Published online 15 March 2017

Nucleic Acids Research, 2017, Vol. 45, No. 8 **4315–4329**
doi: 10.1093/nar/gkx174

Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility

Xi Chen[1], Bowen Yu[2], Nicholas Carriero[3], Claudio Silva[2] and Richard Bonneau[1,2,3,*]

[1]Department of Biology, New York University, New York, NY 10003, USA, [2]Department of Computer Science, New York University, New York, NY 10003, USA and [3]Center for Computational Biology, Flatiron Foundation, Simons Foundation, New York, NY 10010, USA

Received June 20, 2016; Revised February 28, 2017; Editorial Decision March 05, 2017; Accepted March 06, 2017

ABSTRACT

Differential binding of transcription factors (TFs) at *cis*-regulatory loci drives the differentiation and function of diverse cellular lineages. Understanding the regulatory interactions that underlie cell fate decisions requires characterizing TF binding sites (TFBS) across multiple cell types and conditions. Techniques, e.g. ChIP-Seq can reveal genome-wide patterns of TF binding, but typically requires laborious and costly experiments for each TF-cell-type (TFCT) condition of interest. Chromosomal accessibility assays can connect accessible chromatin in one cell type to many TFs through sequence motif mapping. Such methods, however, rarely take into account that the genomic context preferred by each factor differs from TF to TF, and from cell type to cell type. To address the differences in TF behaviors, we developed Mocap, a method that integrates chromatin accessibility, motif scores, TF footprints, CpG/GC content, evolutionary conservation and other factors in an ensemble of TFCT-specific classifiers. We show that integration of genomic features, such as CpG islands improves TFBS prediction in some TFCT. Further, we describe a method for mapping new TFCT, for which no ChIP-seq data exists, onto our ensemble of classifiers and show that our cross-sample TFBS prediction method outperforms several previously described methods.

INTRODUCTION

A diverse host of regulatory factors bind to DNA to regulate gene expression and modulate the accessibility, functional status and structure of chromatin (1,2). These factors form complex regulatory networks that underpin diverse patterns of cellular phenotypes (3,4). Understanding the mechanistic basis of this regulatory, and ultimately phenotypic, diversity has remained a fundamental pursuit in biology and requires complete and accurate mapping of TF binding sites. Learning the patterns and processes of TF binding in a condition/cell type-specific manner is a critical step in identifying multi-factor targeted *cis*-regulatory modules (CRM) that are important in cell fate decisions (4,5), and in understanding the causes and consequences of cellular rewiring in gene regulatory networks (6).

Most TFs are sequence specific (7) and TF sequence preferences are known (and readily available from multiple databases) in the form of position weight matrices (PWMs) (8–12). Today's databases of PWMs cover a large number of factors spanning a multitude of species (8,10,11,13). But TFBS identification based solely on PWM sequence matching is known for a number of problems (14). First, the length of a derived PWM is limited by experimental design, making it, in many cases, insufficient in describing the entirety of binding sequence environment. This is especially true in short probe-based technologies, such as protein binding microarrays, which often lead to the identification of incomplete sites or half sites (technology-dependent biases) (15). Shorter PWMs are invariably statistically more prone to false positive discoveries. Another concern is that degenerative sequences are known to be common among TFBS (16), but they are sometimes unaccounted for in classical PWMs, where terminal degenerative regions of a binding motif are removed. Additionally, motifs derived from *in vitro* high-throughput methods for binding site discovery do not capture patterns in the larger region surrounding motif sites that have recently been shown to be of importance for the bindings of a variety of TFs (17,18). Lastly, high-throughput binding analyses carried out *in vitro* necessarily lack cell-type specificity, thus the binding profiles derived from such *in vitro* analyses do not reflect the dynamic chromatin landscape that promotes biologically meaningful binding events.

The lack of cell-type specificity in many TF binding assays is ameliorated by *in vivo* studies, such as ChIP-Seq. However, ChIP-Seq is carried out one TF and cell-type condition at a time, and its feasibility is often limited by factors such as the requirement to obtain high number of cells as input materials and the availability of high quality antibod-

[*]To whom correspondence should be addressed. Tel: +1 212 998 9516; Email: bonneau@nyu.edu

ies (19,20). With the large number of factors and cell types, and the dynamic nature inherent in all TF binding events, it is challenging to capture the full scope of regulatory interactions for all factors and conditions with ChIP-Seq or any similar single-TF directed experiment.

Eukaryotic transcriptional regulation is shaped by chromatin dynamics, where accessible chromatin sets the stage for various types of regulatory interactions. Experiments that interrogate chromatin accessibility, such as digital genomic footprinting (DGF), DNase-Seq, ATAC-Seq and FAIRE-Seq have been used as promising alternatives to factor-specific ChIP-Seq for the identification of TFBS (21–24). Because chromatin accessibility and nucleosome positioning are critical players enabling both the binding of TFs and the subsequent relay of regulatory information, such as co-factor recruitment and transcriptional machinery assembly, chromatin accessibility-based TFBS prediction methods has allowed cell type-specific predictions of binding sites for many TFs with a single experiment per cell type (25–30). In spite of these advantages, the size and complexity of the mammalian genome, the diversity of TF behaviors (some TFs bind exclusively to nucleosome-free regions while others pioneer nucleosome-bound regions) and the large range of cell types (cell types modulate TF activity, TF-TF interactions and chromosome structure) make large-scale multi-cell type multi-TF binding site inference difficult, especially in a manner that balances method sensitivity and selectivity (31–33).

To address these challenges, we designed a TFBS prediction method that uses sequence-derived genomic features and one chromatin accessibility experiment per cell type to profile TFCT-specific binding activities. Our method has three components: (i) MocapG, a generic unsupervised method that ranks binding probabilities of accessible motif sites based on local chromatin accessibility, (ii) MocapS, which integrates the motif-associated accessibility scores of MocapG with additional genomic features, such as TF footprints, CpG/GC content (sequence features including CpG content, GC content and CpG island), evolutionary conservation and the proximity of TF motifs to transcription start sites (TSS) to train TFCT-specific predictive models under the supervision of ChIP-Seq data and (iii) MocapX, which extends the selectivity of MocapS to more factors and cell types by mapping new TFCT conditions based on genomic feature distance to a nearest TFCT neighbor trained MocapS model using weighted least squares regression. The similarity-weighted ensemble prediction method, MocapX can connect TFCT-specific TFBS prediction models to TFCT pairs not directly queried using ChIP-seq or related methods. This cross-sample prediction framework, although limited to the scope of factors and cell types modeled, addresses the differences between TFCT conditions in TFBS prediction in a data-driven manner, and has the potential to expand the repertoire of putative TFBS with improved accuracy to any factors we have motif information for and in any cell type where chromatin accessibility data is obtainable.

Additionally, we established a cross-assay comparison between model-based predictions using DNase-Seq and ATAC-Seq, in an effort to enable similar binding-site predictions from both of these widely adopted genomic technologies. In building a TFBS prediction method that learns and uses the differences between TF-chromatin interaction patterns, we hope to provide tools that help reveal the mechanistic complexity of mammalian gene regulation and chart the mammalian regulatory landscape spanning multilineage differentiation (Figure 1).

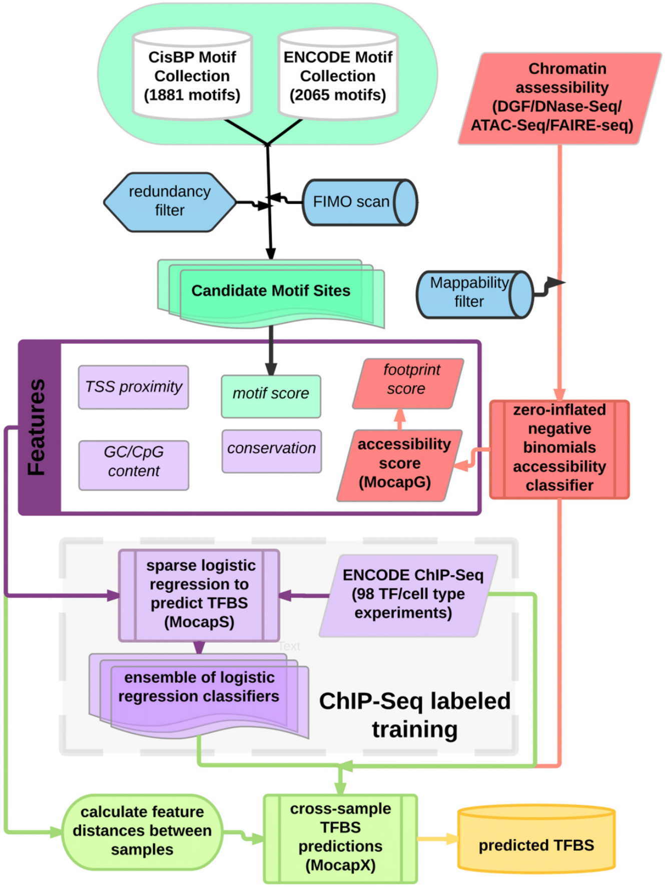
CisBP Motif Collection (1881 motifs)
ENCODE Motif Collection (2065 motifs)
Chromatin assessibility (DGF/DNase-Seq/ ATAC-Seq/FAIRE-seq)
redundancy filter
FIMO scan
Candidate Motif Sites
Mappability filter
Features
TSS proximity
motif score
footprint score
GC/CpG content
conservation
accessibility score (MocapG)
zero-inflated negative binomials accessibility classifier
sparse logistic regression to predict TFBS (MocapS)
ENCODE ChIP-Seq (98 TF/cell type experiments)
ensemble of logistic regression classifiers
ChIP-Seq labeled training
calculate feature distances between samples
cross-sample TFBS predictions (MocapX)
predicted TFBS

Figure 1. Our TFBS prediction pipeline. We compiled a non-redundant set of TF binding motifs, and compute genomic features for all candidate motif sites. We trained sparse logistic regression models to predict binding sites (MocapS) for 98 TFCT conditions, for which ChIP-Seq data is available in ENCODE cell type K562, A549 and Hepg2. True binding sites are defined as motif sites that overlap ChIP-Seq peaks. For a new TFCT condition, binding sites are inferred from either the unsupervised accessibility classifier (Mocap) or a trained sparse logistic regression classifier according to sample mapping using weighted least squares regression (MocapX). Shaded area stands for supervised training steps; unshaded area are steps for data acquisition (top) and making predictions (bottom).

MATERIALS AND METHODS

Obtaining candidate binding sites from motif collections

Human TF motifs (PWMs) were downloaded from the ENCODE motif collection (http://compbio.mit.edu/encode-motifs) and the CisBP motif database (http://cisbp.ccbr.utoronto.ca) (9,10). We combined information from the two motif collections and filtered PWMs representing the same TF using pairwise comparisons based on normalized Euclidean distance (detailed in supplemental materials). The resulting non-redundant set of PWMs was then used to scan the human genome (hg19 assembly) to obtain candidate motif sites genome-wide using FIMO from the MEME Suite with options –max-strand –thresh 1e–3 (34). Overlapping motif sites (where at least half of a motif site overlaps with an adjacent motif of greater or equal length) are further cleaned to keep the motif site with a more significant matching score. Additionally, we excluded motif sites that overlap an ENCODE blacklisted region from downstream analyses (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/).

Chromatin accessibility and ChIP-Seq data processing

DGF and DNase-Seq data were downloaded from ENCODE as aligned reads (35). We filtered out reads with mapping quality <25 and limited the number of mapped cuts per base pair to 50 to reduce the duplication effect caused by technical artifacts. DNase cut counts centered around each motif site were extracted from the processed BAM files with customized scripts (36), and all sites were cleaned with ENCODE mappability tracks prior to modeling to exclude unalignable regions of the genome from downstream analyses.

ATAC-Seq data for Gm12878 was obtained from Buenrostro *et al.* (23). We selected the experiment with the highest sequencing depth (SRR891268) to allow for sufficient sequencing depth for footprint detection. Reads were aligned using Bowtie with the option –best -X2000 -m1. As mentioned in (23), to extract the cut sites from ATAC-Seq BAM files, we offset the + strand read fragment by +4 bp, and the—strand by –5 bp. Similar mapping quality, mappability and 50 reads per base pair upper limit constraints were applied to the ATAC-Seq dataset.

ChIP-Seq broadpeak and narrowpeak tracks were downloaded from ENCODE. In case of multiple ChIP-Seq experiments for the same factor and cell type, tracks were merged to keep the intersections of all available experiments. Tracks flagged based on the quality metrics provided by the ENCODE consortium were excluded (https://genome.ucsc.edu/ENCODE/qualityMetrics.html) (37).

The MocapG model

To obtain cell type-specific accessibility features associated with each motif site, we built a probabilistic model that classifies motif regions with a given number of cuts as either accessible or inaccessible. Briefly, we fit genome-wide accessibility cut count as a mixture of two negative binomial distributions and an additional zero component representing, respectively, accessible, inaccessible and zero-inflated regions of the chromatin.

$$P(c) = \pi_1 P(c|s=1) + \pi_2 P(c|s=2) + \pi_0 P(c|s=0)$$

where

$$\begin{aligned} P(c|s=1) &= F(c|\alpha_1, \tau_1) = \tfrac{\Gamma(c+\alpha_1)}{c!\Gamma(\alpha_1)} \tau_1^{\alpha_1} (1-\tau_1)^c \\ P(c|s=2) &= F(c|\alpha_2, \tau_2) = \tfrac{\Gamma(c+\alpha_2)}{c!\Gamma(\alpha_2)} \tau_2^{\alpha_2} (1-\tau_2)^c \\ P(c|s=0) &\qquad\qquad = \mathbb{1}(c=0) \end{aligned}$$

c is the number of DNase I cut count 100 bp upstream and 100 bp downstream of a specific motif site excluding the motif site itself. α, τ are the mean and dispersion parameters of negative binomial distributions respectively. π_0, π_1 and π_2 correspond to the probability of the zero (s_0), inaccessible (s_1) and accessible (s_2) component, where $\pi_0 + \pi_1 + \pi_2 = 1$. Model parameters were estimated with an Expectation-Maximization (EM) algorithm, which takes a set of random initial parameter values (where $\alpha_1 \ll \alpha_2$ and $\pi_1 > \pi_2$ to avoid label switching) and genome-wide cut count in 200 bp windows as input and outputs model parameters that maximize the log-likelihood function $\log(P(c))$. Each motif region was then assigned a binary accessibility indicator S and a log likelihood score L based on the probability ratio of the region being accessible to inaccessible.

$$L(c) = log \frac{\pi_2 P(c|\alpha_2, \tau_2, \pi_2)}{\pi_1 P(c|\alpha_1, \tau_1, \pi_1)}$$

$$S(c) = \begin{cases} 1 & exp(L(c)) \leq 2 \\ 2 & exp(L(c)) > 2 \end{cases}$$

An $S(c)$ of 2 corresponds to a likelihood ratio where a motif region is at least twice as likely to be accessible than inaccessible.

Calculating other motif-associated genomic features

For all accessible motifs, we assessed the probability that a footprint profile exists around the motif site using a pair of binomial tests adapted from (28). For DNase-Seq, the binomial tests yield scores for strand-imbalance and depletion of DNase I cuts at motif sites. For ATAC-Seq, strand-imbalance is insignificant because of the absence of a size selection step. We merged strand information to assess if a motif region is depleted of cuts as compared to the left and right flanking region respectively. Detailed footprint score calculations are provided in the supplemental methods.

PhastCons and PhyloP scores were calculated for each motif region from conservation tracks downloaded from USCS (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19). TSS proximity features were calculated based on RefSeq Genes. Repeats were labeled for each motif region based on RepeatMasker calls (38). Mapability tracks were downloaded from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/ ENCODE and selected based on corresponding DNase-Seq/ATAC-Seq sequencing length. Mapability scores, GC and CpG content were computed for a region 100bp upstream and downstream of each motif site. CpG islands were calculated as defined in (39,40) (detailed in the supplemental methods). We provide customized scripts for all motif-associated feature calculations.

Training sparse logistic regression classifiers to predict TFBS (MocapS)

We built sparse logistic regression models using the LIBLINEAR package in R to classify motif sites as true (overlapping with ChIP-Seq peaks) or false (not overlapping with ChIP-Seq peaks) binding site based on a list of genomic features and their interaction terms (41).

For all the motif-associated genomic features, we apply a correlation filter to retain, for each trained TFCT condition, 10 features that are most highly correlated with the ChIP-Seq signals based on the Pearson correlation coefficient (PCC). A second correlation filter was applied to the top 10 features and their second-order interaction terms to retain 30 most correlated features. Each TFCT condition was trained on a stratified sample of motif sites representing data from all except the hold-out chromosomes.

We adopt L1 regularization and tune the shrinkage parameter for each TFCT condition by performing 10-fold cross validation to optimize the area under the precision and recall (AUPR) curve. To avoid overfitting, the shrinkage parameter was further tuned such that the resulting classifier is sparser than an optimally performing classifier and still yields a near-optimum (within one standard error of the optimum) cross-validation performance. The final model parameters are estimated by aggregating over 100 bootstrap runs of such sparse logistic regression model fitting for each TFCT condition to reduce estimator bias. Model performance was assessed with data from the held-out chromosome. More details are provided in the supplemental methods.

Cross-sample TFBS prediction (MocapX)

To expand binding site prediction to TFCT conditions where ChIP-Seq data is unavailable, we use robustly weighted least squares regression to derive mapping vector β to match new samples (samples without ChIP-Seq) to trained models for TFBS predictions. The regression problem is such that the derived sample weight $\bar{w}$ and mapping vector $\bar{\beta}$ minimize the following error function

$$\min \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij}\left(\gamma log \frac{Y_{ij} + Y_{ji}}{2} - \sum_{k=1}^{p} X_{ij,k}\beta_k\right)^2$$

$$= \min \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij}(e_{ij})^2$$

where n is the number of TFCT conditions; p is the number of features; $X_{ij,\,k}$ is the binary probability that feature k from TFCT i and TFCT j are derived from the same distribution based on KS test for continuous features and Chi-square test for categorical features; Y_{ij} is the predictive performance (AUPR) of model trained in TFCT i when applied to classify motifs in TFCT j; γ is a hyperparameter that binarizes the target variable Y into sample pairs that cross-predict well and ones that cross-predict poorly; sample pair weight \$w\$$_{ij}$ is fitted through iteratively re-weighted least squares regression using the rlm function in the MASS package in R with the Tukey's bisquare family psi functions (42,43).

We use $\hat{w}$ and $\hat{\beta}$ to compute a sample mapping $i \rightarrow j$ that maximizes $\hat{Y}$ as below

$$\hat{Y} = \arg\max_{i \rightarrow j} \hat{w}_{ij} \sum_{k=1}^{p} X_{ij,k}\hat{\beta}_k$$

Mapping is then assigned if and only if $\hat{Y}$ compares favorably with Y_{MocapG} in TFCT i. More implementation details are provided in the supplemental methods.

Comparison to other TFBS prediction methods

PIQ and CENTIPEDE were given the same motif input and accessibility data (25,26). We also obtained, for CENTIPEDE, PhastCons conservation scores (based on 46-way alignment of placental mammals) and distance to TSS for prior calculations. Because PIQ only makes predictions to sites that carry significant footprint profile, to allow a more complete comparison, we assign unpredicted sites the lowest posterior scores in PIQ. Excluding the unpredicted sites from comparison yields similar results (Supplementary Figure S7E-G). Comparisons are based on three metrics: AUPR, sensitivity at 1% false positive rate (FPR) and areas under the receiver-operating characteristic (AUROC) curve. A peak-level performance comparison is also done to measure the areas under the precision-peak recall curves.

RESULTS

Compiling a large unbiased set of candidate motif sites improves the sensitivity of TFBS prediction

The initial step of our computational pipeline involves the use of a large set of motifs from multiple sources to scan the genome to obtain TF-specific candidate binding sites. There is a many-to-many relationship between TFs and motifs representing them, as some motifs are derived using DNA-binding domains that are shared among several TFs (e.g. the motif for SP7 and SP9 are directly transferred from SP8 because of deep DNA-binding domain homology), whereas some TFs show diverged binding preferences that are matched by two or more distinct motifs (many TFs, e.g. NFKB, have a shared canonical motif accompanied by several non-canonical motifs) (10,15,44,45).

Among the 857 TFs represented by the motif collection, over half have three or more motif representations. The large number of motifs representing the same TF is resulted from either the intrinsic diversity in binding preferences or simply the differences in techniques used to derive them. For example, widely studied TFs, such as REST and CTCF, also tend to have large sets of motif representation, and consequently a large amount of redundancy between motifs (Supplementary Figure S1B). To remove redundancy and resolve these many-to-many mappings prior to predicting binding sites, we introduced a similarity index based on normalized Euclidean distance. We compared the distance between vectorized motifs that represent the same TF and then kept the motifs from each similar cluster as described in the methods section (Supplementary Figure S1).

We used a loose threshold (P-value < 1e–3) to scan for candidate binding sites. This lower threshold does incur a sizable computational cost, but we have found that this high level of inclusiveness is important for the following reasons. First, well-matched motifs only imply TF binding in a minority of cases. Non-concensus, low-affinity binding sites are sometimes required for proper functional readouts (46,47). We found for the majority of TFs, motif matching scores only weakly correlate with ChIP-Seq signal (average PCC is 0.05) (Figure 3A). It was also shown through various studies for many different TFs that PWM matching score is a poor predictor of binding selectivity (25,48). Thus using motif score to preselect for candidate binding sites will lead to sampling bias. Secondly, the number of motif occurrences or the occurrence of clustered motif sites are often important indicators of functional and tissue/cell type-specific binding. Many functional regulatory sites, especially in higher eukaryotes are found in clusters, either of themselves (homotypic clusters) or with each other (e.g. super enhancers) (47,49–51). The fact that degeneracy within a binding site is often tolerated through evolutionary reshuffling, whereas the prevalence of fusion sites and larger functional CRMs persists suggests that the complexity of mammalian TF binding lies, not in the qualitative matching of individual site, but likely emerges from the grammatical arrangement of sites (e.g. enhancer units) (47,52–54). Lastly, not all binding events require the physical interaction between TF and DNA molecule, thus a sequence-specific binding site may be absent at various sites when instead of binding directly and individually to the DNA, TF binds as a part of a protein complex or interacts with parts of the chromatin via chromatin-modifying enzymes (55,56). For all of these reasons loosening the motif scan threshold can help achieve more complete coverage of binding sites on a genome scale (Supplementary Figure S2) and subsequently allow more relevant genomic features, such as chromatin accessibility patterns, to drive the binding site prediction (Figure 3).

Classifying chromatin accessibility landscape with a mixture model of zero-inflated negative binomials recovers patterns of multilineage differentiation

Chromatin accessibility data are rich in cell type-specific regulatory information. To create a baseline method to rank motif sites based on chromatin accessibility (MocapG), we modeled chromatin accessibility states as a mixture of zero-inflated negative binomials, where two distributions for both accessible and inaccessible components of the chromatin are approximated with an EM algorithm. The negative binomial distribution was chosen because it models well the overdispersion commonly found in next-generation sequencing data (25,57,58), and has the flexibility to be applied to genomic regions of various sizes. This choice of distribution is also general enough to describe the signal-to-noise patterns associated with different experimental protocols for measuring accessibility (DGF, DNase-Seq, ATAC-Seq and FAIRE-Seq) (Supplementary Figure S3). Further, we added a zero component into the mixture to assess the amount of zero inflation due to the lack of sequencing depth, e.g. those often found in FAIRE-Seq experiments. For a given accessibility experiment, we learned the component parameters from a cut count (for DGF, DNase-Seq or ATAC-Seq) or fragment count (for FAIRE-Seq) distribution randomly sampled from the mappable regions of the genome. The accessibility for a given genomic region (e.g. a motif region) is then decided based on the log likelihood ratio between the accessible and inaccessible components inferred (Figure 2A).

To test the mixture model' s ability to classify chromatin accessibility and distinguish between developmental cell types, we obtained ENCODE chromatin accessibility data (DGF or DNase-Seq) for 41 normal human cell types at various stages of development. For each cell type, we binned the genome into 400 bp windows, and used our model to classify regions of the genome into a binary profile of ones (accessible) and zeros (inaccessible). We then calculated pair-wise Euclidean distance between cell types and performed hierarchical clustering on the resulting accessibility profile. As shown in Figure 2B, cell types fall into clusters based on germ layer origin, confirming the utility of the classification method across experiments with various sequencing depth. The result also exemplifies the key role chromatin accessibility plays in coordinating cellular differentiation and directing cell fate decisions (59). MocapG was then used to generate our main motif-associated accessibility features in logistic regression training.

TF footprint profiles are condition-dependent

Because of the overarching role of chromatin accessibility in directing TF regulation and distinguishing cell type identity, we next sought to model footprint patterns surrounding accessible motifs to determine if there is evidence of direct physical binding of a TF to each candidate motif site. Using a pair of binomial tests, we assessed whether a footprint profile exists (if the motif site is depleted of cut count as compared to its left and right flanking regions respectively). To examine the strand-specific patterns of footprint scores across TFCT conditions, we plotted the averaged cut count profile surrounding motifs found in ChIP-Seq peaks (Supplementary Figure S6, Appendix). Due to the occlusion of the immediate flanking nucleosomes, the 5′ cut distribution at binding sites tend to be strand-biased, with the positive-strand cuts more likely to accumulate on the left flanking region of a site, while negative-strand cuts more on the right (28). The fragment size selection step in DGF protocol further accentuate this strand imbalance towards the center of binding sites when compared to ATAC-Seq (size selection step is absent from a typical ATAC-seq protocol) (21,23). We thus adopted a strand-specific footprint detection method for DNase I footprints detection and a non-strand-specific method for ATAC-Seq footprint detection (see Materials and Methods).

Previous work has pointed out limitations in footprint modeling, showing that the footprint signature is subject to several systematic biases including enzyme cutting and sequence bias (30,60–64). Corroborating these observations, we found that DNase I cuts tend to result in more conspicuous footprint than Tn5 transposon insertion cuts, as shown in the differences in cut count profiles of factors, such as CTCF and RAD21 (Supplementary Figure S6A-B). As the behavior of these two factors does not vary significantly among cell types measured by DGF (K562, A549 and Hepg2), the observed differences between DGF-measured and ATAC-Seq-measured cell types are likely *bona fide* differences between the enzymatic activities of DNase I used in DGF and Tn5 transposase used in ATAC-Seq (60).

Aside from technical complications, we observed high variability in the shapes and patterns of DNase I footprints

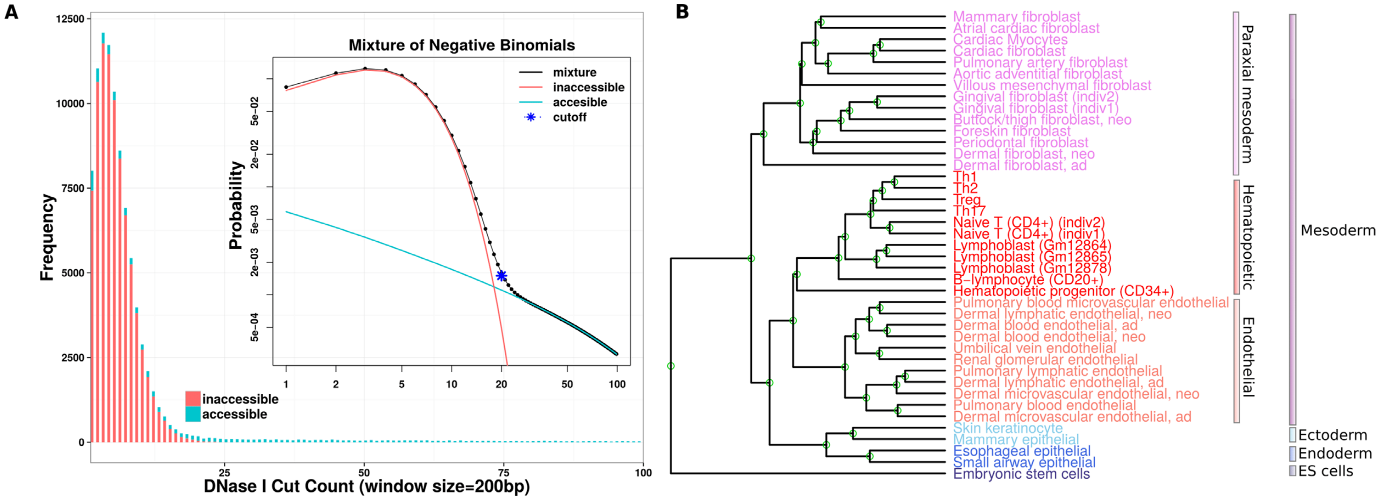

Figure 2. Modelling DNase I cut count as a mixture of negative binomial distributions. (**A**) Distribution of DNase I cut count simulated using zero-inflated negative binomial model parameters derived using an EM algorithm (n = 100 000). Red: cut count from inaccessible regions of the chromatin; blue: cut count from accessible regions of the chromatin. Inset: Cutoff point is determined by the probability ratio between accessible and inaccessible components. X and Y axes are in log scales. (**B**) Hierarchical clustering of the accessibility landscape of ENCODE cell types. Genome is binned into 400 bp (overlapping by 200 bp) windows, and the accessibility of each genomic window is classified using the zero-inflated negative binomial mixture model as 1 s (accessible) and 0 s (inaccessible). Cell types cluster in accordance with their developmental origins.

across factors and cell types. Some factor-specific signatures are attributed to the sequence properties of the motif and/or the binding behavior of the TF. For example, FOXA1, FOXA2, MXI1, CEBPB, NFYA and RFX5 show DNase I footprints that are center-inflated rather than exhibiting a canonical center-depletion (Supplementary Figure S6, Appendix). These factors tend to have A/T rich motifs and three out of the six TFs (FOXA1, FOXA2, NFYA) have been previously implicated as pioneer factors (26,65–68). These inflated footprints might suggest transient binding dynamics, due to the competition between destabilized nucleosome and TFs, or between interacting TFs. Some TFs, e.g. ZNF384 and ARID3A have highly noisy cut patterns surrounding motifs. This transient binding might be due to either poor motif/ChIP-Seq quality or the absence of consensus sequence preference, as these motifs also correspond to the samples with incomplete ChIP-Seq coverage (Supplementary Figure S2). Differences in footprint profiles can also result from cell type-specific TF activities, e.g. due to the presence or absence of a co-binder. Because ChIP-Seq cannot distinguish direct from indirect binding, a TFCT condition lacking an average footprint profile might suggest the dominating effect of indirect binding or low affinity binding of the TF in the tested cell type.

In addition, the lack of an obvious footprint could also be due to the fact that our motif-centric method, by definition, requires footprint discovery be anchored around the center of a known motif, which might not apply to all TF motifs. This is especially true for clustered binding sites where the occurrence of multiple motifs (10–30 bp) within one ChIP-Seq peak region (∼200 bp) makes it difficult to pinpoint with either motif matching, accessibility level or footprint profiling which narrow region corresponds to the actual physical binding of the TF (69). Although centering footprint detection around motifs can absorb part of the variation seen across factors and cell types, the diversity of TF-specific footprint profiles necessitates developing more flexible and condition-specific footprint detection methods, especially for ATAC-Seq footprints.

Building TFCT-specific binding site models

To build condition-specific TFBS prediction models, we selected 19 motif-associated genomic features and trained sparse logistic regression classifiers (MocapS models) under the supervision of ChIP-Seq for 98 TFCTs (representing 52 TFs). Across the samples, local chromatin accessibility scores are the most consistently correlated with ChIP-Seq signal (Supplementary Figure S4) and also the most consistently chosen by our model selection procedure. Here we defined local chromatin accessibility as a relatively small region 100 bp upstream and 100 bp downstream of a motif site, excluding the motif site itself, as we found chromatin accessibility tends to taper off after 100 bp and the motif site itself tends to be depleted of DNase I cuts because of TF footprint (Supplementary Figure S6). In comparison to local chromatin accessibility scores, our binary accessibility feature based on the negative binomial mixture model and footprint scores correlate to ChIP-Seq in a less consistent manner (Supplementary Figure S4).

GC/CpG content-associated features, namely, GC content, CpG count and CpG islands also exhibit more divergence among TFCTs, with TFs, such as E2F6 and HEY1 showing ChIP-Seq correlation to GC/CpG content comparable with accessibility features, while GC/CpG content around MAFK and MAFF motifs are barely correlated with ChIP-Seq (Figure 3A and B, Supplementary Figure S4). Additionally, a binary feature separating motif site based on TSS proximity (1kb) tends to be more correlated with ChIP-Seq signal than the continuous feature distance to TSS (Figure 3A and B, Supplementary Figure S4). This suggests a non-linear relationship exists between the distance of *cis*-element to a nearest TSS and its regulatory

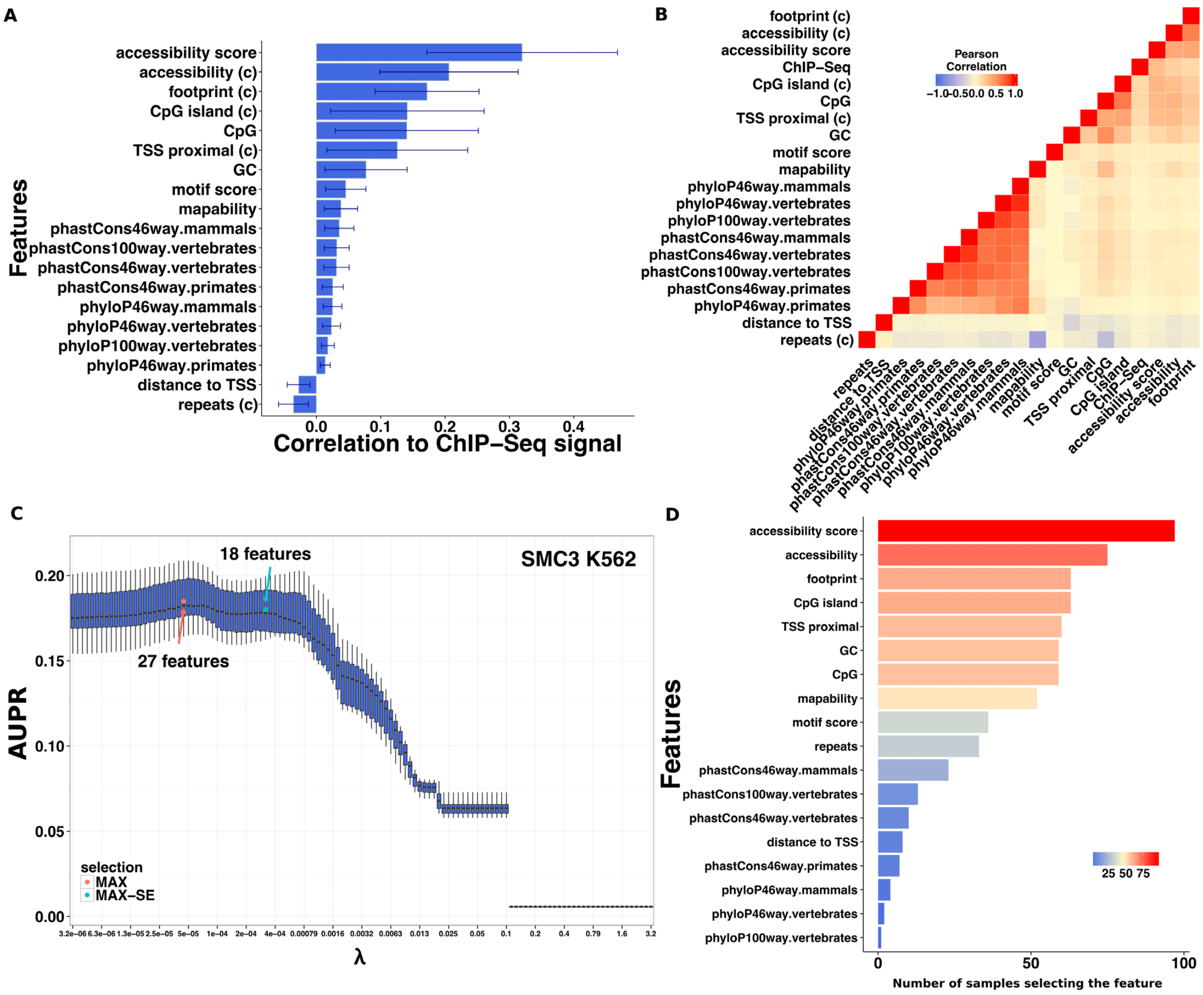

Figure 3. Feature selection and classifier training. (**A**) Genomic features ranked by their correlation to the ChIP-Seq signal. Barplot showing the average correlation for each genomic feature over 98 TFCT samples. Error bars mark average ±- one standard deviation. (**B**) Clustering heatmap showing PCC between genomic features across motif sites. Red: positive correlation, white: no correlation, blue: negative correlation. (**C**) Ten-fold cross-validation performance (AUPR) while adopting different shrinkage parameters λ. We tune the shrinkage parameter to approach maximum AUPR. Red dot marks the shrinkage level (sparsity) that corresponds to the maximum 10-fold cross-validation performance. Green dot corresponds to our selected feature combination–the sparsest model that achieved a near optimum (within one standard error of maximum) cross-validation performance. Example TF: SMC3, cell type: K562. (**D**) Barplot showing the number of times each feature is selected in the 98 trained models. Bar colors are scaled. Red and blue corresponding to more and less commonly selected features respectively.

activity, and that for some TFCT conditions, a disproportional majority of regulatory interactions take place in TSS proximal sites.

Evolutionary conservation scores only weakly correlate with ChIP-Seq signal, despite the common assumption that functional *cis*-regulatory elements are more likely to be conserved (Figure 3A and B and Supplementary Figure S4). Among the two types of conservation scores and three different evolutionary distances we tested, phastCons scores, which measure the probability of a motif site belonging to a conserved element seemed to be more predictive of TFBS than phyloP scores that attempt to capture accelerated rate of evolution. Compared to vertebrate and primate conservation scores, mammalian evolutionary conservation scores were found to be the most correlated with and predictive of TF binding (Figure 3A and D). Mapability scores and repetitive sequences both appeared to correlate with CpG features and accessibility and likely confound binding site predictions (Figure 3B).

To incorporate relevant features in binding site prediction and preclude spurious signals, we filtered genomic features based on their correlation to ChIP-Seq signal (detailed in Materials and Methods) before subjecting them to the sparsity constraints in logistic regression (Figure 3A). Further, we included interaction terms, reasoning that less directly correlated features, such as TSS proximity and evolutionary conservation scores could have modulating effects on TF binding. We used L1-regularization to constrain model sparsity, selecting models that achieve good cross-validation AUPR but also have the potential to generalize well out-of-sample (Figure 3C). For the 98 samples we trained on, there is a general agreement between a feature's correlation with target variable (ChIP-Seq) and the likelihood of the feature being selected into the final model (Figure 3A and D). Among the individual and interaction features, we found features involving local chromatin accessibility to be the most widely selected predictor of TF binding across TFCT conditions (Figure 3D, Supplementary Figure S5).

GC/CpG content surrounding TF motifs modulates TF binding

Sequence features affect TF binding through mechanisms such as altering local DNA shape, affecting nucleosome positioning, DNA methylation and/or influencing co-binding sites (17,70–72). In contrast to the good generalization properties of chromatin accessibility features, sequence features tend to influence TF binding in a TFCT-dependent manner (17). Among all possible k-mers, the most prominent sequence features with predictive power are GC content, CpG dinucleotide frequency and stretches of CpG islands. We reason that the incorporation of key features as such can provide quantitative descriptions of the local sequence/chromatin environment. For example, GC/CpG content can act as a proxy for nucleosome occupancy and DNA methylation, both of which are known regulators of TF binding (73–75).

As we observed high between-sample variation in correlations between GC/CpG features and ChIP-Seq signal (Supplementary Figure S4), GC/CpG feature usage varied across factors and cell types (Supplementary Figure S5C). Among the 98 trained TFCT experiments, we found that GC/CpG sequence features emerge, most frequently, as an interaction term with accessibility features (Supplementary Figure S5B). Given the dominant and relatively universal usage of accessibility features, GC/CpG sequence features seemed to play a modulating role boosting the predictive performance of chromatin accessibility for some TFCT conditions. For example, E2F family factors appeared to prefer CpG rich sequence environment almost universally, whereas the Forkhead family factors, such as FOXA1 and FOXA2, tend to make use of the GC/CpG features in a less straightforward manner.

Further, we saw a moderate correlation between GC/CpG sequence features and TSS proximity (Figure 3B), as well as a significant overlap between the usage of these two types of features across factors and cell types in our trained models (Supplementary Figure S5C). This is consistent with the fact that most promoter regions are GC-rich. In FOXA2 and ZBTB7A, for example, the effect of GC content on nucleosome occupancy and TF binding appears to depend on TSS proximity, with TF binding at TSS proximal sites featuring a more positively correlated relationship with GC/CpG content than TSS distal sites (positive coefficient for interaction features between GC/CpG and TSS proximity and negative coefficient for GC/CpG feature). This is reminiscent of what is observed in macrophage pioneer factor PU.1, where high GC content promotes the stable positioning of nucleosomes and leads to greater nucleosome occupancy at PU.1 sites, but very high GC content (e.g. CpG islands) disfavors nucleosome assembly at proximal sites and low GC content at distal sites (76).

Overall, our trained models show diverse usage of GC and CpG sequence features across factors and cell types (Supplementary Figure S5). Because GC content surrounding motif sites was shown to agree with core motif GC content in a TF family-specific manner, the sequence feature preference could be an indirect result of a homotypic environment for binding and cooperativity (17). Also, extremely GC-poor regions are thermodynamically disfavored for nucleosome positioning due to the stiff property of poly(dA:dT) sequences, so the wide-spread usage of GC content feature could underlie a structural basis for TF binding (77).

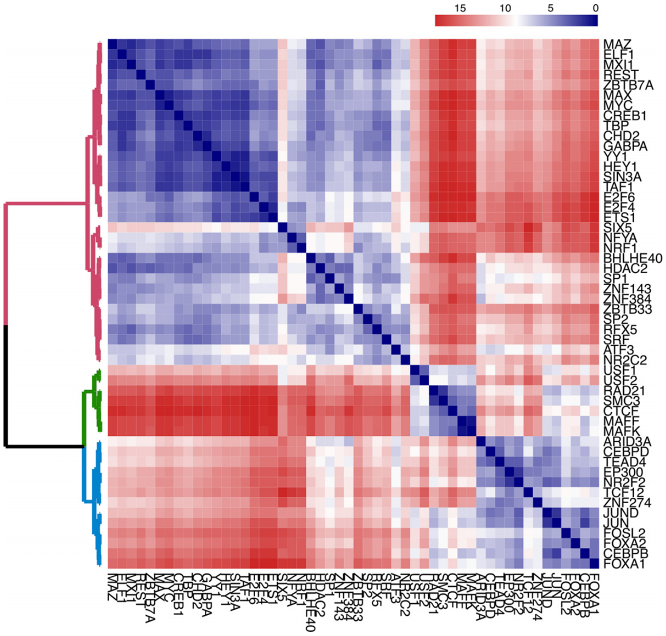

Figure 4. Heatmap clustering TFs based on the Euclidean distance between cross-TF prediction performances (AUPR). Red indicates large Euclidean distance and relatively poor cross-prediction performance between TFs; Blue indicates smaller Euclidean distance and good cross-prediction performance (where cross-prediction is the use of TF's MocapS model to predict another TF's binding). TFs are clustered together if they are more likely to share the same sparse logistic regression models for predicting TFBS. Data from multiple cell types, if available are averaged out for each TF.

Cross-sample TFBS predictions

To extend the usage of our TFCT-specific logistic regression classifiers to new factors and cell types, we next tested whether we can apply our trained models to the binding site predictions of TFCT conditions in the absence of ChIP-Seq data. We first evaluated the extent to which TFCT condition-specific models can cross-predict each other, where cross-prediction is quantified by measuring how well we can use model trained in TF *i* to predict binding for candidate motifs of TF *j*, and vice versa.

To reveal the factor-specific contribution to cross-sample TFBS prediction, we collapsed data for the same factor across cell types and generated a distance matrix between prediction performance across the 52 TFs (Figure 4). Among the TFs we have trained on, factors can be broken into three major clusters based on cross-sample predictive performances (these clusters contain TFs that are well predicted by similar sets of genomic, motif and accessibility features). The largest cluster (red) has accessibility and GC/CpG sequence feature as the most prominent predictors. Cluster 2 (green) features the cohesion complex factors, CTCF, RAD21 and SMC3 together with MAF and USF family factors that all tend to make significant use of motif scores. Cluster 3 (blue) is comprised predominately of enhancer-associated regulators, such as EP300 and TEAD4 whose binding sites are more likely to extend beyond TSS vicinity, and pioneer factors, such as AP1 (JUN/JUND) and Forkhead family factors. We found that, despite the sparsity constraint, more complex multi-featured models are preferred by factors in this cluster.

It is worth noting that part of the cross-sample prediction performance trade-off stems from differing signal-to-noise ratios between samples. There is a high level of cell type-specificity even for the same TFs due to varying levels of TF activities and multi-factored interactions. We found that models trained in cell types where the TF show more true binding sites (ChIP-Seq peaks) tend to be more robust when applied to the same factor in another cell type. This is because TFs with fewer binding sites across the genome tend to produce more biased models. SIX5, for instance, has a very small set of binding sites in K562. This small true class of binding sites, although weight balanced with the false class, can only give us partial information about the factor' s binding preference, so it is often difficult to infer generalizable models or perform fair model evaluation on such datasets. To circumvent this bias and down-weigh such outlier samples, we used iteratively weighted least squares regression to derive a feature mapping vector (Figure 5). As the system is overdetermined ($n \gg p$), we reason that down-weighing noisy outliers (samples with large residuals) can produce more robust sample mapping. Additionally, because not all supervised MocapS models show significantly improved performance over unsupervised Mocap, we constrained the use of this weighted mapping (MocapX) to models that showed significant improvement over unsupervised MocapG to control for uninformative mappings.

The weight distribution in MocapX derived feature vector suggests cross-sample mapping is predominately driven by GC/CpG sequence feature similarities (Figure 5). Among the 98 leave-one-out mapping experiments, we were able to map 35 samples to another TFCT trained model (the rest chose the generic method MocapG). Fourteen of the 35 mappings were between cell types of the same TF, the rest mapped cross-TF. Among the cross-TF mappings, the most prominent one is between CTCF and RAD21 (Supplementary Table S1). This is consistent with the overlapping roles of these two factors in defining chromosomal domains and in interactions with other cohesion complex components, such as SMC3 (78–80). It was shown that CTCF and RAD21 bind to a subset of their accessible motifs and a small number of these binding sites are further influenced by cell type-specific DNA-methylation (72,81). MocapG was found to be rather inaccurate when predicting the binding of these factors. Incorporating motif scores and footprint profiles significantly improves prediction accuracy in these cases.

Other examples of cross-TF mappings are ETS1/YY1, CEBPB/JUN, REST/SIN3A and MAX/MYC (Supplementary Table S1). These factors were previously recognized as binding partners, so their binding sites likely co-cluster and thus share similar sequence environments and model preferences (82–85). As a point of comparison, Figure 5 also demonstrates the peril of wrongly assigning models to samples: randomly selecting MocapS models for cross-sample predictions can negatively impact performance. This underlies the importance of discriminating between condition-specific models. Further, we found that inferring binding for TFs in conditions/cell types where they are inactive, partially active or have altered binding activities tend to confound cross-sample TFBS predictions (Supplementary Table S1 and Figure S7B), potentially highlighting the need to integrate alternate data-types, such as TF expression or approaches that can explicitly account for TF activity in different cell types (http://dx.doi.org/10.1101/051847) (86). Taken together, we show that MocapX, although limited by the number and diversity of TFCT-specific models we have trained, presents a novel framework to generalize trained sparse logistic regression models to an increasing number of TFCTs with improved accuracy.

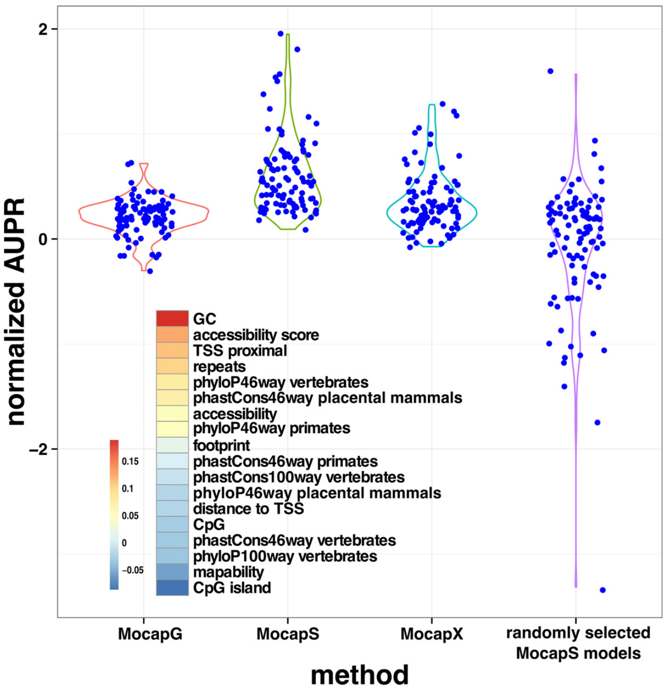

Figure 5. Cross-sample binding site prediction. Violin plot showing the hold-one-out performance for MocapX in comparison to MocapS (with MocapS models trained in the TFCT), MocapG (with local chromatin accessibility feature only) and randomly selected MocapS model (with random mappings between leave-out TFCT and MocapS model ensemble) performance. AUPR scores are normalized (centered at zero) across the four methods in each TFCT condition Inset: Heatmap showing weighted feature vectors that is used to compute distances between new TFCTs and TFCTs for which MocapS models have been trained. If no fit model exists in the trained model pool (no model is predicted to outperform unsupervised MocapG), MocapX will use MocapG for TFBS prediction.

Performance comparisons

We compared the performance of Mocap with two other motif and chromatin accessibility-based TFBS prediction methods CENTIPEDE and PIQ. As motif datasets show significant sample class imbalance, with true binding sites taking up only a fraction of the total candidate sites in our predictions (positive/negative ratio < 0.01), AUROC scores are likely biased toward assessing correct classification of the majority class (which in our case is the non-binding sites that we are less interested in). Thus, we computed the AUPR for each leave-out TFCT to compare the motif-level predictive accuracy among methods (trade-off between precision and recall) (Supplementary Figure 6A and Supplementary Table S1). Additionally, a peak-level performance comparison using the areas under the precision-peak recall curves was also done to access how well each method predicts ChIP-Seq peaks (Supplementary Figure S7D). We also evaluated sensitivity at 1% FPR to evaluate binding site prediction at a low false positive rate cutoff (trade-off between sensitivity and specificity) (Supplementary Figure 6A and Supplementary Table S1).

When compared to CENTIPEDE and PIQ, both our sparse logistic regression training-based MocapS and our extended MocapX methods showed a good balance between sensitivity and specificity across factors and cell types (Supplementary Figure S7A and B). MocapG, which uses simple accessibility cut count ranking, although lacking in the completeness of coverage, remained a robust prediction method across factors and cell types (Figure 6A). But, MocapG did tend to fall short in TFs such as CEBPB, E2F4, RAD21 and MAFF where motif matching scores, sequence features or TSS proximity played major roles (Supplementary Table S1). These cases clearly demonstrate the need to build TFCT-tailored models that use a diverse collection of features.

CENTIPEDE and PIQ both model footprints in great detail. PIQ uses refined technique to model motif and DNase I footprint. CENTIPEDE integrates motif scores, TSS proximity and evolutionary conservation with DNase I footprint in a hierarchical model. Overall, both methods rely heavily on the modeling of footprint profiles. So, when we apply these methods to TFBS prediction with a loosened constraint on motif matching score to improve sensitivity, they often fell short in ranking precision (poor AUPR scores). This points us to the limitation of methods that rely solely or heavily on footprint profiling. First, it is difficult to profile TF footprints even with the aid of factor-specific motifs, especially in a manner that balances specificity with completeness of prediction (due, as we discuss above, to a host of other influences like chromatin context and cell-type specific co-factors). Footprint detection methods, although descriptively useful, tend to be highly biased and condition-specific, thus lack the ability to generalize across factors and cell types and are often unfit for application on a global scale, to facilitate efforts, such as genome-wide CRM detection or gene regulatory network inference. Secondly, part of the variation in footprint profile might be intrinsic to differential TF binding activities. Although both deriving better motif models and applying condition/protocol-specific bias correction using naked DNA could help resolve footprint patterns and improve overall predictive performance, the condition specificity of TF footprints both within and across TFCTs will likely remain (27,30). This again signifies the need for building predictive models that specifically address the differences between TFCTs.

We also note that the poor performance of PIQ and CENTIPEDE are partly due to the fact that both methods are unsupervised learning methods and relied, to an extent, on pre-selecting motif sites, e.g. with high motif matching scores and/or footprint scores, to constrain their prediction problems. Although useful, these pre-selection steps often lead to sampling bias and thus limit their ability to generalize and achieve balanced precision-recall across TFCTs.

Adapting Mocap to predict TFBS using ATAC-Seq

ATAC-Seq has emerged in recent years as an efficient method to assay chromatin accessibility and typically requires much smaller cell sample than DNase-Seq to achieve comparable sequencing depth. This has led to the wide application of ATAC-Seq to clinically relevant rare cell types and conditions that were previously inaccessible to DNase-Seq due to limited cell numbers. The higher sensitivity and stable nature of Tn5 also results in more ambiguous/arbitrary footprint profiles in comparison to DNase I (Supplementary Figure S6, Appendix).

To examine performance across assays, we applied MocapG and MocapX to ATAC-Seq data in human Gm12878 cells and compared the performance of Mocap with that of CENTIPEDE (Figure 6B and Supplementary Table S2). MocapG, because of its simplicity, performs comparably across assays, whereas the differences in footprint profiles limit the specificity of MocapX when applied to ATAC-Seq. This loss in predictive performance is more pronounced for CENTIPEDE, which was designed specifically for DNase footprints. Despite the differences in footprint profiles, MocapX was shown to improve the performance of TFBS predictions for TFs, such as CTCF, RAD21 and USF1, where MocapG tends to perform poorly. This represents, to our knowledge, the first TFCT-specific effort to use ATAC-Seq for TFBS prediction.

DISCUSSION

In this work, we developed a DNA binding motif and chromatin accessibility-based method to predict cell type-specific TF binding. In designing a generalizable TFBS prediction method, we followed several key design principles: (i) we address the differences in binding behaviors between TF-cell type conditions, (ii) we predict TFBS in a manner that balances method precision with recall, and (iii) we employ approaches that improve method scalability. To distinguish between TFs and cell types, we incorporated in our analyses a range of motif-associated genomic features, including motif matching scores, chromatin accessibility, TF footprints, GC/CpG content, TSS proximity and evolutionary conservation. We assessed each feature's contribution to TFBS prediction, and applied model selection to the training of an ensemble of TFCT-specific classifiers integrating these genomic features. We show that incorporation of sequence features, such as GC/CpG content surrounding TF binding motifs, significantly improves predictive performance and helps identify TFCT conditions sharing similar predictive models. To improve the sensitivity of our predictive method (recall), we start with a more complete coverage of binding sites by loosening motif matching threshold; and

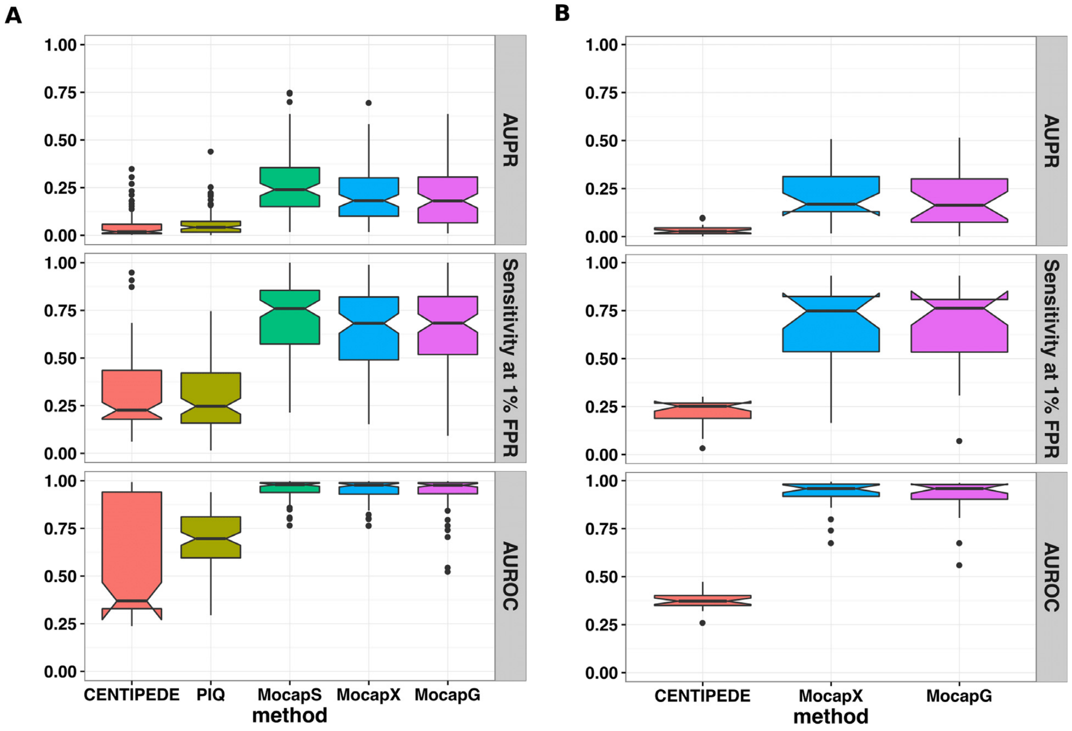

Figure 6. Method comparison between Mocap, CENTIPEDE and PIQ (98 TFCT samples in hold-out chromosome 15). (**A**) Boxplot showing overall performance of CENTIPEDE, PIQ, MocapS, MocapX, and MocapG method in predicting TFBS ($n = 98$). (**B**) Boxplot showing performance of Mocap and CENTIPEDE applied to ATAC-Seq data in Gm12878 ($n = 23$). Performance metrics used are AUPR (top panel), Sensitivity at 1% FPR (middle panel) and AUROC (bottom panel).

to provide more informative ranking of our predicted binding sites (precision), we set the objective function in sparse logistic regression training to optimize AUPR.

Our TFCT-specific models perform favorably in binding site predictions for a range of TFs and in a multitude of cell types in comparison to several previous methods. We show that this specificity and performance can be extended to other TFs that lack ChIP-seq via cross-sample TFBS prediction (MocapX) using condition-specific models trained in other TFCTs. Lastly, to promote scalability, we focused on designing a method that is computationally efficient and relies on only a single type of genomic assay, either DNase-Seq or ATAC-Seq to capture the chromatin dynamics around binding sites. Although a wide variety of functional genomic assays, such as histone modification ChIP-Seq, MNase-Seq and bisulfite sequencing could all contribute to TFBS prediction (87,88), we chose to limit our required inputs to chromatin accessibility experiments, because they remain, to our knowledge, the most generalizable genomic assay in predicting cell type-specific binding sites across factors and conditions. In particular, given that the recent advances in genomic technologies continue to make chromatin accessibility interrogation more widely applicable, chromatin accessibility-based TFBS prediction methods will find more application in global-scale gene regulation analysis, such as cell type-specific CRM identification, multi-lineage regulatory landscape comparison and be used as structural priors in global gene regulatory network inference (http://dx.doi.org/10.1101/051847). Our method combined cell type-specific regulatory information in chromatin accessibility data with a range of genomic evidences in *cis* to drive more accurate binding site prediction. This lessened reliance on multiple expensive data-types in *trans* allows our method to be more readily scaled to a larger number of TFCT conditions.

In addition to building useful classifiers for global-scale TFBS mapping, our study aimed to identify factors that distinguish motif sites that are bound from the vast majority of unbound motif sites and provide mechanistic insights into TF binding dynamics and diversity. While chromatin accessibility assays allowed careful descriptions of the chromatin environment around TF binding sites, the sequence environment that fosters TF binding specificity and cooperativity, in contrast, is arguably harder to unravel and has thus remained by and large a conjecture that needs to be disentangled and tested in a more systematic fashion (89). Our trained sparse logistic regression models encapsulate some of the diverse combinations of sequence features that lead to TF-specific binding, such as GC/CpG content plays in binding site prediction. Additional features that describe motif-proximal sites, such as k-mer features and predictors of DNA shape , need to be investigated in more diverse biological contexts (70,90,91). For CpG features in particular, cell type-specific methylation assays, such as BS-sequencing could bring more functional relevance to its predictive modeling (48). Our study also provides a framework for examining co-clustered binding sites in a relatively unbiased manner; a key avenue of investigation, as binding site clustering is required for activation at many well studied loci (4,49,92,93).

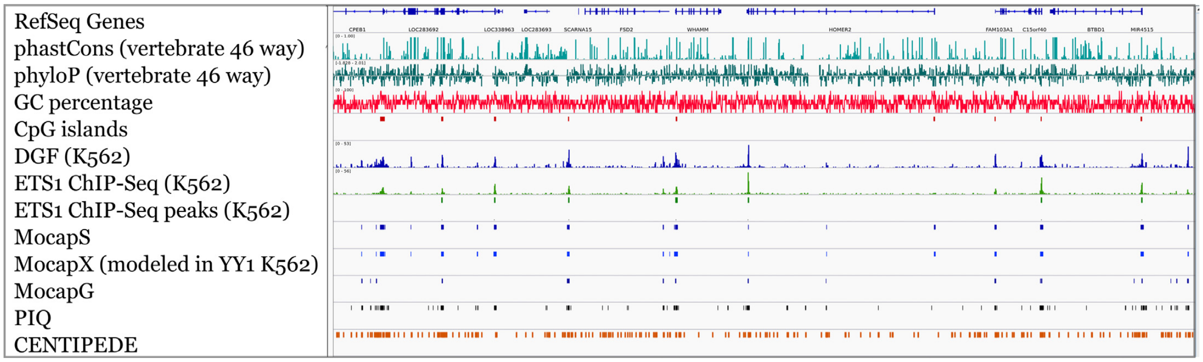

Figure 7. Genome browser view of predictions made by different methods. Tracks highlight region 85081291–85557900 on chromosome 15 for binding site predictions of ETS1 in K562. We standardized MocapG, MocapS and MocapX (modeled with YY1 in K562) prediction scores into z scores and used a cutoff of $z > 3$. Cutoff for PIQ and CENTIPEDE are 700 and 0.99 as suggested.

Our definition of true binding sites might have included motifs due to their proximity to an actual binding site, but are not actually physically bound. Several reasons preclude an accurate definition of true binding site in our current data framework. First, the use of clustered or fused binding sites in mammalian species is prevalent. This means that there are often times multiple motifs, slight variations of known motifs, permuted arrangements of motifs clustered under one ChIP-Seq peak, so it often requires the support of new data evidence, such as higher resolution ChIP-exo or ChIP-nexus and in-depth analysis of raw sequencing reads to deconvolve the exact bound sites from ChIP-Seq enriched regions (94,95). Second, as our analyses (Supplementary Figure S2) suggest it is common to observe ChIP-Seq peaks that do not contain any high-quality motifs due to indirect binding, non-specific binding or ChIP-Seq-associated technical artifacts (19,96), so for most, if not all, TFCTs, there is not a one-to-one correspondence between known motifs and ChIP-Seq peaks. Lastly, methods that attempt to preselect motif site with motif matching scores or footprint patterns as a way to resolve closely spaced binding sites often risk missing *bona fide* binding sites due to the rather frequent presence of fused or permuted motif sites that deviate from canonical motif patterns or low-residence time binding events that lack conspicuous footprint profiles. As our study is not aimed at resolving closely spaced binding sites (under ChIP-Seq peaks), the motif regions identified in our study can be further refined by *de novo* motif discovery tools such as SeqGL, GEM and ChIPMunk (94,95,97–99). Refined motif models could in turn facilitate more accurate depiction of TF footprints and improve binding site predictions (Figure 7).

Evolutionary conservation represents yet another type of data that has long been associated with functional TF binding. We tested a range of conservation features in this work, including both measures of cross-species divergence and population-level polymorphism (e.g. SNP density) from ENCODE and 1000 genomes respectively (100). We found that SNP density at motif sites appear to have a diminished effect on binding site prediction (unpublished data), in comparison to cross-species divergence (101,102). Among the cross-species conservation scores, mammalian conservation (both phyloP scores and phastCons scores) seemed the more relevant evolutionary distance than vertebrates or primates conservation (as evidenced by their selection in models for multiple TFs during MocapS training). This perhaps suggests a shift in balance between the regulatory conservation within mammalian species and the site divergence experienced among primates pointing to a precarious relationship between binding site conservation and divergence during evolution (54,103).

As technical advances in genomics and statistics enable the accurate and large-scale mapping of a large number of TFBS, predictive methods that combine sequence features with chromatin accessibility modeling represent a promising direction for resolving the myriad binding sites across a diverse array of TFs and cell types. Large-scale mapping of TFBS, when connected with gene expression data, will in turn promote a better and more systematic understanding of mammalian gene regulation and enable large scale network inference via the generation of detailed structural priors (53,86,104).

AVAILABILITY

Our core methods are freely available as an R package. Detailed implementation and example data can be found at https://github.com/xc406/Mocap.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Christoph Hafemeister, Christian Muller, Aviv Madar, Emily Miraldi and other members of the Bonneau lab for helpful discussions. We would also like to thank Dennis Shasha and Jonathan Flowers for useful suggestions and Shenglong Wang and other members of the NYU HPC team for excellent technical support.

FUNDING

NIH; NSF; Simons Foundation. Funding for open access charge: Simons Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Mitchell,P.J. and Tjian,R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **245**, 371–378.
2. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
3. van Steensel,B. (2005) Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet.*, **37**, S18–S24.
4. Junion,G., Spivakov,M., Girardot,C., Braun,M., Gustafson,E.H., Birney,E. and Furlong,E.E. (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, **148**, 473–486.
5. Davidson,E. (2010) *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*, Elsevier Science, NY.
6. Blais,A. and Dynlacht,B.D. (2005) Constructing transcriptional regulatory networks. *Genes Dev.*, **19**, 1499–1511.
7. Tjian,R. (1978) The binding site on SV40 DNA for a T antigen-related protein. *Cell*, **13**, 165–179.
8. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110.
9. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.
10. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
11. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2014) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
12. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Soboleva,A.V., Kasianov,A.S., Ashoor,H., Ba-alawi,W., Bajic,V.B., Medvedeva,Y.A., Kolpakov,F.A. *et al.* (2016) HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, **44**, D116–D125.
13. Matys,V., Fricke,E., Geffers,R., Goessling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
14. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
15. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
16. Zhang,C., Xuan,Z., Otto,S., Hover,J.R., McCorkle,S.R., Mandel,G. and Zhang,M.Q. (2006) A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res.*, **34**, 2238–2246.
17. Dror,I., Golan,T., Levy,C., Rohs,R. and Mandel-Gutfreund,Y. (2015) A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.*, **25**, 1268–1280.
18. Levo,M., Zalckvar,E., Sharon,E., Machado,A.C.D., Kalma,Y., Lotam-Pompan,M., Weinberger,A., Yakhini,Z., Rohs,R. and Segal,E. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.*, **25**, 1018–1029.
19. Gilfillan,G.D., Hughes,T., Sheng,Y., Hjorthaug,H.S., Straub,T., Gervin,K., Harris,J.R., Undlien,D.E. and Lyle,R. (2012) Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics*, **13**, 645.
20. Park,P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
21. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.
22. Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protoc.*, **2010**, doi:10.1101/pdb.prot5384.
23. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
24. Giresi,P.G., Kim,J., McDaniell,R.M., Iyer,V.R. and Lieb,J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.
25. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
26. Sherwood,R.I., Hashimoto,T., O'Donnell,C.W., Lewis,S., Barkal,A.A., van Hoff,J.P., Karun,V., Jaakkola,T. and Gifford,D.K. (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.*, **32**, 171–178.
27. Kähärä,J. and Lähdesmäki,H. (2015) BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics*, **31**, 2852–2859.
28. Piper,J., Elze,M.C., Cauchy,P., Cockerill,P.N., Bonifer,C. and Ott,S. (2013) Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.*, **41**, e201.
29. Gusmao,E.G., Dieterich,C., Zenke,M. and Costa,I.G. (2014) Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, **30**, 3143–3151.
30. Yardımcı,G.G., Frank,C.L., Crawford,G.E. and Ohler,U. (2014) Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.*, **42**, 11865–11878.
31. Slattery,M., Zhou,T., Yang,L., Machado,A.C.D., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
32. Zinzen,R.P., Girardot,C., Gagneur,J., Braun,M. and Furlong,E.E. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.
33. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.
34. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
35. Dunham,I., Birney,E., Herrero,J., Wilder,S.P., Keefe,D., Beal,K., Flicek,P., Johnson,N., Sobraland,D., Kundaje,A. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
36. Yu,B., Doraiswamy,H., Chen,X., Miraldi,E., Arrieta-Ortiz,M.L., Hafemeister,C., Madar,A., Bonneau,R. and Silva,C.T. (2014) Genotet: An interactive web-based visual exploration framework to support validation of gene regulatory networks. *Visual. Comput. Graph. IEEE Trans.*, **20**, 1903–1912.
37. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
38. Tarailo-Graovac,M. and Chen,N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, 4–10.
39. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
40. Saxonov,S., Berg,P. and Brutlag,D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 1412–1417.

41. Fan,R.-E., Chang,K.-W., Hsieh,C.-J., Wang,X.-R. and Lin,C.-J. (2008) LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.
42. Andersen,R. (2008) *Modern Methods for Robust Regression*, Sage, p. 152.
43. Huber,P.J. (1981) Wiley series in probability and mathematics statistics. *Robust Stat.*, 309–312.
44. Wong,D., Teixeira,A., Oikonomopoulos,S., Humburg,P., Lone,I.N., Saliba,D., Siggers,T., Bulyk,M., Angelov,D., Dimitrov,S. *et al.* (2011) Extensive characterization of NF-κB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biol.*, **12**, 1–19.
45. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
46. Ramos,A.I. and Barolo,S. (2013) Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.*, **368**, 20130018.
47. Crocker,J., Abe,N., Rinaldi,L., McGregor,A.P., Frankel,N., Wang,S., Alsawadi,A., Valenti,P., Plaza,S., Payre,F. *et al.* (2015) Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, **160**, 191–203.
48. Xu,T., Li,B., Zhao,M., Szulwach,K.E., Street,R.C., Lin,L., Yao,B., Zhang,F., Jin,P., Wu,H. *et al.* (2015) Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res.*, **43**, 2757–2766.
49. Gotea,V., Visel,A., Westlund,J.M., Nobrega,M.A., Pennacchio,L.A. and Ovcharenko,I. (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.*, **20**, 565–577.
50. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-André,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
51. Whyte,W.A., Orlando,D.A., Hnisz,D., Abraham,B.J., Lin,C.Y., Kagey,M.H., Rahl,P.B., Lee,T.I. and Young,R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
52. Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.
53. Ciofani,M., Madar,A., Galan,C., Sellars,M., Mace,K., Pauli,F., Agarwal,A., Huang,W., Parkurst,C.N., Muratet,M. *et al.* (2012) A validated regulatory network for Th17 cell specification. *Cell*, **151**, 289–303.
54. Hardison,R.C. and Taylor,J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.*, **13**, 469–483.
55. Benveniste,D., Sonntag,H.-J., Sanguinetti,G. and Sproul,D. (2014) Transcription factor binding predicts histone modifications in human cell lines. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 13367–13372.
56. Siggers,T., Duyzend,M.H., Reddy,J., Khan,S. and Bulyk,M.L. (2011) Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.*, **7**, 555.
57. Rashid,N.U., Giresi,P.G., Ibrahim,J.G., Sun,W. and Lieb,J.D. (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.*, **12**, R67.
58. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
59. Stergachis,A.B., Neph,S., Reynolds,A., Humbert,R., Miller,B., Paige,S.L., Vernot,B., Cheng,J.B., Thurman,R.E., Sandstrom,R. *et al.* (2013) Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, **154**, 888–903.
60. Vierstra,J. and Stamatoyannopoulos,J.A. (2016) Genomic footprinting. *Nat. Methods*, **13**, 213–221.
61. He,H.H., Meyer,C.A., Chen,M.-W., Zang,C., Liu,Y., Rao,P.K., Fei,T., Xu,H., Long,H., Liu,X.S. *et al.* (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods*, **11**, 73–78.
62. Koohy,H., Down,T.A. and Hubbard,T.J. (2013) Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS One*, **8**, e69853.
63. Madrigal,P. (2015) On accounting for sequence-specific bias in genome-wide chromatin accessibility experiments: recent advances and contradictions. *Front. Bioeng. Biotechnol.*, **3**, 144.
64. Sung,M.-H., Guertin,M.J., Baek,S. and Hager,G.L. (2014) DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell*, **56**, 275–285.
65. Ang,S.-L., Wierda,A., Wong,D., Stevens,K.A., Cascio,S., Rossant,J. and Zaret,K.S. (1993) The formation and maintenance of the definitive endoderm lineage in the mouse: involvement of HNF3/forkhead proteins. *Development*, **119**, 1301–1315.
66. Iwafuchi-Doi,M. and Zaret,K.S. (2014) Pioneer transcription factors in cell reprogramming. *Genes Dev.*, **28**, 2679–2692.
67. Zaret,K.S. and Carroll,J.S. (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, **25**, 2227–2241.
68. Iwafuchi-Doi,M., Donahue,G., Kakumanu,A., Watts,J.A., Mahony,S., Pugh,B.F., Lee,D., Kaestner,K.H. and Zaret,K.S. (2016) The pioneer transcription factor FoxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation. *Mol. Cell*, **62**, 79–91.
69. Mahony,S. and Pugh,B.F. (2015) Protein–DNA binding in high-resolution. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 269–283.
70. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.
71. Valouev,A., Johnson,S.M., Boyd,S.D., Smith,C.L., Fire,A.Z. and Sidow,A. (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516–520.
72. Maurano,M.T., Wang,H., John,S., Shafer,A., Canfield,T., Lee,K. and Stamatoyannopoulos,J.A. (2015) Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.*, **12**, 1184–1195.
73. Medvedeva,Y.A., Khamis,A.M., Kulakovskiy,I.V., Ba-Alawi,W., Bhuyan,M. S.I., Kawaji,H., Lassmann,T., Harbers,M., Forrest,A.R., Bajic,V.B. *et al.* (2014) Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, **15**, 119.
74. Tillo,D. and Hughes,T.R. (2009) G+ C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, **10**, 1.
75. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.
76. Barozzi,I., Simonatto,M., Bonifacio,S., Yang,L., Rohs,R., Ghisletti,S. and Natoli,G. (2014) Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol. Cell*, **54**, 844–857.
77. Iyer,V. and Struhl,K. (1995) Poly (dA: dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.*, **14**, 2570.
78. Phillips-Cremins,J.E., Sauria,M.E., Sanyal,A., Gerasimova,T.I., Lajoie,B.R., Bell,J.S., Ong,C.-T., Hookway,T.A., Guo,C., Sun,Y. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.
79. Seitan,V.C., Faure,A.J., Zhan,Y., McCord,R.P., Lajoie,B.R., Ing-Simmons,E., Lenhard,B., Giorgetti,L., Heard,E., Fisher,A.G. *et al.* (2013) Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res.*, **23**, 2066–2077.
80. Sofueva,S., Yaffe,E., Chan,W.-C., Georgopoulou,D., Rudan,M.V., Mira-Bontenbal,H., Pollard,S.M., Schroth,G.P., Tanay,A. and Hadjur,S. (2013) Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.*, **32**, 3119–3129.
81. Wang,H., Maurano,M.T., Qu,H., Varley,K.E., Gertz,J., Pauli,F., Lee,K., Canfield,T., Weaver,M., Sandstrom,R. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, **22**, 1680–1688.
82. Gaston,K. and Fried,M. (1995) CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and Surf-2 genes. *Nucleic Acids Res.*, **23**, 901–909.
83. Hong,S., Skaist,A.M., Wheelan,S.J. and Friedman,A.D. (2011) AP-1 protein induction during monopoiesis favors C/EBP: AP-1 heterodimers over C/EBP homodimerization and stimulates FosB transcription. *J. Leukocyte Biol.*, **90**, 643–651.

84. Huang,Y., Myers,S.J. and Dingledine,R. (1999) Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nat. Neurosci.*, **2**, 867–872.
85. Nair,S.K. and Burley,S.K. (2003) X-ray structures of Myc-Max and Mad-Max recognizing DNA: molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, **112**, 193–205.
86. Arrieta-Ortiz,M.L., Hafemeister,C., Bate,A.R., Chu,T., Greenfield,A., Shuster,B., Barry,S.N., Gallitto,M., Liu,B., Kacmarczyk,T. *et al.* (2015) An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Mol. Syst. Biol.*, **11**, 839.
87. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
88. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
89. Dror,I., Rohs,R. and Mandel-Gutfreund,Y. (2016) How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays*, **38**, 605–612.
90. Abe,N., Dror,I., Yang,L., Slattery,M., Zhou,T., Bussemaker,H.J., Rohs,R. and Mann,R.S. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell*, **161**, 307–318.
91. Chiu,T.-P., Comoglio,F., Zhou,T., Yang,L., Paro,R. and Rohs,R. (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, **32**, 1211–1213.
92. Tsankov,A.M., Gu,H., Akopian,V., Ziller,M.J., Donaghey,J., Amit,I., Gnirke,A. and Meissner,A. (2015) Transcription factor binding dynamics during human ES cell differentiation. *Nature*, **518**, 344–349.
93. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
94. Guo,Y., Papachristoudis,G., Altshuler,R.C., Gerber,G.K., Jaakkola,T.S., Gifford,D.K. and Mahony,S. (2010) Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, **26**, 3028–3034.
95. Guo,Y., Mahony,S. and Gifford,D.K. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.
96. Teytelman,L., Thurtle,D.M., Rine,J. and van Oudenaarden,A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 18602–18607.
97. Setty,M. and Leslie,C.S. (2015) SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. *PLoS Comput. Biol.*, **11**, e1004271.
98. Kulakovskiy,I.V., Boeva,V., Favorov,A.V. and Makeev,V. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
99. Kulakovskiy,I., Levitsky,V., Oshchepkov,D., Bryzgalov,L., Vorontsov,I. and Makeev,V. (2013) From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinformatics Computat. Biol.*, **11**, 1340004.
100. McVean,G.A., Abecasis,G.R., Bentley,D.R., Chakravarti,A., Clark,A.G., Donnelly,P., Eichler,E.E., Flicek,P., Lunter,G., Marchini,J.L. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
101. Lindblad-Toh,K., Garber,M., Zuk,O., Lin,M.F., Parker,B.J., Washietl,S., Kheradpour,P., Ernst,J., Jordan,G., Mauceli,E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.
102. Ward,L.D. and Kellis,M. (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, **337**, 1675–1678.
103. Dowell,R.D. *et al.* (2010) Transcription factor binding variation in the evolution of gene regulation. *Trends Genet.: TIG*, **26**, 468.
104. Greenfield,A., Hafemeister,C. and Bonneau,R. (2013) Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, **29**, 1060–1067.