ref. 41)), resulting in the protection of TF-occupied DNA from nuclease attack. This assumes (i) that the occupancy of a given TF on its genomic template is in thermodynamic equilibrium, which (unlike for protein complexes[42]) has validity in models of TF binding, a largely energy-independent reversible reaction occurring in a preaccessible regulatory site[43], with accessibility remaining remarkably stable within the timescale of a typical DNase I experiment (~3 min); and (ii) that the kinetics of DNase I cleavage are 'single hit', which severely limits DNase I concentration guarantees that the reaction will never proceed to completion and thus that the total number of TF sites available will not change appreciably. Differences in occupancy should in principle be explained by different kinetic parameters of DNA binding and reflected directly in DNase I cleavage data (**Supplementary Box 1**).

Genomic footprinting with a digital readout

Classical DNase I footprinting assays were limited by (i) lack of scale, (ii) the difficulty of quantifying relative nucleotide-level cleavage events, (iii) lack of direct comparability between different regions owing to variable experimental conditions (e.g., as specific activities of different radioactive probes), (iv) lack of standardized detection of footprints and (v) the extreme difficulty of *in vivo* application to intact nuclei. Many of these limitations were surmounted by the advent of DGF[8], which combines massively parallel sequencing and computational analysis to quantify millions of individual DNase I cleavage events in a footprinting experiment (**Fig. 1b**), thus enabling direct quantification of relative cleavage rates at all nucleotide positions genome-wide. First demonstrated in the yeast genome[8], the DGF approach has been applied extensively to human[9,10,12,44–46], mouse[12] and, more recently, plant[11] cells and tissues. In principle, attaining sequencing depths of tens to hundreds of cleavage events within an ~200–400-bp regulatory region provides the basis for robust and systematic detection of DNA-protein interactions; in practice, however, these densities are achieved only in DNase I hypersensitive sites (DHSs).

Visualization of genomic DNA footprinting data

Primary genomic footprinting data are readily visualized in a genome browser as per-nucleotide cleavage, either absolute or normalized to sequencing depth (e.g., per million mapped genomic reads) (**Fig. 1b**). Individual footprints are defined as short stretches of contiguous nucleotides over which cleavage deviates from the expectation (discussed below); typically, these are readily distinguished given sufficient local sequencing depth (**Fig. 2**).

Genomic footprinting data can also be visualized as a TF-centric heat map to display the occupancy spectrum of thousands of individual TF recognition sites, or as aggregated or averaged cleavage profiles. In some cases, these cleavage profiles closely parallel the topology of the protein-DNA interface[8,10].

Figure 3 illustrates (using data from ref. 10) the relationships among TF occupancy, aggregated DNase I cleavage profiles, intrinsic cleavage preferences and individual DNase I footprints. A typical TF may have thousands of consensus recognition sites within DHSs that can be sorted by the ratio of DNase I cleavage occurring in the recognition site versus that in flanking regions and rendered using a heat map of per-nucleotide DNase I cleavage (**Fig. 3a**). In such a rendering, the most highly occupied sites are at the top, and sites with low or no occupation are at the bottom. The aggregated averaged cleavage profiles of measured and modeled DNase I cleavage for the most and least occupied recognition sites differ markedly (**Fig. 3b**). **Figure 3c** shows the same data as **Figure 3a**, but with measured versus expected cleavage propensities superimposed. Subtracting expected from measured cleavage further clarifies the morphology of aggregated cleavage profiles (**Fig. 3d**).

The distinction between aggregated or averaged cleavage profiles compiled from thousands of motif matches and classical DNase I footprints that are individually resolved on the genome has engendered some confusion in the literature[47,48]. Averaged profiles are not true footprints, as they combine occupied and unoccupied templates. Importantly, aggregated cleavage profiles created from thousands of candidate TF recognition sites would be expected to differ markedly for TFs with high versus low average template occupancy. In the former case, the aggregated profile would predominantly reflect occupied sites; in the latter, aggregate profiles would chiefly encompass unoccupied motifs and hence highlight the structural preferences of DNase I. **Figure 3e** demarcates individual *de novo*–defined TF footprints (using the FOS algorithm[10]) in the context of a typical motif match–based aggregated cleavage heat map. Detection of DNase I footprints at such high-occupancy recognition sites is readily achieved for all classes of TF DNA-binding domains (**Supplementary Fig. 1**).

© 2016 Nature America, Inc. All rights reserved.

© 2016 Nature America, Inc. All rights reserved.

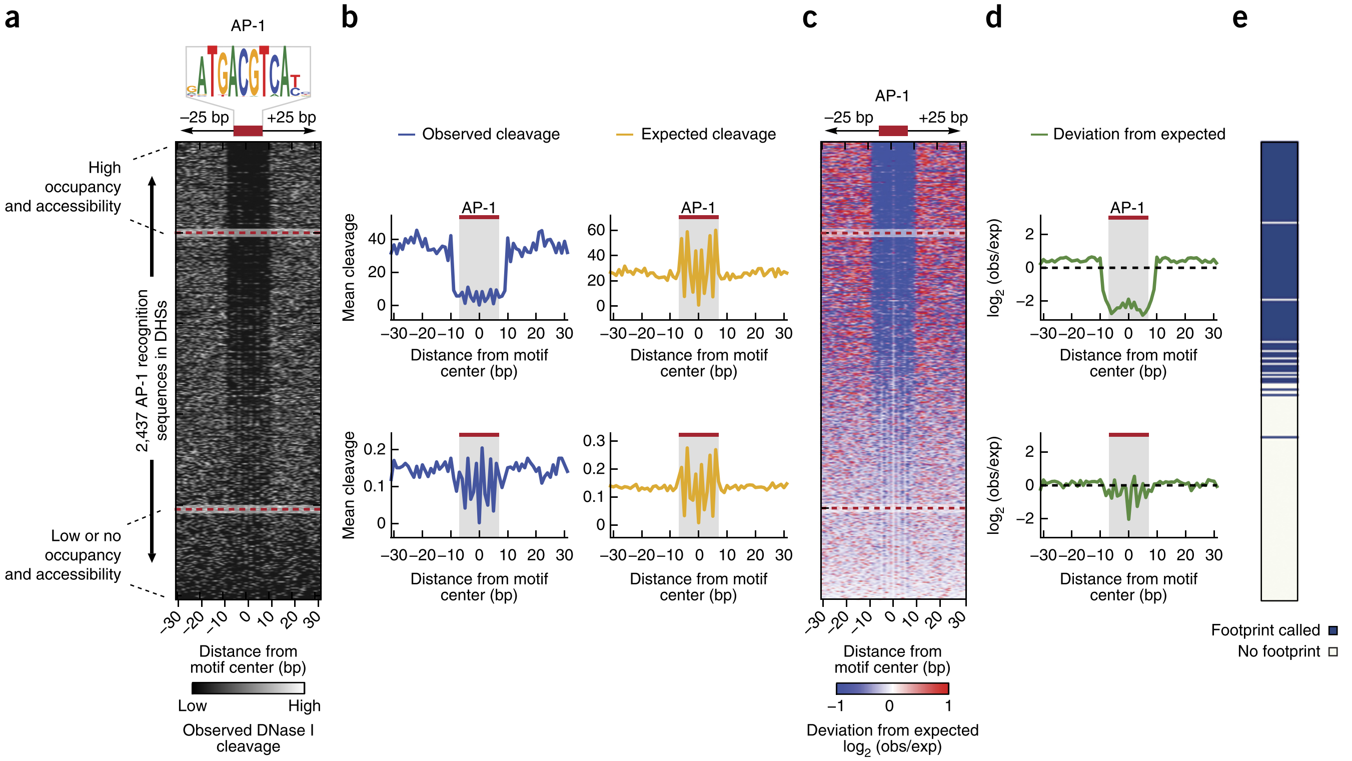


Figure 3 | An illustrative example of interpretation of DNase I cleavage to determine TF occupancy. (**a**) Heat map of per-nucleotide DNase I cleavages surrounding AP-1 recognition sequences within DHSs in regulatory T cells, sorted by decreasing cleavage density in a ±25-bp window. Red dashed lines demarcate the 20% most and least accessible recognition sequences. (**b**) Aggregate profiles of observed and expected mean per-nucleotide DNase I cleavages of the 20% most (top) and least (bottom) accessible recognition sequences. For computation of expected cleavages, observed cleavages were reassigned with respect to a 6-mer preference model. (**c**) Heat map of the ratio of observed (obs) cleavages to expected (exp) cleavages surrounding AP-1 recognition sequences sorted as in **a**. (**d**) Aggregate profiles of the $\log_2$ observed/expected ratio of the 20% most (top) and least (bottom) accessible recognition sequences. (**e**) Footprints identified at AP-1 recognition sequences. Blue tick marks indicate that the recognition sequence has an associated DNase I footprint (from ref. 10).

Experimental considerations

A successful genomic footprinting experiment is predicated on dense mapping of DNase I cleavages in regulatory DNA. This can readily be achieved via deep sequencing of a high-quality DNase I library, wherein a high proportion of all mapped tags (typically >50%) lie within DHSs. Key practical experimental considerations crucial to the success of a genomic footprinting experiment are library quality, complexity, sequencing depth and sequencing mode (**Supplementary Table 1**).

Quality. In a typical mammalian cell type, the core TF binding regions constitute ~1% of the genomic landscape[15,49,50] encompassed within ~150,000 DHSs. Treatment of nuclear chromatin with DNase I releases small (<<125 bp) fragments[51] from DHSs. Poor experimental enrichment of these small fragments results in an excess of uninformative sequencing tags and a low signal-to-noise ratio (SNR). DNase I libraries with a low SNR generally fail to yield discernable footprints even with deep sequencing and should be arrested at an early stage of the quality-control process. A basic metric for measuring SNR is the SPOT score (for “signal proportion of tags”), which is now routinely computed for all DNase-seq data sets produced by major consortia such as the Roadmap Epigenomics[52] and ENCODE[49] Projects. For a DNase I library to be suitable for genomic footprinting, SPOT scores typically should exceed 0.4 (i.e., at least 40% of mapped cleavages within DHSs) and ideally should be greater than 0.5.

Complexity and sequencing depth. Large-scale *de novo* footprint detection in the human genome or one of similar size requires at least 200 million uniquely mapping reads from a high-quality, high-complexity library, with data from the ENCODE project frequently exceeding 500 million uniquely mapped cleavages. Notably, sequencing requirements for simpler tasks, such as determining the occupancy status of specific predetermined TF recognition elements, are estimated to be much lower[53] (~30–60 million mapped reads). DNase I library complexity refers to the proportion of DNA fragments in the sequencing pool derived from unique genomic cleavage events and can be quantified via the incorporation of a unique molecular identifier[54]. Critically, because low-complexity libraries result from experimental manipulations that would be expected to result in skewed fragment populations, they cannot simply be rescued by deeper sequencing and discarding of duplicate reads at the analysis stage. Indeed, caution in removing duplicate reads should be exercised with high-complexity libraries, as DNase I cleavage is stereotypical in nature, leading to the release of identical fragments derived from distinct chromatin templates.

Single- versus paired-end sequencing. Under standard cation conditions (Ca^{2+} or Mg^{2+}), DNase I–released DNA fragments reflect four distinct single-stranded cleavage events. Initial genomic footprinting experiments[8] used single-end sequencing, which, owing to modification of the 3′ end during blunt end generation, enabled mapping of one of the four cleavage events. Paired-end sequencing measures two of the four cleavage events—one at each end of the fragment—providing visibility into joint TF occupancy events on the same chromatin template or joint occupancy of TFs and nucleosomes[51].

Analytical strategies and considerations

Conceptually, analyses of genomic footprinting data can be divided into footprint-centric and TF recognition sequence–centric approaches. The former focus on *de novo* detection and annotation of DNase I footprints, whereas the latter attempt to quantify TF occupancy at defined genomic locations. In contrast to classical *in vitro* footprinting approaches, which routinely incorporated side-by-side naked DNA controls, genomic footprinting experiments frequently lack an empirically derived expected cleavage profile that can properly control for both structural features (particularly chromatin and/or nucleosomal architecture) and structural cleavage preferences. Thus a key analytical consideration is the proper modeling of background cleavage distributions, which is discussed in detail below.

***De novo* detection of TF footprints.** Several algorithms for *de novo* annotation of DNase I footprints have been developed (**Table 1**). A major difficulty inherent to *de novo* footprint detection is that the basic parameters defining a TF-DNA interaction are unknown *a priori* and must be learned simultaneously with the footprint detection process. Many *de novo* strategies make use of a windowing approach for comparing observed cleavage densities in a central region to those within the adjacent flanking sequences (**Fig. 4a**) and commonly use free parameters that define both footprint and flanking-sequence widths, which are fully enumerated during the detection process and greedily selected. For example, the 'footprint occupancy score' approach[10] uses this method to minimize the ratio of total cleavages in a central window to those in two adjacent windows. The Wellington[55] algorithm uses a similar windowing strategy but incorporates information on DNA strands and scores the central window by performing a binomial test using the cleavage rate within the flanking windows as the expected cleavage rate. Differing from these *ad hoc* greedy windowing approaches, additional methods are available that make use of probabilistic frameworks such as hidden Markov models (HINT[56]) and dynamic Bayesian networks (DBFP[57]) to model per-nucleotide cleavage data.

Determining TF occupancy at defined recognition sites. Genomic matches to consensus recognition-sequence models for hundreds of TFs can be generated readily with algorithms such as FIMO[58]. TF occupancy at such defined recognition sites can be quantified under the assumption that occupied recognition elements have significantly different cleavage rates than expected (**Fig. 4b** and **Table 1**). Notably, this task is considerably more simple than *de novo* footprint identification because using consensus motif matches as a prior encodes latent topological features of binding, such as location and the expected width of the of the TF-DNA interaction. Because of this, a common strategy is to explicitly model the per-nucleotide cleavage profile around recognition elements for a given TF and use machine-learning approaches for classification. For example, CENTIPEDE[44] models bound and unbound cleavage profiles using multinomial distributions containing parameters for each nucleotide in the window surrounding recognition elements. The learning strategy used by CENTIPEDE is representative of a broad class of unsupervised approaches for DGF analysis that perform model parameter optimization and classification simultaneously and directly using the observed data (**Table 1**). This approach is contrasted by supervised learning methods, such as FLR[59] and BinDNase[53], that require positive and negative labeled training data (based on, for example, ChIP-seq peaks or other occupancy measures) to fit model parameters that are subsequently used for the discrimination of TF recognition sequences of unknown occupancy status.

Notably, although ChIP-seq can be used to determine the genomic chromatin occupancy pattern of a defined TF, it cannot by itself discriminate direct DNA occupancy from indirect occupancy arising from proximity ligation to another DNA-bound moiety. Currently, direct occupancy is inferred from identified sites that contain a probabilistic match to the consensus recognition motif for the TF. However, this is frequently

Table 1 | Computational methods for DGF analysis

De novo footprint detection	Strategy	Explicit modeling of DNase cleavage preferences
Hesselberth *et al.*[8] (2009)	Binomial distribution of cleavages in a target region versus in flanks; greedy search	No
DBFP[57] (2010)	Dynamic Bayes network	No
Footprint occupancy score[10] (2012)	Minimize ratio of cleavages in a target region versus in flanks; greedy search	No
Wellington[55] (2013)	Binomial distribution of cleavages in a target region versus in flanks; greedy search	No
DNase2TF[48] (2014)	Binomial distribution of cleavages in a target region versus in the local background; greedy search	Yes; naked DNA dinucleotide model
HINT[56] (2014)	Hidden Markov model	No
TF recognition sequence occupancy	**Strategy**	**Explicit modeling of DNase cleavage preferences**
CENTIPEDE[44] (2011)	Hierarchical mixture model using a multinomial to model cleavage profiles; unsupervised learning	No
MILLIPEDE[66] (2013)	Logistic regression; supervised learning	No
PIQ[45] (2014)	Bayesian inference; unsupervised learning	No
FLR[59] (2014)	Mixture model of occupied and unoccupied multinomial cleavage profiles; supervised learning	Yes; naked DNA 6-mer model used as prior
BinDNase[53] (2015)	Logistic regression; supervised learning	No

© 2016 Nature America, Inc. All rights reserved.

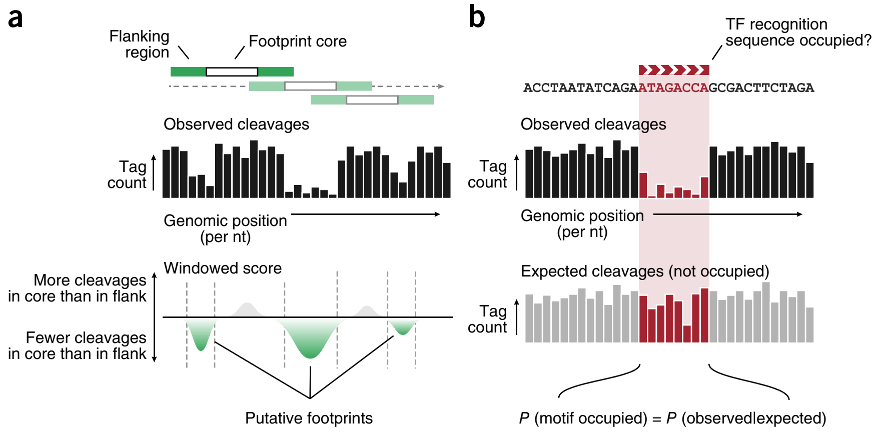

a
Flanking region
Footprint core
Observed cleavages
Tag count
Genomic position (per nt)
Windowed score
More cleavages in core than in flank
Fewer cleavages in core than in flank
Putative footprints
b
TF recognition sequence occupied?
ACCTAATATCAGAATAGACCAGCGACTTCTAGA
Observed cleavages
Tag count
Genomic position (per nt)
Expected cleavages (not occupied)
Tag count
P (motif occupied) = P (observed|expected)

Figure 4 | *De novo* versus TF recognition site–directed analysis of TF occupancy. (**a**) Conceptual strategy for *de novo* delineation of TF footprints in digital genomic data. Footprints are defined when the number of observed cleavages decreases over a short stretch of contiguous nucleotides relative to the number in adjacent flanking regions. Additional corrections for primary DNA structure-directed DNase I cleavage preferences (**Fig. 5**) can be incorporated explicitly or applied in a subsequent step. *De novo* footprint detection is critically dependent on cleavage density over the regulatory region (which in turn depends on sequencing depth and sample SNR; **Supplementary Table 1**). (**b**) Strategy for determining TF occupancy at a predefined candidate recognition element (e.g., defined by a match to a consensus sequence motif). Motif occupancy is determined via application of a statistical framework that assesses the probability of the observed cleavage given an expected model. The total number of cleavage events required for statistically robust categorization of occupancy events is lower than that needed for *de novo* detection because the nucleotide stretch is predefined.

© 2016 Nature America, Inc. All rights reserved.

inconclusive because of the incomplete state and degeneracy of recognition-sequence databases. ChIP-seq can be used to illuminate genomic footprinting data by resolving the assignment of specific TFs to footprints in cases where distinct TFs utilize highly similar recognition elements. However, the high frequency of likely indirect occupancy observed with most TFs, combined with a propensity for artifactual enrichment at highly active loci[60] and poor reflection of binding kinetics[61], precludes ChIP-seq from serving as a true gold standard for the evaluation of footprinting data.

DNase I cleavage preferences. The precise and sensitive detection of TF footprints hinges on the ability to correctly model the expected DNase I cleavage rates at unoccupied regions. This presents a considerable challenge because of the magnitude of variation in accessibility among different DHSs, as well as variability in cleavage rates at adjacent bases due to DNase I shape preference. In principle, a suitable null model would (i) broadly reflect the per-nucleotide cleavage variability observed among adjacent sites, (ii) account for the structural features of DHSs and (iii) correctly model the variance from these two features, with residual variation in the expected model reflecting the protection or potentiation of nucleotide cleavage events by a TF.

Most current computational approaches for footprint analysis model DNase I cleavage in unoccupied DNA as a uniform process, such that each unoccupied nucleotide is expected to have an equal probability of cleavage (**Table 1**). However, DNase I has long been known to exhibit sequence and/or structural cleavage preferences[19–21,62], and the recent availability of extensive cleavage data from DNase I treatment of deproteinated DNA has enabled robust modeling and visualization of these preferences[21]. DNase I cleavage propensities have the potential to vary ~1,000-fold between adjacent hexamers (although the median is <10-fold), and genome-wide, the top 25% most highly cleaved hexamers constitute ~60% of the cleavages in both naked DNA and chromatin (**Supplementary Fig. 2**).

In certain circumstances, DNase I cleavage preferences could potentially affect the delineation of footprints or confound interpretations of aggregated per-nucleotide cleavage patterns[47,48,62]. However, the types and magnitudes of such effects have not been fully delineated, and the impact of sequence preference may differ considerably depending on the analytical strategy. For example, a possible consequence of sequence preference for *de novo* footprint detection could result from a series of overlapping hexamers that each display decreased cleavage preference, which could be further compounded if they are flanked by hexamers with increased preference[48]. However, this situation seems to be rare, as DNase I cleavage bias has only a negligible effect on the classification performance of occupancy at predetermined TF recognition elements[53]. It is also possible that the impact of sequence bias is limited because discriminative features learned from training data used to distinguish occupied versus unoccupied recognition elements implicitly encompass sequence preference information[59]. Nevertheless, relevant to both analytical strategies is the proper handling of sequence preference when interpreting genomic footprinting data.

The per-nucleotide cleavage variance attributable to the intrinsic sequence preferences of DNase I can be modeled effectively *in silico* via cleavage reassignment with respect to the relative sequence preference within small sliding windows (**Fig. 5a**). In contrast to uniform reassignment of per-nucleotide cleavages, reassignment with respect to the 6-mer sequence preference of DNase I within local windows closely tracks the observed data (**Fig. 5b**). However, although sequence preference explains much of the per-nucleotide variability in observed DNase I cleavage counts, a significant amount of variance remains that is not effectively modeled by current footprinting analysis tools.

Overdispersion. Conventional approaches to modeling high-throughput sequencing data rely on a non-negative distribution such as the Poisson to form a basis for the statistical detection of differences in count data. A strong assumption made when using the Poisson distribution in the context of DGF data is that the variance in cleavage rates at individual nucleotides or in a recognition sequence is equal to the mean sequencing depth at the same site. However, the empirical variance in DGF data is much greater than the mean sequencing depth[57,63], even after sequence preference has been accounted for (**Fig. 5c**). As a result, statistical tests that assume equal mean and variance might not be appropriate because of overdispersion, which is likely to lead to overly

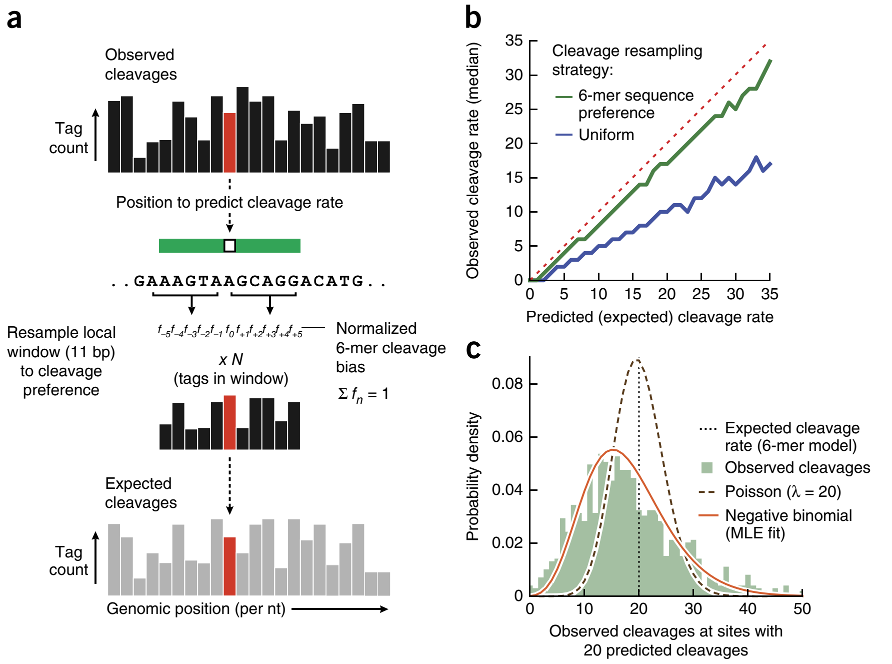


Figure 5 | Modeling variation in DNase I cleavages rates due to primary DNA structure. (**a**) Strategy for determining expected cleavage counts from observed data. Observed cleavage counts are reassigned within a local window (±5 bp) according to the relative sequence preference of the overlapping 6-mer (derived from empirical data in ref. 21), using only the center position as the corresponding expected cleavage rate. The window is stepped in 1-bp increments to compute the expected cleavage rates genome-wide. (**b**) Median observed cleavage rates versus predicted rates using the strategy described in **a**. Local resampling of DNase I cleavages using a sequence-preference model outperforms uniform shuffling. The dashed red line indicates a hypothetical preference model that fully predicts the observed cleavage rates. (**c**) The negative binomial model is more effective at fitting the variation in observed per-nucleotide rates than the commonly used Poisson distribution. Shown is a density histogram of the observed cleavage rates at sites for which there are 20 expected cleavages on the basis of the 6-mer model from **b**. The brown dashed line indicates the probability density of the Poisson distribution with a mean of 20. The orange line shows the probability density of the negative binomial fitted to the observed data.

© 2016 Nature America, Inc. All rights reserved.

optimistic significance values and impede proper calibration of false discovery rates[57].

A robust strategy for managing overdispersion in many types of sequencing data is the use of probability distributions that (unlike the Poisson) model variance independently of the mean. For example, the negative binomial distribution allows for an extra parameter that can be used to account for the variance inherent to DNase I cleavage rates and can be efficiently fitted to represent the observed data (**Fig. 5c**). Accordingly, correctly modeling DGF data with higher fidelity via the incorporation of DNase I sequence preference and accounting for excess experimental variation should vastly increase sensitivity, specificity and reproducibility both for *de novo* footprint identification and for quantifying TF recognition-sequence occupancy.

Challenges and future prospects

Genomic footprinting has provided unique vistas into the organization and function of *cis*-regulatory DNA. However, the technology is still in its infancy, with many unresolved technical, analytical and interpretive challenges. Two key interpretive challenges are the assignment of specific TFs to individual footprints and the recognition and handling of low-affinity binding sites and low-occupation TFs.

Although *de novo* detection of TF footprints can be accomplished without knowledge of the underlying genomic sequence, biological interpretation requires the identification of the causative TF(s). Currently, this is done *post hoc* via comparison with overlapping TF recognition-sequence matches. However, many TFs recognize highly similar recognition motifs, leading to ambiguous assignments. The scope of this problem is presently unclear, as it is unknown to what extent it derives from poorly defined recognition sequences versus truly overlapping recognition specificities. This challenge should be significantly diminished by recent concerted efforts to build comprehensive recognition-sequence databases[64], which are rapidly approaching coverage of the human TF repertoire. Likewise, the systematic derivation of heteromeric TF sequence preferences will undoubtedly illuminate instances where TF assignment is ambiguous because of the intrinsic degeneracy of the TF-to-motif relationship or modulation of recognition specificities by protein-protein interactions[65].

Characterizing the true dynamic range of *in vivo* TF occupancy events presents a further challenge. Although genomic footprinting could in principle simultaneously detect all TF binding in the genome, both classical and more recent studies have suggested that at certain genomic occupancy sites, some TFs, such as the glucocorticoid receptor, may leave very weak or noncanonical footprints that differ in many respects from the signatures of other factors[47,48]. It is likely that most such observations result chiefly from the relative predominance of low-affinity recognition sites; however, other, as yet unclarified aspects of protein structure or behavior might also be contributors.

Though all assays have limits in terms of sensitivity, the aforementioned challenges do not fundamentally limit the application of genomic footprinting. Indeed, there is ample room for experimental and analytical innovation of footprinting methodologies. Certainly, the development of novel computational and statistical methodologies that can precisely model cleavage action will increase sensitivity and the dynamic range for detecting TF occupancy events. Additionally, refinements in molecular methods will likely enable profiling of all four DNase I cleavage events that give rise to DNase I–released fragments, which will substantially affect our understanding of protein-DNA configuration, particularly in the context of the enhanced TF recognition repertoires now on the horizon. One trend clearly in favor of footprinting is the continued fold increases in the throughput of DNA sequencing, which will soon make billion-read cleavage maps routine, with corresponding improvements in both sensitivity and resolution.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
This work was supported by the US National Human Genome Research Institute (grant U54HG007010 to J.A.S.).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ptashne, M. & Gann, A. *Genes and Signals* (Cold Spring Harbor Laboratory Press, 2002).
2. Galas, D.J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
3. Church, G.M., Ephrussi, A., Gilbert, W. & Tonegawa, S. Cell-type-specific contacts to immunoglobulin enhancers in nuclei. *Nature* **313**, 798–801 (1985).
4. Jackson, P.D. & Felsenfeld, G. A method for mapping intranuclear protein-DNA interactions and its application to a nuclease hypersensitive site. *Proc. Natl. Acad. Sci. USA* **82**, 2296–2300 (1985).
5. Zinn, K. & Maniatis, T. Detection of factors that interact with the human beta-interferon regulatory region *in vivo* by DNAase I footprinting. *Cell* **45**, 611–618 (1986).
6. Ephrussi, A., Church, G.M., Tonegawa, S. & Gilbert, W. B lineage–specific interactions of an immunoglobulin enhancer with cellular factors *in vivo*. *Science* **227**, 134–140 (1985).
7. Becker, P.B., Ruppert, S. & Schütz, G. Genomic footprinting reveals cell type-specific DNA binding of ubiquitous factors. *Cell* **51**, 435–443 (1987).
8. Hesselberth, J.R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
9. Boyle, A.P. *et al.* High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–464 (2011).
10. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
11. Sullivan, A.M. *et al.* Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.* **8**, 2015–2030 (2014).
12. Stergachis, A.B. *et al.* Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**, 365–370 (2014).
13. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).
14. Galas, D.J. The invention of footprinting. *Trends Biochem. Sci.* **26**, 690–693 (2001).
15. Gross, D.S. & Garrard, W.T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
16. Tullius, T.D. Physical studies of protein-DNA complexes by footprinting. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 213–237 (1989).
17. Hampshire, A.J., Rusling, D.A., Broughton-Head, V.J. & Fox, K.R. Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. *Methods* **42**, 128–140 (2007).
18. Weiss, B., Live, T.R. & Richardson, C.C. Enzymatic breakage and joining of deoxyribonucleic acid. V. End group labeling and analysis of deoxyribonucleic acid containing single stranded breaks. *J. Biol. Chem.* **243**, 4530–4542 (1968).
19. Ehrlich, S.D., Bertazzoni, U. & Bernardi, G. The specificity of pancreatic deoxyribonuclease. *Eur. J. Biochem.* **40**, 143–147 (1973).
20. Dingwall, C., Lomonossoff, G.P. & Laskey, R.A. High sequence specificity of micrococcal nuclease. *Nucleic Acids Res.* **9**, 2659–2673 (1981).
21. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. USA* **110**, 6376–6381 (2013).
22. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
23. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
24. Johnson, A.D., Meyer, B.J. & Ptashne, M. Interactions between DNA-bound repressors govern regulation by the lambda phage repressor. *Proc. Natl. Acad. Sci. USA* **76**, 5061–5065 (1979).
25. Payvar, F. *et al.* Sequence-specific binding of glucocorticoid receptor to MTV DNA at sites within and upstream of the transcribed region. *Cell* **35**, 381–392 (1983).
26. Dynan, W.S. & Tjian, R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**, 79–87 (1983).
27. Church, G.M. & Gilbert, W. Genomic sequencing. *Proc. Natl. Acad. Sci. USA* **81**, 1991–1995 (1984).
28. Mueller, P.R. & Wold, B. In vivo footprinting of a muscle specific enhancer by ligation mediated PCR. *Science* **246**, 780–786 (1989).
29. Warshawsky, D. & Miller, L. Mapping protein-DNA interactions using in vivo footprinting. *Methods Mol. Biol.* **127**, 199–212 (1999).
30. Weston, S.A., Lahm, A. & Suck, D. X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. *J. Mol. Biol.* **226**, 1237–1256 (1992).
31. Drew, H.R. & Travers, A.A. DNA structural variations in the *E. coli* tyrT promoter. *Cell* **37**, 491–502 (1984).
32. Melgar, E. & Goldthwait, D.A. Deoxyribonucleic acid nucleases. II. The effects of metals on the mechanism of action of deoxyribonuclease I. *J. Biol. Chem.* **243**, 4409–4416 (1968).
33. Campbell, V.W. & Jackson, D.A. The effect of divalent cations on the mode of action of DNase I. The initial reaction products produced from covalently closed circular DNA. *J. Biol. Chem.* **255**, 3726–3735 (1980).
34. Lutter, L.C. Precise location of DNase I cutting sites in the nucleosome core determined by high resolution gel electrophoresis. *Nucleic Acids Res.* **6**, 41–56 (1979).
35. Rhodes, D. & Klug, A. Helical periodicity of DNA determined by enzyme digestion. *Nature* **286**, 573–578 (1980).
36. Stamatoyannopoulos, J.A., Goodwin, A., Joyce, T. & Lowrey, C.H. NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO J.* **14**, 106–116 (1995).
37. Ptashne, M. *A Genetic Switch* (Cold Spring Harbor Laboratory Press, 2004).
38. Dabrowiak, J.C., Goodisman, J. & Ward, B. Quantitative DNA footprinting. *Methods Mol. Biol.* **90**, 23–42 (1997).
39. Pellerin, I., Schnabel, C., Catron, K.M. & Abate, C. Hox proteins have different affinities for a consensus DNA site that correlate with the positions of their genes on the hox cluster. *Mol. Cell. Biol.* **14**, 4532–4545 (1994).
40. Renda, M. *et al.* Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J. Biol. Chem.* **282**, 33336–33345 (2007).
41. N'soukpoé-Kossi, C.N., Diamantoglou, S. & Tajmir-Riahi, H.A. DNase I-DNA interaction alters DNA and protein conformations. *Biochem. Cell Biol.* **86**, 244–250 (2008).
42. Coulon, A., Chow, C.C., Singer, R.H. & Larson, D.R. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat. Rev. Genet.* **14**, 572–584 (2013).
43. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
44. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).
45. Sherwood, R.I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014).
46. Siersbæk, R. *et al.* Molecular architecture of transcription factor hotspots in early adipogenesis. *Cell Rep.* **7**, 1434–1442 (2014).
47. He, H.H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2014).
48. Sung, M.-H., Guertin, M.J., Baek, S. & Hager, G.L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell* **56**, 275–285 (2014).
49. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

© 2016 Nature America, Inc. All rights reserved.

50. Sabo, P.J. *et al.* Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods* **3**, 511–518 (2006).
51. Vierstra, J., Wang, H., John, S., Sandstrom, R. & Stamatoyannopoulos, J.A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat. Methods* **11**, 66–72 (2014).
52. Roadmap Epigenomics Consortium. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
53. Kähärä, J. & Lähdesmäki, H. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics* **31**, 2852–2859 (2015).
54. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
55. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, e201 (2013).
56. Gusmao, E.G., Dieterich, C., Zenke, M. & Costa, I.G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* **30**, 3143–3151 (2014).
57. Chen, X., Hoffman, M.M., Bilmes, J.A., Hesselberth, J.R. & Noble, W.S. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics* **26**, i334–i342 (2010).
58. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
59. Frank, C.L., Crawford, G.E. & Ohler, U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.* **42**, 11865–11878 (2014).
60. Teytelman, L., Thurtle, D.M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. USA* **110**, 18602–18607 (2013).
61. Lickwar, C.R., Mueller, F., Hanlon, S.E., McNally, J.G. & Lieb, J.D. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484**, 251–255 (2012).
62. Koohy, H., Down, T.A. & Hubbard, T.J. Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS One* **8**, e69853 (2013).
63. Hashimoto, T.B., Edwards, M.D. & Gifford, D.K. Universal count correction for high-throughput sequencing. *PLoS Comput. Biol.* **10**, e1003494 (2014).
64. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
65. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
66. Luo, K. & Hartemink, A.J. Using DNase digestion data to accurately identify transcription factor binding sites. *Pac. Symp. Biocomput.* **2013**, 80–91 (2013).

© 2016 Nature America, Inc. All rights reserved.