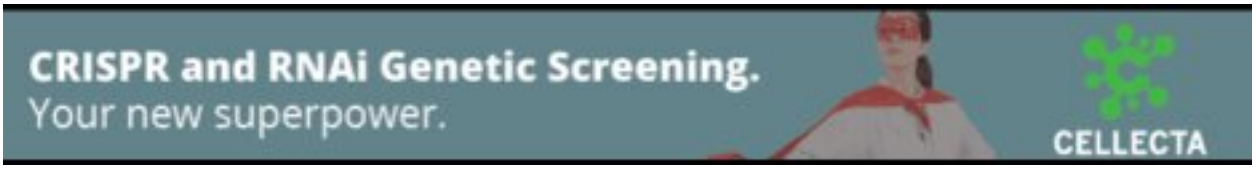
Profiling the quantitative occupancy of myriad transcription factors across conditions by modeling chromatin accessibility data

Kaixuan Luo, Jianling Zhong, Alexias Safi, et al.

Genome Res. 2022 32: 1183-1198 originally published online May 24, 2022
Access the most recent version at doi:10.1101/gr.272203.120



Method

Profiling the quantitative occupancy of myriad transcription factors across conditions by modeling chromatin accessibility data

Kaixuan Luo,[1,2,3,4] Jianling Zhong,[1,2,3] Alexias Safi,[2,5] Linda K. Hong,[2,5] Alok K. Tewari,[6] Lingyun Song,[2,5] Timothy E. Reddy,[1,2,7,8,9] Li Ma,[1,10] Gregory E. Crawford,[1,2,5] and Alexander J. Hartemink[1,2,3,11]

[1]Computational Biology & Bioinformatics Graduate Program, Duke University, Durham, North Carolina 27708, USA; [2]Center for Genomic and Computational Biology, Duke University, Durham, North Carolina 27708, USA; [3]Department of Computer Science, Duke University, Durham, North Carolina 27708, USA; [4]Department of Human Genetics, The University of Chicago, Chicago, Illinois 60637, USA; [5]Department of Pediatrics, Duke University Medical Center, Durham, North Carolina 27710, USA; [6]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA; [7]Department of Biostatistics and Bioinformatics, [8]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina 27710, USA; [9]Department of Biomedical Engineering, [10]Department of Statistical Science, [11]Department of Biology, Duke University, Durham, North Carolina 27708, USA

Over a thousand different transcription factors (TFs) bind with varying occupancy across the human genome. Chromatin immunoprecipitation (ChIP) can assay occupancy genome-wide, but only one TF at a time, limiting our ability to comprehensively observe the TF occupancy landscape, let alone quantify how it changes across conditions. We developed TF occupancy profiler (TOP), a Bayesian hierarchical regression framework, to profile genome-wide quantitative occupancy of numerous TFs using data from a single chromatin accessibility experiment (DNase- or ATAC-seq). TOP is supervised, and its hierarchical structure allows it to predict the occupancy of any sequence-specific TF, even those never assayed with ChIP. We used TOP to profile the quantitative occupancy of hundreds of sequence-specific TFs at sites throughout the genome and examined how their occupancies changed in multiple contexts: in approximately 200 human cell types, through 12 h of exposure to different hormones, and across the genetic backgrounds of 70 individuals. TOP enables cost-effective exploration of quantitative changes in the landscape of TF binding.

[Supplemental material is available for this article.]

Genes are expressed differently in different types of cells and under different conditions. This response of a cell's gene expression to its internal and external context is enacted in large part through the tuned occupancy of transcription factors (TFs) across the genome. To understand how TFs regulate gene expression, it is critical to determine how likely they are to be present at sites in the genome over time, and how that likelihood changes across varying genetic backgrounds, different cell types, and dynamic environmental conditions. We can measure the quantitative occupancy of one TF throughout the genome using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), provided that a selective antibody exists for the TF. Although The ENCODE Project Consortium has generated such data for hundreds of human TFs, the data are typically from only a small number of cell types because of a major limitation of ChIP-seq: A separate experiment is required for each TF in each cell type under each condition. Profiling the time-varying genome-wide occupancy of a large set of TFs across a broad range of cell types and conditions is currently impractical because it would require thousands of antibodies and millions of separate ChIP experiments.

An alternative strategy for profiling genome-wide TF occupancy is to exploit chromatin accessibility data like DNase-seq or ATAC-seq, which many groups and consortia have generated for a large number of cell types and experimental conditions (Neph et al. 2012; Thurman et al. 2012; Buenrostro et al. 2013; Klemm et al. 2019). The primary advantage of this strategy is that a single DNase- or ATAC-seq experiment can be used to profile the occupancy of many different TFs at once, and a number of methods using this strategy have been proposed in recent years (Pique-Regi et al. 2011; He et al. 2012, 2014; Luo and Hartemink 2013; Sherwood et al. 2014; Zhong et al. 2014; Kähärä and Lähdesmäki 2015; Raj et al. 2015; Zeng et al. 2016; Keilwagen et al. 2019; Quang and Xie 2019; Li et al. 2019a; Schreiber et al. 2020a).

Although multiple methods have been developed to predict TF binding (for an overview of the modeling frameworks used by a number of these methods, see Supplemental Table S1), many require data types beyond DNase- or ATAC-seq (Keilwagen et al. 2019; Quang and Xie 2019; Li et al. 2019a; Schreiber et al. 2020a), making them less efficient at profiling TF occupancy across multiple cell types, conditions, or genetic backgrounds than

Corresponding author: amink@cs.duke.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.272203.120.

methods requiring only one data type. Furthermore, most existing methods model TF occupancy in a binary fashion—each TF is simply considered present or absent at each location in the genome—not effectively using the wealth of quantitative information available in the data (Hoffman et al. 2012). Although peak-calling is widely applied to genomic sequencing data, and binary peak calls are simple to use when training classification models, we know that at different genomic locations, TFs show different levels of occupancy (likelihood of being bound at that location across the cells in a population) in accordance with prevailing energetic and thermodynamic conditions, including competition with other TFs and nucleosomes (Narlikar et al. 2007; Gordân and Hartemink 2008; Wasson and Hartemink 2009; Li et al. 2011; Lickwar et al. 2012; Neph et al. 2012). Also, growing evidence suggests quantitative levels of TF occupancy can play a significant role in regulating gene expression (Gertz et al. 2009; Segal and Widom 2009; McDaniell et al. 2010; Tewari et al. 2012). Therefore, it is important that statistical models be developed with a quantitative perspective, allowing us to monitor subtle changes in TF occupancy over time across different genetic backgrounds, cell types, and conditions.

Here, we describe a novel method called TF occupancy profiler (TOP) that integrates chromatin accessibility data from DNase- or ATAC-seq with information about TF binding specificity (in our case, TF motifs) to predict the quantitative occupancy of multiple TF motifs genome-wide. TOP follows the site-centric strategy used by CENTIPEDE (Pique-Regi et al. 2011), and later by MILLIPEDE (Luo and Hartemink 2013), BinDNase (Kähärä and Lähdesmäki 2015), and msCentipede (Raj et al. 2015). This strategy models the DNase- or ATAC-seq profiles (a.k.a. footprints) around motif matches and is therefore limited to predictions at genomic locations with motif matches.

In contrast to earlier methods like CENTIPEDE (Pique-Regi et al. 2011), PIQ (Sherwood et al. 2014), and msCentipede (Raj et al. 2015), TOP is supervised, meaning we can use available ChIP-seq data to train it to high accuracy. Importantly, and in contrast to earlier methods like MILLIPEDE (Luo and Hartemink 2013) and BinDNase (Kähärä and Lähdesmäki 2015), TOP uses a Bayesian hierarchical regression framework, obtaining both TF-specific and TF-generic model parameters by borrowing information across the full spectrum of training TFs and cell types. The hierarchical nature of TOP is significant because it enables us to predict the occupancy of TFs for which we lack training data, including ones that have never before been profiled with ChIP.

We aim to evaluate TOP's performance and explore the potential applications of TOP's quantitative predictions in three different contexts. First, we predict the genome-wide quantitative occupancy of hundreds of TF motifs across many human cell types, constructing a cell type specificity map for different TFs and identifying TFs with selective binding and differential occupancy across cell types. Second, we assess TOP's ability to elucidate the dynamics of TF occupancy across treatment conditions using time course data from human cells exposed to steroid hormones. Finally, to evaluate TOP's utility across varying genetic backgrounds, we predict quantitative TF occupancy for TF motifs in 70 Yoruba lymphoblastoid cell lines (LCLs) and map thousands of genetic variants associated with quantitative TF occupancy across individuals (which we term "topQTLs"). These topQTLs suggest specific mechanistic explanations for the functional impact of genetic variants within regulatory regions. In summary, TOP offers a cost-effective strategy for profiling the quantitative occupancy of hundreds of TF motifs using only a single chromatin accessibility experiment, markedly enhancing our ability to explore quantitative changes in TF occupancy across cell types, conditions, and genetic variants.

Results

Bayesian hierarchical regression accurately predicts quantitative TF occupancy from chromatin accessibility data

Training TOP entailed two basic steps, as illustrated in Figure 1. First, following the site-centric strategy, we used motif matches to enumerate candidate binding sites and extracted DNase- and/or ATAC-seq along with ChIP-seq data centered on each site for training. Second, we fit our Bayesian hierarchical regression model on spatially binned DNase- and/or ATAC-seq data. Owing to its hierarchical nature, once TOP is trained, we can use it to predict occupancy for any TF motif in any cell type or condition for which we have DNase- or ATAC-seq data, generating TF occupancy estimates for ChIP-seq experiments that have not been performed.

Specifically, in the first step, for each TF, we identified candidate binding sites by motif scanning with a permissive threshold (using FIMO) (Grant et al. 2011). Then, for each cell type, we considered (normalized) DNase and/or ATAC cleavage events occurring within 100 bp of the candidate binding site. Similarly, we quantified TF occupancy in terms of ChIP-seq read counts within 100 bp of the candidate binding site, and this served as the target of our regression when training TOP. We simplified the chromatin accessibility data into predictive features using five bins that aggregate the number of cleavage events occurring within the motif itself, as well as within two nonoverlapping flanking regions upstream and downstream; this is the same binning scheme used in the MILLIPEDE model (Luo and Hartemink 2013) and markedly reduces the potential impact of DNase digestion or Tn5 insertion bias (Luo and Hartemink 2013; He et al. 2014; Sung et al. 2014; Raj et al. 2015; Martins et al. 2018; Li et al. 2019b).

As an alternative, we tried extracting DNase features using wavelet-transformed multiscale signals from coarse to fine spatial resolution. However, after variable selection using Lasso (Tibshirani 1996), we found only the coarsest resolutions yielded significant features for predicting TF occupancy, whereas fine-resolution features were essentially irrelevant (Supplemental Fig. S1). Moreover, the simpler MILLIPEDE binning scheme achieved comparable or better prediction accuracy than optimally selected wavelet features (Supplemental Fig. S2). As an added benefit, when fitting TOP to a large number of different TFs across many diverse cell types, the five-bin scheme showed superior computational efficiency and better generality in capturing common features across TFs and cell types. Thus, the results that follow are all based on chromatin accessibility data aggregated into five bins.

We chose to use a Bayesian hierarchical model because it allows statistical information to be borrowed across TFs and cell types. TOP's hierarchical structure has three levels (Supplemental Fig. S3). The bottom level of the hierarchy contains model parameters specific to each TF × cell type combination for which ChIP-seq training data are available. In the middle level, one set of TF-specific but cell type–generic model parameters is shared across all training cell types for each TF. Finally, the top level has one set of TF-generic parameters jointly learned from all TFs. In other words, we obtain more general model parameters as we move to higher levels of the hierarchy. Once its parameters have been trained, TOP can quickly estimate occupancy for TFs (or TF families sharing a motif) in new cell types or conditions with DNase- or

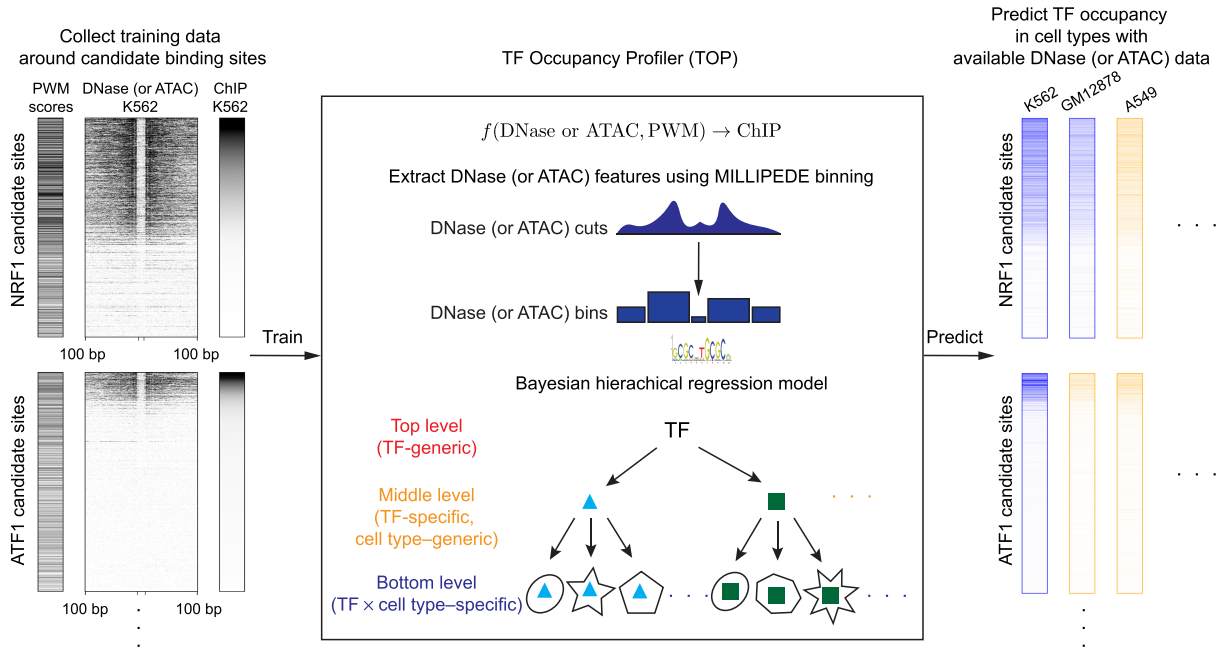


Figure 1. Schematic outline of the TF occupancy profiler (TOP) workflow. (*Left*) Collect training data. For a sequence-specific TF with a known PWM, compute its candidate binding sites throughout the genome. Then, around each of those sites, collect ChIP-seq and DNase- and/or ATAC-seq data from the same cell type. (*Center*) Extract DNase or ATAC features using MILLIPEDE binning and fit a Bayesian hierarchical regression model to the training data. Bottom-level models in the hierarchy make predictions in a TF × cell type–specific manner; middle-level models extend prediction in a TF-specific manner to new cell types; and the top-level model extends prediction in a TF-generic manner to new TFs. (*Right*) Predict TF occupancy at candidate binding sites across cell types. Blue columns indicate a cell type for which ChIP-seq measurements are available, allowing us to evaluate the predictive accuracy of our bottom-level models. Orange columns indicate a cell type for which we make novel predictions of TF occupancy using middle-level parameters of the hierarchical model.

ATAC-seq data by using a model from the appropriate level of the hierarchy.

We evaluated TOP's performance in terms of its fit to quantitative TF occupancy as measured experimentally by ChIP (Fig. 2). To avoid overfitting and allow easier comparison with other methods, we trained models using odd chromosomes and tested prediction performance using even chromosomes for all the comparisons shown in Figures 2 and 3. TOP predicted quantitative occupancy with varying degrees of accuracy across different TFs (Figs. 2A, 3). In light of technical differences and possible batch effects between DNase- and ATAC-seq data, or between DNase-seq data generated by different ENCODE laboratories, we trained separate hierarchical models for ATAC-seq and for DNase-seq data from Duke and from Washington (UW). Our results show comparable performance between DNase- and ATAC-seq (Figs. 2B,C, 3). We also fit a joint model using both DNase- and ATAC-seq features (with five DNase bins and five ATAC bins) but found only slight improvement over models trained using just one of DNase- or ATAC-seq, suggesting that collecting both DNase- and ATAC-seq data in an effort to improve prediction accuracy is unnecessary.

In general, although bottom-level models achieved the highest prediction accuracy, middle-level models performed equally well, and top-level models performed nearly as well (Fig. 2B–D). This indicates that for a TF that has been profiled with ChIP in some cell type, we can use the TF's middle-level model to predict its occupancy in any other cell type with available DNase- or ATAC-seq data. In addition, even for TFs that have never been profiled with ChIP, the top-level TF-generic model will still tend to provide good predictions of quantitative occupancy. Our predicted occupancy accurately matched quantitative ChIP-seq occupancy in various cell types and allowed us to explore TF occupancy in cell types like the embryonic stem cell line H9ES in which no TF ChIP data have been published to date (Fig. 2E). The quantitative predictions produce composite landscapes that sensitively reflect cell type–specific changes in TF occupancy.

Because our goal is to efficiently and accurately predict quantitative TF occupancy for candidate binding sites using only a single chromatin accessibility experiment, when comparing with alternative existing methods, we focused on site-centric methods that use only TF motif information and DNase (or ATAC) data, including CENTIPEDE (Pique-Regi et al. 2011), msCentipede (Raj et al. 2015), MILLIPEDE (Luo and Hartemink 2013), and BinDNase (Kähärä and Lähdesmäki 2015) (for a discussion of why these were chosen, see Methods). Although these predict TF binding in a site-centric framework, they only predict probabilities of TF binding rather than ChIP-seq read counts. However, because the CENTIPEDE paper showed a substantial correlation between its TF binding predictions (posterior log odds) and ChIP-seq read counts (sqrt transformed), we used the posterior log odds of TF binding as a proxy for quantitative ChIP-seq predictions. We also included total chromatin accessibility and cell-average ChIP occupancy at the candidate sites as baselines. Our results indicate TOP achieves greater—and in some cases markedly greater—accuracy than all the other methods on both DNase- and ATAC-seq data (Fig. 2B; Supplemental Fig. S4).

One important advantage of TOP is that it enables TF occupancy predictions across cell types or TFs that have not been profiled before. To show this, we trained a separate version of TOP by holding out ChIP training data for a random subset of TFs (and their related TF family members) and cell types. We then

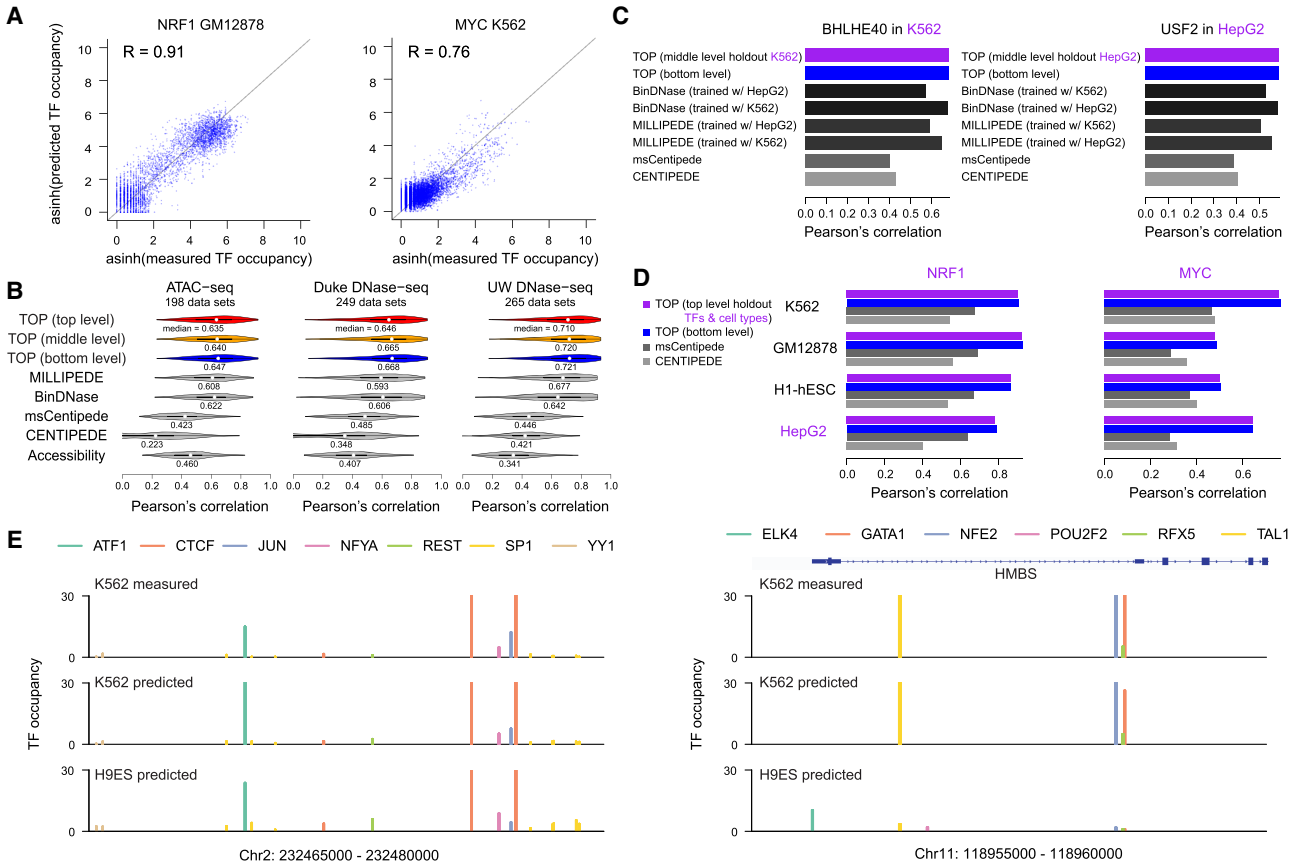


Figure 2. Evaluation of TOP results. (*A*) Scatter plots show predicted versus measured TF occupancy for test chromosomes of a specific TF in a specific cell type using Duke DNase-seq data, with dots representing the candidate binding sites across the genome. Model performance differs among TFs, as seen in the two examples. (*B*) Separately for ATAC-seq and DNase-seq data from Duke and UW protocols, violin plots show distribution of Pearson's correlations between predicted and measured TF occupancy (asinh transformed) in test chromosomes. Predictions were made with TOP models at each level of the hierarchy, in comparison with CENTIPEDE, msCentipede, MILLIPEDE, and BinDNase (using log odds of binding probability as a quantitative measurement of TF occupancy), as well as with total accessibility around candidate sites. (*C*) Comparing prediction performance in scenarios in which ChIP training data are missing for a cell type (indicated in purple). (*Left*) Predicting TF occupancy with data from K562 held out from training. (*Right*) Predicting TF occupancy with data from HepG2 held out from training. TOP is trained without any held-out data. MILLIPEDE and BinDNase are trained using a different cell type and also using data that was held out for TOP, showing that TOP performs as well without the held-out data as these methods do with it. msCentipede and CENTIPEDE are unsupervised, so do not require training data; however, their performance is poor. For more, see Supplemental Figure S5. (*D*) Comparing prediction performance in scenarios in which ChIP training data are missing for both TFs and a cell type (indicated in purple). Data from NRF1 and MYC (and all their TF family members with similar motifs), as well as from HepG2, were held out from training. MILLIPEDE and BinDNase were not included in this comparison as they require training data from the exact TFs. For more, see Supplemental Figure S6. (*E*) Predicted TF occupancy landscapes for two genomic regions in K562 and H9ES cell types. For K562, ChIP-seq data for these TFs are available and are displayed for comparison; for H9ES, no published ChIP-seq data are available so TOP provides a novel view of TF occupancy in this embryonic stem cell line. (*Left*) An example genomic region where the occupancy landscape did not change markedly between K562 and H9ES. (*Right*) An example genomic region near the *HMBS* gene (involved in heme biosynthesis) where GATA1, TAL1, and NFE2 showed clear cell type–specific occupancy.

compared predictions made using TOP with the held-out training set with the full training set, as well as the other methods. Our results indicate TOP achieves similar performance with parameters trained using the held-out training set as with the full training set and consistently outperforms all other methods (Fig. 2C,D; Supplemental Figs. S5, S6). In summary, these results show TOP's superior performance in making predictions across TFs and cell types.

TOP is trained to predict quantitative TF occupancy, which is one of its motivating applications and what distinguishes it from existing methods that are trained to make binary predictions about whether sites are bound or unbound. However, to allow comparison with those existing methods, we can instead train TOP with a logit link function. This logistic version of TOP performs as well or better in the binary prediction context as the best supervised methods (MILLIPEDE and BinDNase) and notably better than unsupervised methods (CENTIPEDE and msCentipede) or total accessibility (Supplemental Fig. S7). The incorporation of a hierarchical framework not only contributes to its improvement over MILLIPEDE and BinDNase but additionally enables predictions for TFs and cell types that are not present during training.

TOP reveals a spectrum of predictability across TFs and cell types

Across TFs, we observed a spectrum of predictability of TF occupancy, as indicated by the blue squares in Figure 3. Predictability was correlated with the degree of DNase depletion at the motif (Supplemental Fig. S8). For TFs with higher prediction accuracy, like NRF1 and ATF1, we observed clear profiles of depletion within motif regions and elevation at nearby flanking regions

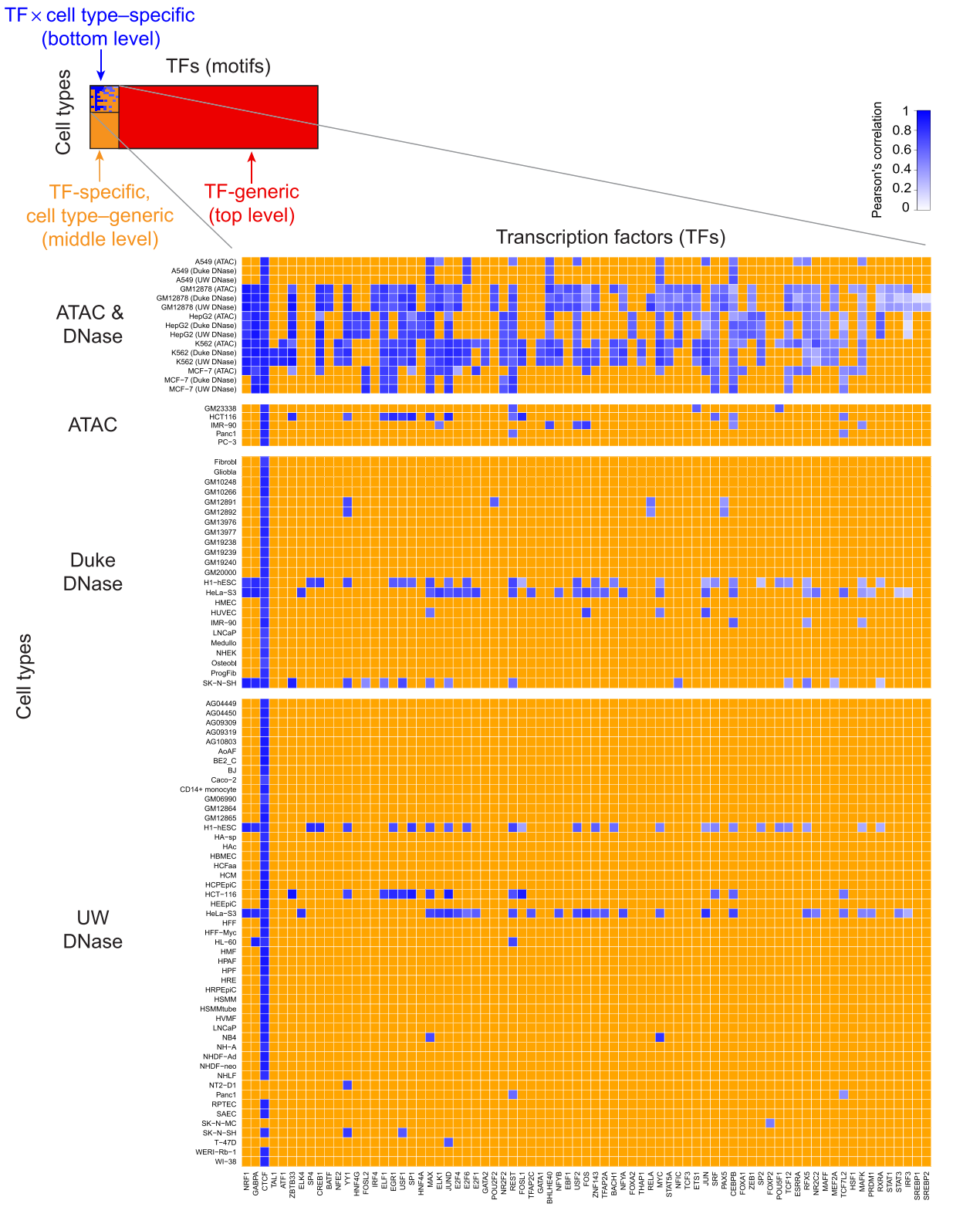


Figure 3. TOP provides quantitative TF occupancy profiles for hundreds of TF motifs across many cell types. (*Top, left*) Blue squares represent the TF × cell type combinations profiled with ChIP-seq as part of the ENCODE Project. For each of these TFs, we used a middle-level (TF-specific, cell type–generic) TOP model to generate new occupancy predictions across the rest of the cell types (orange squares). We then used a top-level (TF-generic) model to generate new occupancy predictions for the remainder of the TF motifs (red squares). (*Zoomed inset*) In the case of TF × cell type combinations with available ChIP-seq data, we computed the accuracy of TOP predictions; shades of blue indicate the correlation between predicted and measured occupancy. In this submatrix, TFs (columns) were sorted by average accuracy, revealing a spectrum of predictability. TFs toward the *left* were on average more predictable, whereas TFs to the *right* were less predictable.

(Supplemental Fig. S9), suggesting direct TF–DNA contact. Many of these TFs have previously been classified as pioneer factors, which directly open up chromatin and keep it open to allow other TFs to bind nearby (Sherwood et al. 2014). In contrast, TFs with lower prediction accuracy, like STATs and SREBPs, showed less marked elevation at nearby flanking regions and weak or no depletion at motif regions (Supplemental Fig. S9). Weaker DNase depletion profiles may result from transient binding with short residence time—known to occur with nuclear receptors and the AP-1 complex (Voss et al. 2011; Lickwar et al. 2012; Sung et al. 2014; Goldstein et al. 2017)—or from ChIP data that include many indirect binding events. For some TFs, we observed a high prediction accuracy in most cell types but a lower prediction accuracy in just one or two cell types. DNase profiles in the latter cases showed markedly weaker depletion (Supplemental Fig. S10). Many of those cases may be related to a low level of expression for the TF in those cell types.

TOP uses PWM scores to provide a priori information about how likely a site is to be bound in any cell type or condition. However, in the absence of genetic variation, the PWM score of a particular site does not change across cell types or conditions, so TOP's ability to quantify changes in TF occupancy in such situations depends entirely on changes in the chromatin accessibility data. As expected, when we compared them as single features, the overall level of DNase cleavage was almost always more correlated with ChIP-seq occupancy across cell types than was the PWM score (Supplemental Fig. S11).

Having established the reliability of TOP's predictions, we applied it to data from different contexts to illustrate the biological insights that arise from its ability to efficiently predict and compare quantitative occupancy for myriad TF motifs across conditions; each of the remaining three subsections explores one of these applications: changes in TF occupancy across different cell types, in response to dynamic environmental conditions, and in the context of genetic variation. In these applications, we focused on DNase-seq data, instead of ATAC-seq data, because DNase-seq data are currently available with matching ChIP-seq data in more cell types from ENCODE and are also available across 70 individuals from Yoruba LCLs to study genetic effects on predicted TF occupancy.

TOP maps out the cell type specificity of TF occupancy

TFs regulate gene expression in a cell type–specific manner. To assess TF occupancy differences across cell types, we constructed a cell type differential occupancy map to reveal distinct patterns in how TFs direct the gene regulation programs of different cell types. For each TF, we calculated the percentage of candidate sites in each cell type showing occupancy significantly higher or lower than the mean across cell types (FDR < 10%); we then clustered TFs on the basis of this measure of cell type specificity (Fig. 4A). Some TFs—including TAL1, GATA1, and NRF1—displayed large differences in occupancy among cell types, whereas the occupancy of other TFs—like the SPs—was quite cell type invariant (Fig. 4B). Lending credence to these results, we successfully recovered TFs known to be specifically or differentially expressed in certain cell types. For instance, as expected, we saw that POU5F1 (also known as OCT4) occupancy was significantly higher in stem cells; hepatocyte nuclear factors (HNFs) were higher in liver cells; GATAs were higher in K562; and REST was lower in medulloblastoma, etc.

To explore the relationship between a TF's concentration (here approximated by its gene expression level) and its occupancy, we computed the correlation between each TF's average level of occupancy in each cell type with its gene expression level in that same cell type, and observed several categories of TFs with different relationships (Fig. 4C; Supplemental Table S4). Many TFs showed significant positive correlations between their gene expression level and average occupancy, most of which are known to be cell type–specific TFs, such as FOSL2, HNF4A, FOXA1, and POU5F1 (Supplemental Fig. S12A). Three TFs (BATF, BHLHE40, and ZEB1, all known repressors) showed significant negative correlations (Supplemental Fig. S12B). This result is somewhat unexpected because increased levels of expression might generally be expected to correlate with increased occupancy (even for repressors), but many layers of post-transcriptional regulation and complicated regulatory dynamics make causal interpretation problematic. Future experiments probing the dynamics of expression and binding will be necessary to shed light on this result.

TOP monitors the dynamics of TF occupancy during hormone response

Nuclear hormone receptors are TFs specifically activated in response to hormone exposure. Once activated, they bind to specific hormone response elements (HREs) where they regulate gene expression, often in conjunction with the binding of cofactors and remodeling of the chromatin structure. Glucocorticoid (GC) receptor (GR; encoded by the *NR3C1* gene), androgen receptor (AR), and estrogen receptor (ER; including *ESR1* [also known as *ER-alpha*] and *ESR2* [also known as *ER-beta*] gene) are type I nuclear receptors, playing critical roles in immune response or reproductive system development, and are heavily involved in many types of cancer. To investigate TF occupancy dynamics in response to glucocorticoid, androgen, or estrogen stimulation, we predicted TF occupancy using DNase-seq data collected under each of these treatment conditions. For GC treatment, we conducted DNase-seq experiments in A549 cells (human alveolar adenocarcinoma cell line) over 12 time points from 0 to 12 h of GC exposure (McDowell et al. 2018). For androgen treatment, we collected DNase-seq data in LNCaP cells (human prostate adenocarcinoma cell line) over four time points from 0 to 12 h following androgen induction (Tewari et al. 2012). For estrogen treatment, we used published DNase-seq data before and after estrogen induction in two kinds of cells: Ishikawa (human endometrial adenocarcinoma cell line) and T-47D (human ductal carcinoma cell line) (Gertz et al. 2013).

We identified sites with significantly differential TF occupancy before and after estrogen induction, as well as over the full time courses for GC and androgen treatment. We then ranked TFs based on the percentage of sites showing significantly increased or decreased occupancy in response to treatment. We grouped TFs with similar motifs together using RSAT clusters (Castro-Mondragon et al. 2017) and present results for all significant clusters in Figure 5 (results for individual TFs in Supplemental Fig. S13).

We observed different sets of TFs enriched in response to GC, androgen, and estrogen. In the list of most dynamic clusters for GC response (Fig. 5A), GR was ranked at the top—consistent with recent results showing that motif-driven GR binding is the most predictive feature of GC-inducible enhancers (Vockley et al. 2016; McDowell et al. 2018)—followed closely by CEBP (McDowell et al. 2018). FOX and GATA clusters appeared next, and in both cases, although we identified more sites whose occupancy increased over the time course, we also detected a significant number that decreased.

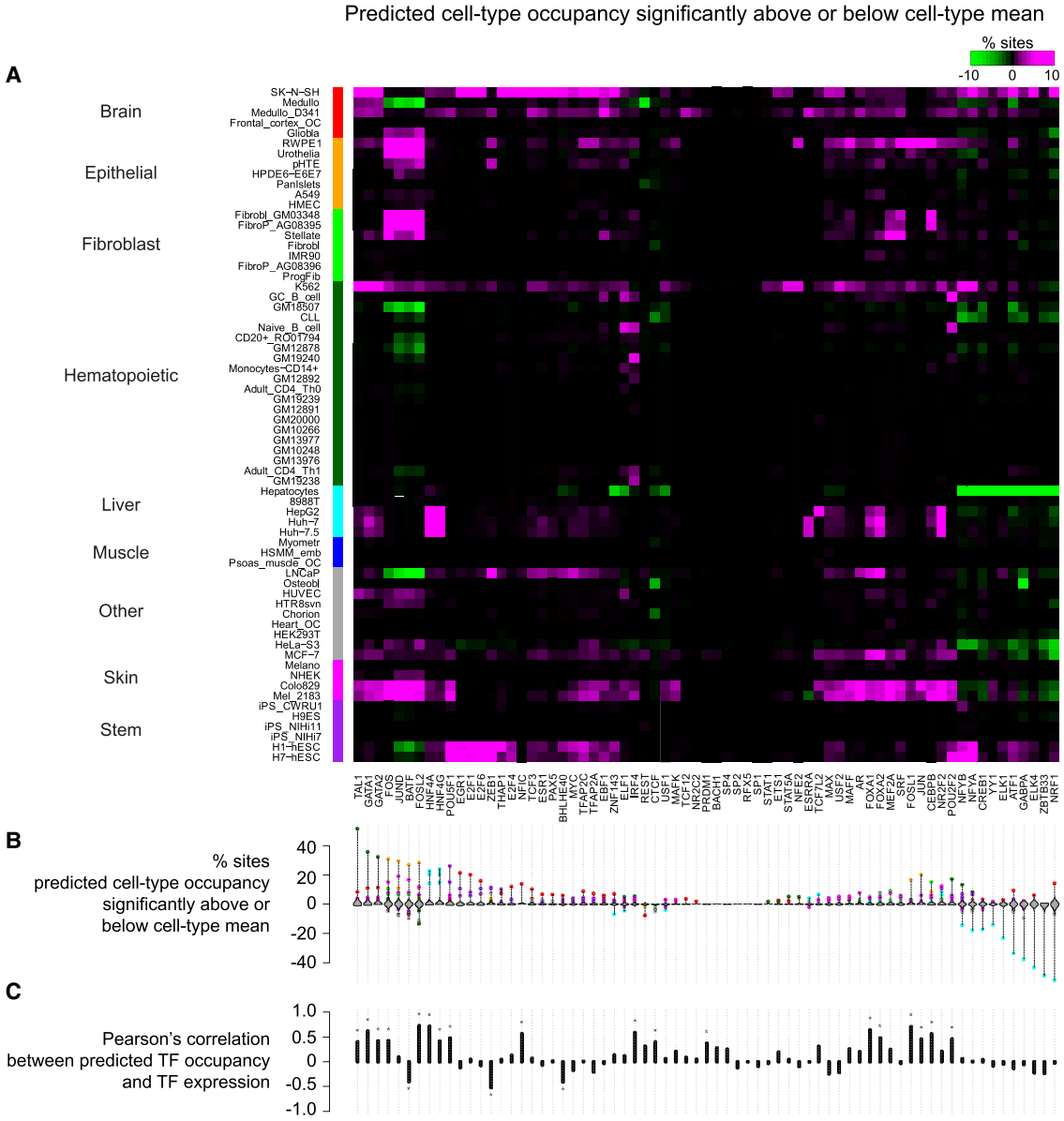


Figure 4. Cell type–specificity matrix of TFs. (*A*) Percentage of sites with predicted cell type occupancy significantly above or below predicted cell type mean occupancy (FDR < 10%). Cell types were first grouped by lineage (ordered alphabetically) (Sheffield et al. 2013) and, within each lineage group, were ordered by hierarchical clustering. TFs were ordered by hierarchical clustering (with optimal leaf ordering) (Bar-Joseph et al. 2001). (*B*) Violin plots shown for each TF the distribution across cell types of the percentage of sites showing significantly differential occupancy. Colored dots highlight cell types with at least 3% of sites showing significantly differential occupancy (color reflects lineage of cell type; e.g., liver and brain show frequent differential occupancy). (*C*) Pearson's correlation across cell types between average predicted occupancy and gene expression of each TF (in this plot, we used only Duke DNase data because corresponding gene expression was measured in each of the cell types).

Among TFs whose occupancy was predicted to be significantly responsive to androgen treatment (Fig. 5B), AR was at the top of the list, followed by the FOX cluster. Clusters showed very few sites with decreasing occupancy along the time course. These observations are consistent with our previous findings that androgen induction mainly leads to an increase in chromatin accessibility and that AR and FOXA1 are key TFs with increased occupancy (Tewari et al. 2012). The fact that the occupancy of many AR and FOXA1 sites increased gradually over the duration of the time course (Supplemental Fig. S14A) highlights the importance of a quantitative perspective of TF occupancy.

In the case of estrogen induction (Fig. 5C), ER was ranked at the top, followed closely by the NFY cluster. The DNase digestion profiles flanking NFYA binding sites showed oscillation patterns similar to those observed within nucleosomes (Supplemental Fig. S15; Zhong et al. 2016). This is consistent with previous reports that NFYA has nucleosome-like properties and plays an important role in maintaining chromatin structure (Nardini et al. 2013; Sherwood et al. 2014).

That different TFs were enriched in these lists may be partly due to cell type differences but also suggests different utilization of cofactors for GR, AR, and ER binding in response to hormone

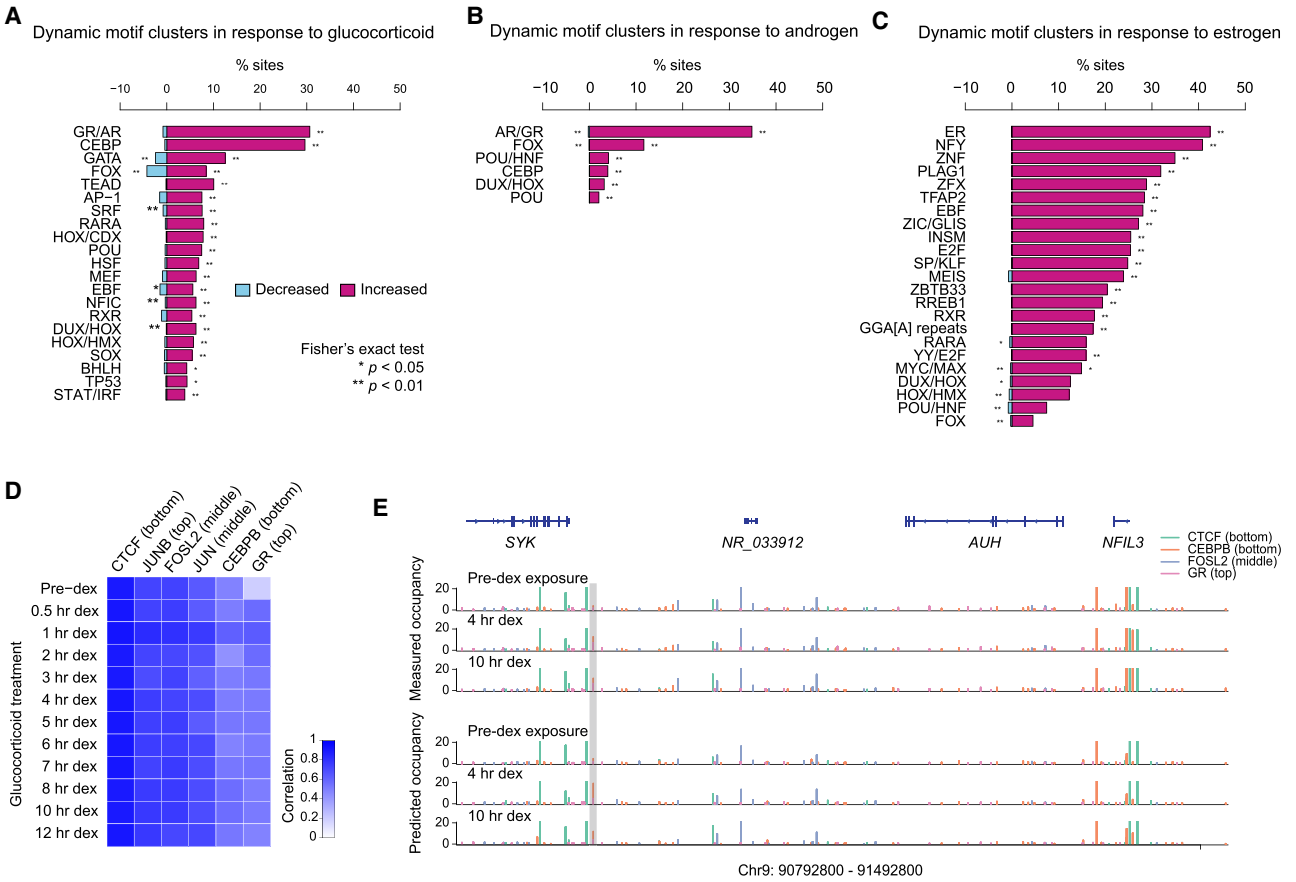


Figure 5. TF occupancy dynamics in response to hormone stimulation. (*A*) Motif clusters were ranked by the percentage of candidate sites whose predicted occupancy showed either a linear increasing or decreasing trend along the 12 time points of glucocorticoid (GC) treatment. Only significant dynamic motif clusters (*P*-value < 0.05) are listed. (*B*) Similar to *A*, but along the four time points of androgen treatment. (*C*) Similar to *A* and *B*, but before and after estrogen treatment. Because DNase data were collected at 12 time points during treatment with GC, at four time points with androgen, and at only two time points with estrogen, numbers are not necessarily comparable between different experiments in *A*, *B*, and *C*. (*D*) Prediction accuracy for six TFs was evaluated afterward using subsequently generated ChIP-seq data (McDowell et al. 2018). Shades of blue indicate the correlation between predicted and measured occupancy for each of the six TFs at each time point. Columns (TFs) were sorted by average accuracy across the 12 time points. (*E*) Measured and predicted TF occupancy landscapes of CTCF, CEBPB, FOSL2, and GR in an example genomic region on human Chromosome 9. Predicted occupancy corresponded well with measured occupancy across time, for example, revealing in the highlighted region how CEBPB occupancy at this site increased following GC treatment.

stimulation. We observed that PWM scores were significantly higher in sites with increased occupancy than sites of unchanged occupancy for GR, AR, and ER but not for CEBPB, FOXA1, or NFYA (Supplemental Fig. S14B), indicating that motif strength for GR, AR, and ER may play a role in prioritizing the selection of binding sites in response to hormone stimulation. This accords with results indicating that GR motif strength is predictive of GC-induced enhancer function (Vockley et al. 2016).

To independently validate our occupancy predictions with data not seen during training, we compared our predictions throughout the GC time course with ChIP-seq data collected in the same experiment (Fig. 5D,E; McDowell et al. 2018). We computed the correlation between measured and predicted occupancies for CTCF, JUNB, FOSL2, JUN, CEBPB, and GR. Across all six TFs and 12 time points, the average correlation was 0.70. Over the time course, it was lowest before treatment (0.63) but otherwise consistent (between 0.68 and 0.72). Among TFs, predictions were the most accurate for CTCF (0.91)—not surprising given how predictable we observed it to be (Fig. 3)—and least for GR (0.52). Two reasons for the lower accuracy of GR are that we used a top-level model because GR was not included in our training set (owing to potential quality concerns with the GR ChIP-seq data from ENCODE) and that GR is known to have a weak DNase footprint (Goldstein et al. 2017). The correlation is particularly low before treatment (time point 0), consistent with observations that many GR binding sites occur at regions of the genome that are already open before GC exposure (Reddy et al. 2012). We also noticed that some subtle and transient dynamics of TF occupancy measured by ChIP-seq were not captured by predictions based on DNase-seq data (Supplemental Fig. S16).

TOP identifies genetic variants associated with predicted TF occupancy (topQTLs) and provides mechanistic interpretations for dsQTLs

A large majority of genetic variants associated with complex traits are located in noncoding genomic regions (Hindorff et al. 2009), suggesting roles in transcriptional regulation. To elucidate this, it is imperative that we continue to identify genetic variants affecting TF occupancy and chromatin dynamics. To examine whether

TOP is capable of sensitively distinguishing quantitatively differential TF occupancy across individuals or genetic variants, we predicted CTCF occupancy in LCLs from two trio studies, one from a CEPH Utah (CEU) family and one from a Yoruba from Ibadan (YRI) family (McDaniell et al. 2010; The 1000 Genomes Project Consortium 2010). TOP successfully identified differential CTCF occupancy between individuals across CEU and YRI families (Fig. 6A) and was sensitive enough to capture quantitative differences in CTCF occupancy between allele genotypes at allele-specific sites within CEU and YRI families (Fig. 6B).

Encouraged by this result, we extended our predictions of genome-wide quantitative occupancy to approximately 1500 TF motifs across 70 Yoruba LCLs using TOP applied to previously published genotype and DNase-seq data (Degner et al. 2012). With the resulting TF occupancy profiles across 70 individuals, we applied a QTL mapping strategy to identify genetic variants whose genotypes were significantly associated with changes in predicted TF occupancy, which we called "topQTLs." Because genetic variants that change TF motifs often affect TF binding occupancy by changing DNA binding affinity (Deplancke et al. 2016), we focused our attention on SNPs within TF motif matches because these have the highest potential for causal interpretation.

We compared topQTLs within motif matches to a subset of dsQTLs that we call "localizable dsQTLs," dsQTLs that fall inside the 100-bp windows with which they are linked and also lie within TF motif matches (Fig. 6C). Of the 1230 reported dsQTLs that were localizable, 943 of them were topQTLs (this number increased to 1000 when using FDR < 20%, whereas 1141 [93%] were associated with a significant change in predicted TF occupancy under a less stringent threshold of P-value < 0.05). In so doing, we identified genetic variants associated with the occupancy of TFs that are likely to drive observed changes in chromatin accessibility, providing a more mechanistic interpretation for localizable dsQTLs. Moreover, we identified more than 6000 additional topQTLs that were not reported as dsQTLs. Among RSAT-clustered motifs, CTCF, STAT/IRF, SP/KLF, ETS, AP-1, POU, NF-kB, and RREB1 motifs had the greatest number of topQTLs (each well over 200); most of these factors are known to be active in LCLs and critical for immune cell development (Degner et al. 2012; Tehranchi et al. 2016). Figure 6D shows three sample topQTLs: one for NF-kB that is a nonlocalizable dsQTL, another for NF-kB that is a localizable dsQTL, and one for CTCF that is not reported as a dsQTL.

That CTCF had the largest number of topQTLs, over 1300, is noteworthy because CTCF plays a key role in chromosomal looping and commonly demarcates the boundaries of topologically associating domains (TADs) (Rao et al. 2014). A genetic variant that disrupts a CTCF motif not only may have a significant impact on occupancy at loop anchor sites but also could disrupt TAD boundaries. Such disruption has been shown experimentally and pathologically to dysregulate the chromatin landscape and the expression of genes within the affected TAD (Lupiáñez et al. 2016).

Tehranchi et al. (2016) conducted pooled ChIP-seq experiments for five TFs (NF-kB, SPI1, JUND, STAT1, and POU2F1 [also known as OCT1]) across 60 LCLs, and identified SNPs showing pre-ChIP versus post-ChIP differential allele frequencies, which they called binding QTLs (bQTLs). Among our topQTLs that are also identified as bQTLs, we observed a strong directionality agreement between allele preference in topQTLs and bQTLs (Supplemental Fig. S17). However, the overlap between bQTLs and topQTLs is low. This can happen when bQTLs reside in regions without detectable changes in chromatin accessibility, for instance, when a TF shows transient binding or low residency, like NF-kB. Alternatively, it can happen when bQTLs do not reside within a match to any TF motif; indeed, <0.9% of bQTLs reside within their own motif matches (Tehranchi et al. 2016), which may reflect the inability of ChIP to distinguish direct versus indirect binding. Although <0.9% of bQTLs are located in their own motif matches, >10% of bQTLs can be explained by the topQTLs of different motifs (Supplemental Fig. S17), suggesting that topQTLs may help us gain a mechanistic understanding about direct binding at these loci. Consistent with Tehranchi et al. (2016), CTCF topQTLs were enriched in bQTLs, highlighting the functional significance of CTCF and its topQTLs (Supplemental Fig. S17).

Discussion

We introduce TOP to accurately predict quantitative ChIP-seq occupancy using chromatin accessibility data from DNase- or ATAC-seq. TOP effectively learns both TF-specific and TF-generic model parameters among TFs and across cell types using a Bayesian hierarchical regression framework. TOP uses a supervised learning strategy, trained with existing ChIP-seq data, TF binding specificity data (motifs), and DNase- or ATAC-seq data, yet can accurately predict TF occupancy for new conditions, cell types, or TFs owing to its hierarchical structure. In contrast to traditional ways of analyzing ChIP-seq data by calling peaks in order to label genomic regions as bound or unbound, TOP adopts a quantitative perspective, allowing us to predict the level of TF occupancy along a continuum. This opens up a new way to investigate quantitative changes in TF occupancy across cell types, treatment conditions, and developmental time courses. TOP is general in that it can predict occupancy for sequence-specific TFs of interest with new DNase- or ATAC-seq data in any cell type or condition without requiring a new ChIP-seq experiment. For example, TOP's ability to use time course DNase data over 12 h of GC treatment served as a cost-effective strategy to study the temporal dynamics of TF occupancy: By doing one DNase-seq experiment at each time point, we obtained occupancy predictions for hundreds of TF motifs, allowing us to screen for TFs showing significant changes in occupancy. For example, TOP results suggest a significant role for FOX and GATA factors in GC-induced transcriptional response (Fig. 5A). Although TOP was trained on human data, it could equally be trained and applied in other organisms. As an example, we were able to successfully predict quantitative Reb1 occupancy (ChIP-exo) in the yeast genome from DNase-seq data (Supplemental Fig. S18). As a resource for the community, we provide predicted occupancy for hundreds of JASPAR TF motifs across hundreds of ENCODE cell types, throughout a 12-h time course of GC exposure, and across 70 LCLs. TOP makes use of existing ChIP-seq data generated by ENCODE and other projects and learns a model that extends to more cell types and conditions, allowing people to generate preliminary TF occupancy estimates for ChIP-seq experiments that have not yet been performed.

Recently, other methods have emerged for the imputation of missing epigenomic data (histone modifications, chromatin accessibility, etc.) using the many types of available data generated by ENCODE (Ernst and Kellis 2015; Durham et al. 2018; Schreiber et al. 2020b). Our approach shares a goal with these imputation methods in trying to predict unmeasured data using models trained on existing data sets across multiple cell types. However, we note a few major distinctions between our approach and these recent imputation strategies. First, our approach requires only DNase- or ATAC-seq and TF ChIP-seq data for training and requires

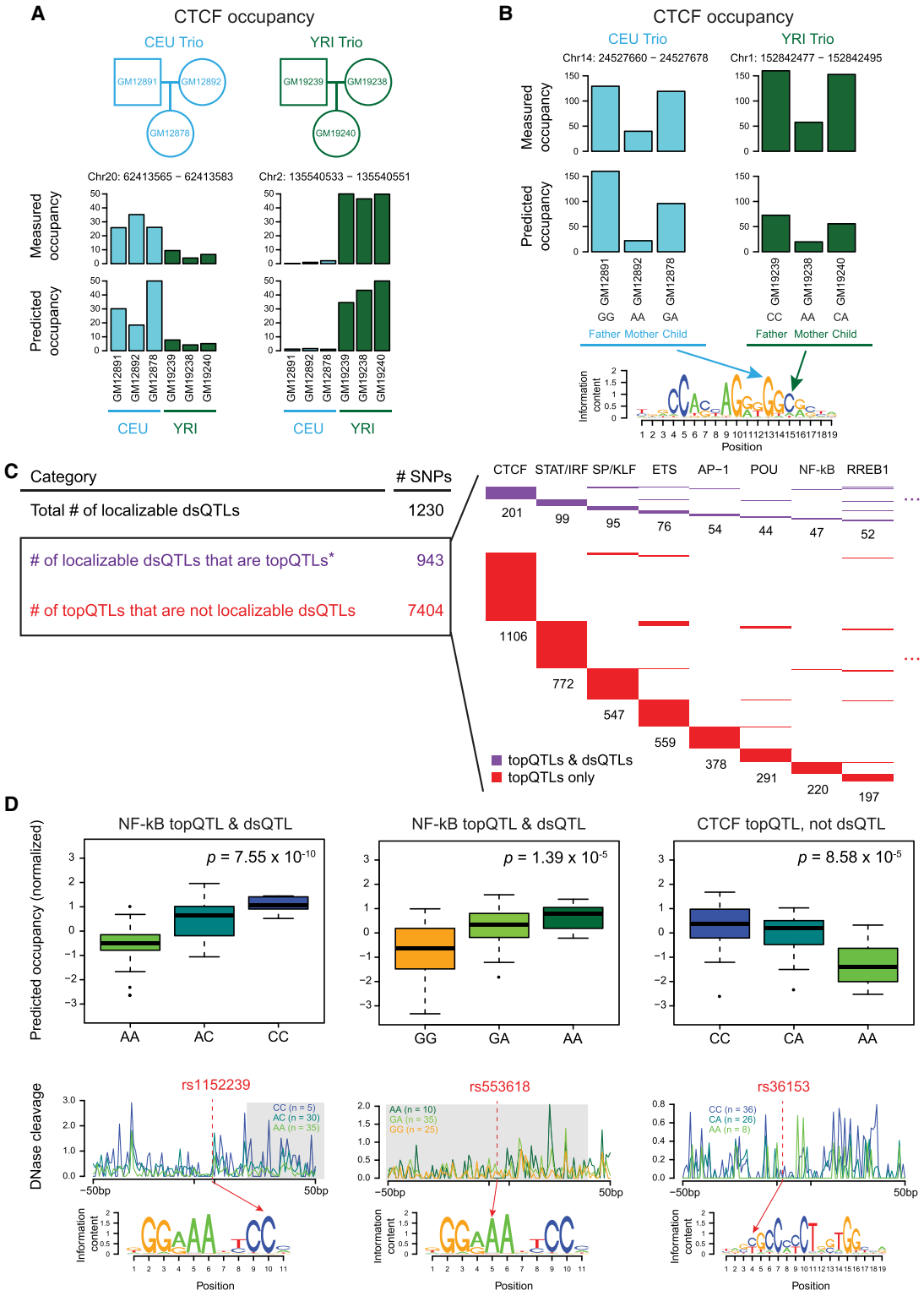


Figure 6. TF occupancy profile QTLs (topQTLs). (*A*) Measured and predicted CTCF occupancy at two individual-specific example loci with significantly differential CTCF occupancy between CEU and YRI families. (*B*) Measured and predicted CTCF occupancy at two allele-specific example loci with significantly differential CTCF occupancy within CEU and YRI families. (*C*) Intersections of topQTLs with "localizable dsQTLs" (those within their own 100-bp windows and also within motif matches). topQTLs were defined with FDR < 10%; (*) with FDR < 20%, the number of localizable dsQTLs that are also topQTLs becomes 1000. (*Right*) Largest motif clusters for topQTLs are displayed in the matrix; each row represents one topQTL that can be explained by one or more motif clusters in the columns. (*D*) Examples of topQTLs showing normalized allele-specific predicted occupancy; average DNase digestion profiles within 50 bp of the motif for each allele (significant dsQTL windows shaded in gray); and SNP locations within motifs. The CTCF topQTL overlapped a measured CTCF ChIP-seq peak in multiple LCLs but was not identified as a dsQTL.

only DNase- or ATAC-seq data to make predictions. In contrast, existing imputation methods often require a large variety of existing assays (DNase- or ATAC-seq, RNA-seq, histone modification ChIP-seq, etc.), which may not be readily available, especially in studies that profile new cell types or treatment conditions. Second, our strategy predicts TF occupancy only at candidate binding sites (based on low stringency motif matches), whereas existing imputation approaches attempt to impute a TF's ChIP-seq signal across the entire genome. Each strategy has its own advantages. The site-centric strategy we adopt here is able to more effectively model the DNase or ATAC-seq features around motif matches but will miss binding sites that do not match known motifs. In contrast, methods that impute a TF's ChIP-seq signal across the entire genome are not limited to candidate binding sites but will devote statistical power and computational effort to genomic locations where a given TF is unlikely to bind, which is the overwhelming majority. Third, TOP uses a Bayesian hierarchical regression framework to model chromatin accessibility features, which allows for easier interpretation (by examining the regression coefficients learned from the model) than more complex methods, especially those involving deep neural networks or tensor factorization. Last but not least, our hierarchical model is able to predict the binding of TFs that have never been assayed by ChIP, a significant advantage over imputation methods that require ChIP-seq training data for TFs of interest.

The fact that TOP predicts TF occupancy only at candidate binding sites is a limitation because it has been observed that for many TFs, a large number of their ChIP-seq peaks do not have motif matches (Deplancke et al. 2016). On the other hand, this does have some benefits, because distinguishing direct from indirect binding can be difficult using ChIP assays. By focusing only on motif matches, our results can be viewed as predictions of TF occupancy that are explainable by direct binding. Indeed, TOP could be used to suggest direct-binding TFs that may be mediating the indirect binding of other TFs in ChIP experiments (Gordân et al. 2009). As a last observation, because motif quality directly affects which genomic locations are selected as candidate binding sites and because PWM scores also factor into TOP predictions, better TF binding affinity models should improve TOP's predictions in the future. Another limitation of TOP (and other motif-based methods) is that it is hard to precisely distinguish TF family members without any additional information about their expression in the monitored cell type or their binding profiles in comparable cell types. Therefore, we note that the TOP model predictions and applications would be limited to TF families, or more accurately, motifs.

Our study provides a comprehensive survey of quantitative occupancy for multiple TFs across multiple cell types. Our results show a wide spectrum of predictability across TFs, providing a reference for the reliability of computational predictions. Although the occupancy of many TFs can be predicted at high accuracy, our results also suggest that computational predictions based on motif information and chromatin accessibility data may not be sufficient to reproduce ChIP measurement for many TFs. In addition, from our analysis of the GC treatment time course data, we found some subtle and transient dynamics of TF occupancy measured by ChIP-seq were not captured by predictions based on DNase-seq data, especially for TFs with more transient binding, like GR. Therefore, ChIP experiments will still be needed to provide insight into TF occupancy, especially for TFs that are harder to predict.

Our approach can be viewed as complementary to ChIP-based exploration of TF occupancy. It is not intended to recapitulate all experimentally detected ChIP-seq signals. Rather, TOP makes use of existing ChIP-seq data that have been generated by ENCODE and other projects, and amplifies to more cell types and conditions, allowing people to generate preliminary TF occupancy estimates for ChIP-seq experiments that have not yet been performed. Instead of doing one ChIP-seq experiment for every TF in a particular cell type or condition, TOP needs only one DNase- or ATAC-seq experiment to predict the genome-wide occupancy of many TFs. TOP can therefore be used to screen and identify TFs showing significant changes in occupancy, enabling the prioritization of future ChIP experiments for a small number of key TFs. The modeling strategy we present here offers a foundational and cost-effective approach for profiling the quantitative occupancy of myriad TFs across diverse cell types, dynamic conditions, and genetic variants.

Methods

Candidate binding site selection

We defined candidate TF binding sites by PWM scanning across the genome using FIMO (Grant et al. 2011). When training or applying our model, we included as candidate sites all motif matches with P-value $< 10^{-5}$. Similar to CENTIPEDE (Pique-Regi et al. 2011) and MILLIPEDE (Luo and Hartemink 2013), we filtered out candidate sites if >10% of the nucleotides in the surrounding window (100 bp flanking each side of the motif) were unmappable or overlapped with ENCODE blacklist regions.

When training the regression model, if the training TF had more than one motif, we manually selected one based on which was the most representative motif for that TF in the Factorbook database (Wang et al. 2012). After training, we used the model parameters estimated at various levels of the TOP hierarchy to make occupancy predictions for motifs from JASPAR. The motifs selected for training the model, along with the full list of all motifs used for prediction in this paper, are provided in Supplemental Tables S2 and S3.

Normalization and data preprocessing

To account for differences in sequencing depth across experiments in different cell types or conditions, DNase-seq, ATAC-seq, and ChIP-seq data were normalized by library size. This simple library size normalization is flexible for downstream analysis. We considered other types of normalization methods, including quantile normalization, trimmed mean of M-values (TMM), etc. However, these methods assume different experiments will have the same distribution of reads across peaks (or a subset of common peaks) among all the experiments, which is too strong an assumption in our case—especially, for example, when comparing hormone receptor binding before and after hormone induction—and leads to a high number of false negatives in the GR, AR, and ER analyses.

To address the offsets inherent in ATAC-seq reads, we shifted their start positions to align the signal across strands, thereby obtaining more accurate Tn5 binding locations (Buenrostro et al. 2013).

Feature extraction using binning versus wavelet coefficients

We systematically evaluated different features of DNase or ATAC cleavage events in an attempt to avoid overfitting (Raj et al. 2015) and any possible influence of sequence bias arising from a DNase- or ATAC-seq experiment. First, we tried extracting multiresolution features of DNase digestion data using wavelet multiresolution decomposition. Wavelet methods provide a natural

approach to extract the multiresolution information contained in both DNase cut magnitude and detail profiles. We decomposed DNase-seq data using Haar wavelets (Haar 1910) with the wavethresh package (version 4.6.8, https://cran.r-project.org/web/packages/wavethresh/index.html) in R (R Core Team 2020). The detail signals were extracted at different resolution levels through the mother wavelet coefficients, whereas the scales of cuts at different resolution levels were represented by the father wavelet coefficients. We started with windows of size 128 bp around the motif center but later focused on 64-bp windows around the motif centers, because that was where the majority of the largest mother wavelet coefficients were located. Then we fit regression models with mother wavelet coefficients and log-transformed father wavelet coefficients at multiple resolution levels as predictors, together with PWM score, and conducted variable selection with Lasso. Variable selection results suggested the scale of DNase cuts (represented by the father wavelet coefficients) was the most significant feature for predicting TF occupancy (Supplemental Fig. S1), consistent with previous findings (Cuellar-Partida et al. 2012; Luo and Hartemink 2013; He et al. 2014). In contrast, very few spiky DNase signals (represented by the mother wavelet coefficients) were selected. Worse, some of the fine details in the DNase signal in the motif region might arise from sequence-specific DNase digestion bias (Cuellar-Partida et al. 2012; Luo and Hartemink 2013; He et al. 2014; Sung et al. 2014; Yardımcı et al. 2014; Martins et al. 2018).

In a previous work, we developed MILLIPEDE, a model that divides the motif region and its flanking regions upstream and downstream into various distinct bins (Supplemental Fig. S2A; Luo and Hartemink 2013). Following the binning scheme of MILLIPEDE, we compared different binning models, from the most complicated M12 model to the simplest M1 model, and evaluated their performance in comparison to an optimally selected wavelet model. Supplemental Figure S2B shows the prediction performance of all these models for four TFs in K562 cells using five-fold cross-validation. In summary, different binning models led to roughly similar prediction performances and were generally comparable to a model using optimally selected wavelet features. Based on these empirical observations of DNase digestion profiles around motifs, we simplified the process of feature extraction by using a more flexible binning scheme in place of the rigid dyadic splitting of the wavelet framework. Based on these results, we chose M5 binning—which effectively summarizes the number of DNase cleavage events in the motif region, nearby flanking regions, and distal flanking regions on both sides of the motif—to capture the chromatin accessibility features. It is simple enough to fit into the Bayesian hierarchical regression framework and still yield easily interpretable TF-specific and TF-generic signatures.

Bayesian hierarchical regression model

We designed the hierarchical model to have three levels, with cell types nested within TF branches. (In principle, we could expand the hierarchical model to have an additional branch with parameters for each cell type, i.e., a cell type–specific but TF-generic model. However, we expect a TF to have similar model parameters in different cell types. Also, the majority of TFs have not been profiled with ChIP-seq in many cell types, so we would likely have insufficient data to estimate cell type–specific parameters for most cell types.)

ChIP-seq count data are typically fit using a negative binomial distribution, which uses an extra parameter to model the overdispersion in ChIP-seq data better than a Poisson distribution. We found a Gaussian linear model on asinh-transformed ChIP-seq data to be a better choice for fitting our Bayesian hierarchical model (asinh transformation is similar to log transformation but handles zero values more gracefully; we used it successfully in our NucID model [Zhong et al. 2016]). This choice has the added benefit of applying to noninteger data, which arise whenever we average counts over replicate experiments or conduct data normalization. It also performed well in predicting the ENCODE ChIP-seq signal values, which are fold-over-control values (also noninteger). We compared the prediction accuracy of our Gaussian linear model on asinh-transformed ChIP-seq data against the alternative of negative binomial regression on ChIP-seq count data and observed very close agreement. We ultimately decided to use the Gaussian distribution because it has a nice conjugacy property, allowing posteriors to be estimated through Gibbs sampling and thereby providing a computational advantage over a negative binomial distribution.

The basic regression model for modeling the asinh-transformed ChIP-seq occupancy $y_{t,c,i}$ that is observed when TF t occupies its candidate binding site i in cell type c can be briefly summarized as

$$y_{t,c,i} \sim \text{Normal}(\mu_{t,c,i}\ ,\ v_{t,c}),$$

where

$$\mu_{t,c,i} = \beta_{t,c}^{(0)} + \sum_{j=1}^{J} \beta_{t,c}^{(j)} \times \text{D}_{t,c,i,j} + \beta_{t,c}^{J+1} \times \text{PWM}_i.$$

$\text{D}_{t,c,i,j}$ represents DNase (or alternatively ATAC) feature j for site i of TF t in cell type c, and J is the number of DNase (or ATAC) features in the model (in our final model, we use M5 binning so $J=5$).

The Bayesian hierarchical model is specified as follows:

$$\begin{aligned}
\beta_{t,c}^{(j)} &\sim \text{Normal}(b_t^{(j)},\ 1) \quad \forall j \in \{0,\ 1,\ \ldots,\ j+1\} \\
b_i^{(j)} &\sim \text{Normal}(B^j,\ 1) \quad \forall j \in \{0,\ 1,\ \ldots,\ j+1\} \\
B^j &\sim \text{Normal}(0,\ 1) \quad \forall j \in \{0,\ 1,\ \ldots,\ j+1\} \\
\frac{1}{v_{t,c}} &\sim \text{Gamma}(\tau_t^2,\ \tau_t) \\
\tau_t &\sim \text{Gamma}(T^2,\ T) \\
T &\sim \text{Gamma}(1,\ 1).
\end{aligned}$$

We used the consensus Monte Carlo algorithm (Scott et al. 2016), a parallel technique to reduce the running time of the Gibbs sampler while maintaining predictive performance. Briefly, we split all data randomly into 10 equal parts. Gibbs samplers were run on each part separately in parallel, and then posterior samples from the ten Gibbs samplers were averaged to get the final posterior samples for each model's parameters. Once trained, the model can be quickly applied to predict TF occupancy across many conditions.

To evaluate model performance, we used odd chromosomes for training and even chromosomes for testing. Figure 2 shows the prediction performance of the TOP model using M5 binning on the test data. For completeness, we also experimented with several other versions of the Bayesian hierarchical model: (1) We tested allowing the variances of the different beta parameters to be learned from the data using an inverse gamma hyperprior, but the results were essentially unchanged; (2) we implemented a model with 12 bins (using the M12 binning scheme) (Luo and Hartemink 2013) and observed very close prediction performance with the M5 model with five bins; (3) we tried a more complicated version of the model with an extra level of the hierarchy to model the heterogeneity of DNase- or ATAC-seq replicates and found very similar performance with the simpler model trained with the replicates pooled together; and (4) for cell types for which both DNase- and ATAC-seq data were available, we tried jointly modeling both DNase and ATAC features, using five DNase bins and five

ATAC bins together with PWM score as model covariates. The joint model achieved only marginal improvement over models trained by either DNase- or ATAC-seq data alone. This indicates collecting both DNase- and ATAC-seq data from the same cell type or condition is not necessary to make more accurate predictions.

When we wanted to compare how TOP performs in predicting binary TF binding (bound vs. unbound), we implemented a separate hierarchical logistic version of the model. Instead of quantitative ChIP-seq occupancy, we trained this model with binary ChIP-seq peak labels. Under the logistic regression framework, the binary status $y_{t,\,c,\,i}$ of a TF t being bound at its candidate binding site i in cell type c follows a Bernoulli distribution with a binding probability $p_{t,\,c,\,i}$:

$$y_{t,c,i} \sim \text{Bernoulli}(p_{t,c,i}),$$

where

$$\log\left(\frac{p_{t,c,i}}{1-p_{t,c,i}}\right) = \beta_{t,c}^{(0)} + \sum_{j=1}^{J} \beta_{t,c}^{(j)} \times \text{D}_{t,c,i,j} + \beta_{t,c}^{j+1} \times \text{PWM}_i.$$

We used the same DNase (or ATAC) features and the same normal priors for the beta regression coefficients as specified above.

Comparison of prediction accuracy with existing methods in a site-centric framework

We compared TOP with four existing methods in a site-centric framework. CENTIPEDE and msCentipede predict TF binding probabilities using an unsupervised generative framework to model the DNase- or ATAC-seq footprint profiles around candidate sites (motif matches) without ChIP-seq training data. msCentipede improves on CENTIPEDE by using a multiscale model framework to better model heterogeneity across sites and replicates. We ran CENTIPEDE and msCentipede on DNase- or ATAC-seq data in each TF–cell type combination under default parameter settings. CENTIPEDE was run on DNase or ATAC data after pooling the replicate samples. msCentipede was run on individual DNase or ATAC replicates to better capture heterogeneity (its investigators showed that replicates are beneficial to its accuracy). Because the CENTIPEDE paper showed a substantial correlation between its TF binding predictions (posterior log odds) and ChIP-seq read counts (sqrt transformed), we used posterior log odds of TF binding probabilities as predicted quantitative occupancy. In contrast to CENTIPEDE and msCentipede, MILLIPEDE (Luo and Hartemink 2013) adopts a supervised learning strategy using a logistic regression framework with binary ChIP-seq peaks as training labels and TF-generic binning to extract DNase digestion features (similar to TOP). BinDNase (Kähärä and Lähdesmäki 2015) is a later method that is very similar to MILLIPEDE but allows each TF to have its own DNase binning scheme. We ran MILLIPEDE (with M5 binning) and BinDNase on DNase or ATAC data in each TF–cell type combination under default parameter settings. As with CENTIPEDE and msCentipede, we used the log odds of TF binding probabilities as predicted quantitative occupancy.

For all the comparisons shown in Figures 2 and 3, we trained models using data from the odd chromosomes and evaluated the prediction performance using data from the even chromosomes as the test set. We used Pearson's correlations between predicted and measured TF occupancy (asinh transformed) to evaluate prediction performance. In the binary (bound vs. unbound) context, we evaluated the prediction results using binary ChIP labels and computed metrics of area under ROC curve (AUROC) and area under precision recall curve (AUPR) (Supplemental Fig. S7).

We did not include PIQ (Sherwood et al. 2014) in our comparison, because msCentipede has already been shown to significantly outperform PIQ when it has access to DNase replicates (Raj et al. 2015). GERV (Zeng et al. 2016) is a statistical method that learns a *k*-mer-based model to predict TF binding using ChIP-seq and DNase-seq data and scores genetic variants by quantifying the changes of predicted ChIP-seq reads between the reference and alternative allele. Like TOP, it tries to predict quantitative TF occupancy, but its main goal is to score genetic variants that affect TF binding, and it treats DNase signals as a binary feature (open vs. closed), which would not be effective in capturing quantitative changes in DNase signals across dynamic conditions. Also, as a *k*-mer-based method, it does not adopt the site-centric framework that we and the other methods do. For these reasons, we did not include GERV in our comparison. We focused our attention on methods within the site-centric framework that use only chromatin accessibility data (and DNA sequence information). Thus, we did not include methods from the ENCODE DREAM Challenge, as they use additional training features including gene expression (RNA-seq) and in vitro DNA shape parameters, and predict binary TF binding events based on ChIP-seq peaks.

Differential occupancy comparison across cell types

We used the edgeR package (Robinson et al. 2010) to identify sites with significantly differential occupancy across cell types. For each TF at each candidate binding site, we tested the cell type effect by contrasting the predicted occupancy in each cell type (using DNase replicate samples) against the cell type mean. Sites with predicted occupancy of less than one read per million across the cell types were filtered out from the test, and then sites with a significant cell type effect (FDR < 10%) were selected.

When comparing predicted occupancy across cell types, potential influences from copy number variation (CNV) could lead to false positives. However, because our method predicts TF occupancy using DNase data and because CNV affects both DNase-seq and ChIP-seq counts in a consistent manner (CNV would lead to higher occupancy in both measured and predicted ChIP-seq reads in a higher copy number region), our predictions should still agree with measured occupancy. To deal with CNV influences while comparing across cell types, instead of directly correcting CNV on both DNase-seq and ChIP-seq data within the regression model, it is easier to do CNV adjustment as a postprocessing procedure on the predicted occupancy using input ChIP-seq data. However, because not all these cell types have input ChIP-seq data available, we did not perform CNV corrections in this study (input correction could be performed in those cell types for which input ChIP-seq data are available).

DNase-seq data across hormone treatment conditions

DNase-seq data from LNCaP cells exposed to androgen were collected in our laboratories. Data from before induction (time point 0) and after 12 h were previously published (Tewari et al. 2012) and are available from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) repository under accession GSE34780. DNase-seq data from the 45-min and 4-h treatments, along with more samples from before induction, were generated for this study (see Data access).

LNCaP cells were obtained from ATCC. Cells were maintained using the protocol described at http://genome.ucsc.edu/ENCODE/protocols/cell/human/LNCaP_Crawford_protocol.pdf. Before stimulation with either androgen (R1881, methyltrienolone) or vehicle (ethanol) for varying time durations, cells were grown in RPMI-1640 medium with 10% char-coal:dextran-stripped medium for 60 h. Androgen was added to the culture medium for a final concentration of 1 nM in all experiments. Isolation of total DNA,

cleavage with DNase I (henceforth, DNase), and subsequent preparation of sequencing libraries were performed as previously described (Song and Crawford 2010). Replicates from 12 h of androgen exposure were previously sequenced on the Illumina GAIIx platform, whereas replicates from the 45-min and 4-h time points were sequenced for this study on the Illumina HiSeq 2000 platform. Sequenced reads were aligned to the genome and further processed as previously described (Boyle et al. 2008; Tewari et al. 2012; Yardımcı et al. 2014).

DNase-seq data from A549 cells exposed to the GC hormone dexamethasone were collected in our laboratories. Detailed methods are provided in our paper (McDowell et al. 2018).

DNase-seq data from Ishikawa and T-47D cells before and after estrogen exposure were collected by others and previously published (Gertz et al. 2013); we downloaded their published data.

Differential occupancy comparison across hormone treatment conditions

In the androgen treatment analysis, we combined DNase-seq data from an earlier study (Tewari et al. 2012) with three replicates of uninduced samples and two replicates of 12-h androgen-induced samples (using the Illumina GAIIx sequencing platform), and DNase-seq data generated in this study with two replicates of uninduced samples, two replicates of 45-min induced samples, and two replicates of 4-h induced samples (using the Illumina HiSeq 2000 sequencing platform). AR ChIP-seq data collected in an earlier study with 4-h androgen induction in LNCaP cells (Massie et al. 2011) matched with our DNase-seq data of 4-h androgen induction were included in the training data set for AR in the hierarchical model. To screen for TFs showing dynamic occupancy changes, we predicted genome-wide TF occupancy at approximately 1500 JASPAR motifs. For each motif, we used edgeR to test linear, quadratic, and cubic trends of TF occupancy changes over the time course of uninduced and 45-min, 4-h, and 12-h induced conditions, adjusting for the batch effect from different sequencing platforms (GAIIx vs. HiSeq sequencing). Sites with a predicted occupancy of less than one read per million across the conditions were filtered out from the test, and then sites with significant linear, quadratic, or cubic trend of TF occupancy over the time course (FDR < 10%) were selected. Very few sites were found to have a significant quadratic or cubic trend, so we focused on sites with a significant linear trend.

In the estrogen treatment analysis, we used previously published DNase-seq and ChIP-seq data generated in Ishikawa (endometrial cancer cell line; previously mislabeled as ECC-1) and T-47D (breast cancer cell line) cells before and after estrogen induction (Gertz et al. 2013). ER (ESR1) ChIP-seq data from estrogen induced conditions were matched with the corresponding DNase data and included in the training data set in the hierarchical model. Occupancy predictions were made for each TF using its middle-level parameters in DNase-seq replicate samples in Ishikawa and T-47D, before and after estrogen stimulation. For each TF, we used edgeR to test for differential occupancy, where we considered both cell type effect (Ishikawa vs. T-47D) and treatment effect (estrogen induced vs. uninduced). Sites with a predicted occupancy of less than one read per million across the conditions were filtered out from the test, and then sites with treatment effect significantly higher or lower than zero (FDR < 10%) were selected.

In the GC treatment analysis, we used DNase-seq data collected in our laboratories from A549 cells (human alveolar adenocarcinoma cell line) over 12 time points from 0 to 12 h following exposure to the GC hormone dexamethasone (McDowell et al. 2018). For each TF, we used edgeR to test linear, quadratic, and cubic trends of TF occupancy changes over the 12 time points of GC treatment. Sites with a predicted occupancy of less than one read per million across the conditions were filtered out from the test, and then sites with significant linear, quadratic, or cubic trend of TF occupancy over the time course (FDR < 10%) were selected. Very few sites were found to have a significant quadratic or cubic trend, so we focused on sites with a significant linear trend.

After selecting sites with significant differential occupancy, we ranked TF motifs based on the percentage of sites showing significantly increased or decreased occupancy in response to treatment. Similar motifs were grouped together using RSAT (Castro-Mondragon et al. 2017) to simplify downstream interpretation and visualization.

topQTL mapping

We predicted genome-wide TF occupancy for about 1500 JASPAR motifs using previously published genotype information and DNase data generated from LCLs from 70 individuals (Degner et al. 2012). For each motif, we focused on those motif matches that had a SNP inside (we scanned for motif matches with *P*-value < 10^{-4} in this analysis to include more candidate genomic locations). When making predictions across the 70 LCLs using both PWM scores and DNase data, we fixed the PWM scores for candidate sites to be the average of the PWM scores calculated from the two homozygous genotypes for that SNP, in order to avoid using PWM scores twice: in both occupancy predictions and QTL association testing. We mapped topQTLs by testing the associations between genotypes and predicted TF occupancy across the 70 individuals using a linear model with R package MatrixEQTL (Shabalin 2012). For each TF motif, we selected the top 10% of candidate sites with the highest predicted occupancy for QTL mapping and downstream analysis (we tested the top 10%, 20%, …, 100% sites, and found the top 10% sites tended to maximize the number of QTLs detected after multiple testing correction). To facilitate comparison with dsQTLs, we followed the same data processing procedures as described by Degner et al. (2012), including z-score standardization, GC content correction, quantile normalization, and regressing out four principal components (PCs). We mapped *cis*-topQTLs by testing SNPs within motif matches. For each TF motif, genetic variants with significant associations to predicted TF occupancy (FDR < 10%) were identified as topQTLs for that motif and were the basis of all subsequent analysis in Figure 6. Similar motifs were grouped together using RSAT (Castro-Mondragon et al. 2017) to simplify downstream interpretation and visualization.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE157473. TOP is implemented in R, and is available at GitHub (https://github.com/HarteminkLab/TOP) and as Supplemental Code. Precomputed genome-wide quantitative TF occupancy, pretrained TOP model parameters, and links to other code resources are available via http://users.cs.duke.edu/~amink/software/.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank David MacAlpine, Raluca Gordân, Galip Gürkan Yardımcı, Jason Belsky, Ian McDowell, Chris Vockley, and Tony

D'Ippolito for helpful comments during the development of this work or in response to drafts of the manuscript. This work was funded in part by National Institutes of Health (NIH) grants U01-HG007900, R01-GM118551, and R35-GM141795.

Author contributions: K.L., G.E.C., and A.J.H. designed the study. K.L., J.Z., L.M., G.E.C., and A.J.H. conducted or supervised analyses. A.S., L.K.H., A.K.T., L.S., T.E.R., and G.E.C. conducted or supervised experiments. K.L. and A.J.H. wrote the paper, with assistance from J.Z.

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073. doi:10.1038/nature09534

Bar-Joseph Z, Gifford DK, Jaakkola TS. 2001. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17**(Suppl 1)**:** S22–S29. doi:10.1093/bioinformatics/17.suppl_1.S22

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24:** 2537–2538. doi:10.1093/bioinformatics/btn480

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 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. doi:10.1038/nmeth.2688

Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M, van Helden J. 2017. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res* **45:** e119. doi:10.1093/nar/gkx314

Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL. 2012. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28:** 56–62. doi:10.1093/bioinformatics/btr614

Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482:** 390–394. doi:10.1038/nature10808

Deplancke B, Alpern D, Gardeux V. 2016. The genetics of transcription factor DNA binding variation. *Cell* **166:** 538–554. doi:10.1016/j.cell.2016.07.012

Durham TJ, Libbrecht MW, Howbert JJ, Bilmes J, Noble WS. 2018. PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition. *Nat Commun* **9:** 1402. doi:10.1038/s41467-018-03635-9

Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33:** 364–376. doi:10.1038/nbt.3157

Gertz J, Siggia ED, Cohen BA. 2009. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature* **457:** 215–218. doi:10.1038/nature07521

Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM, et al. 2013. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* **52:** 25–36. doi:10.1016/j.molcel.2013.08.037

Goldstein I, Baek S, Presman DM, Paakinaho V, Swinstead EE, Hager GL. 2017. Transcription factor assisted loading and enhancer dynamics dictate the hepatic fasting response. *Genome Res* **27:** 427–439. doi:10.1101/gr.212175.116

Gordân R, Hartemink AJ. 2008. Using DNA duplex stability information for transcription factor binding site discovery. *Pac Symp Biocomput* **2008:** 453–464.

Gordân R, Hartemink AJ, Bulyk ML. 2009. Distinguishing direct versus indirect transcription factor–DNA interactions. *Genome Res* **19:** 2090–2100. doi:10.1101/gr.094144.109

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018. doi:10.1093/bioinformatics/btr064

Haar A. 1910. Zur theorie der orthogonalen funktionensysteme. *Math Ann* **69:** 331–371. doi:10.1007/BF01456326

He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. 2012. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res* **22:** 1015–1025. doi:10.1101/gr.133280.111

He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, et al. 2014. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* **11:** 73–78. doi:10.1038/nmeth.2762

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* **106:** 9362–9367. doi:10.1073/pnas.0903103106

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9:** 473–476. doi:10.1038/nmeth.1937

Kähärä J, Lähdesmäki H. 2015. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics* **31:** 2852–2859. doi:10.1093/bioinformatics/btv294

Keilwagen J, Posch S, Grau J. 2019. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol* **20:** 505. doi:10.1186/s13059-018-1614-y

Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20:** 207–220. doi:10.1038/s41576-018-0089-8

Li X-Y, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD. 2011. The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol* **12:** R34. doi:10.1186/gb-2011-12-4-r34

Li H, Quang D, Guan Y. 2019a. Anchor: *trans*-cell type prediction of transcription factor binding sites. *Genome Res* **29:** 281–292. doi:10.1101/gr.237156.118

Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. 2019b. Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* **20:** 45. doi:10.1186/s13059-019-1642-2

Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD. 2012. Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **484:** 251–255. doi:10.1038/nature10985

Luo K, Hartemink AJ. 2013. Using DNase digestion data to accurately identify transcription factor binding sites. *Pac Symp Biocomput* **2013:** 80–91. doi:10.1142/9789814447973_0009

Lupiáñez DG, Spielmann M, Mundlos S. 2016. Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet* **32:** 225–237. doi:10.1016/j.tig.2016.01.003

Martins AL, Walavalkar NM, Anderson WD, Zang C, Guertin MJ. 2018. Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Res* **46:** e9. doi:10.1093/nar/gkx1053

Massie CE, Lynch A, Ramos-Montoya A, Boren J, Stark R, Fazli L, Warren A, Scott H, Madhu B, Sharma N, et al. 2011. The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis. *EMBO J* **30:** 2719–2733. doi:10.1038/emboj.2011.158

McDaniell R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328:** 235–239. doi:10.1126/science.1184655

McDowell IC, Barrera A, D'Ippolito AM, Vockley CM, Hong LK, Leichter SM, Bartelt LC, Majoros WH, Song L, Safi A, et al. 2018. Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Res* **28:** 1272–1284. doi:10.1101/gr.233346.117

Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, Vonrhein C, Moras D, Romier C, Bolognesi M, et al. 2013. Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell* **152:** 132–143. doi:10.1016/j.cell.2012.11.047

Narlikar L, Gordân R, Hartemink AJ. 2007. Nucleosome occupancy information improves *de novo* motif discovery. In *Research in Computational Molecular Biology (RECOMB 2007)* (ed. Speed T, Huang H), Vol. **4453** of *Lecture Notes in Computer Science*, pp. 107–121. Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-71681-5_8

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489:** 83–90. doi:10.1038/nature11212

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21:** 447–455. doi:10.1101/gr.112623.110

Quang D, Xie X. 2019. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* **166:** 40–47. doi:10.1016/j.ymeth.2019.03.020

Raj A, Shim H, Gilad Y, Pritchard JK, Stephens M. 2015. msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLoS One* **10:** e0138030. doi:10.1371/journal.pone.0138030

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680. doi:10.1016/j.cell.2014.11.021

R Core Team. 2020. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Reddy TE, Gertz J, Crawford GE, Garabedian MJ, Myers RM. 2012. The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. *Mol Cell Biol* **32:** 3756–3767. doi:10.1128/MCB.00062-12

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Schreiber J, Bilmes J, Noble WS. 2020a. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biol* **21:** 82. doi:10.1186/s13059-020-01978-5

Schreiber J, Durham T, Bilmes J, Noble WS. 2020b. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol* **21:** 81. doi:10.1186/s13059-020-01977-6

Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE. 2016. Bayes and big data: the consensus Monte Carlo algorithm. *Intl J Manage Sci Engin Manage* **11:** 78–88. doi:10.1080/17509653.2016.1142191

Segal E, Widom J. 2009. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* **10:** 443–456. doi:10.1038/nrg2591

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28:** 1353–1358. doi:10.1093/bioinformatics/bts163

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* **23:** 777–788. doi:10.1101/gr.152140.112

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32:** 171–178. doi:10.1038/nbt.2798

Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010:** pdb.prot5384. doi:10.1101/pdb.prot5384

Sung M-H, Guertin MJ, Baek S, Hager GL. 2014. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell* **56:** 275–285. doi:10.1016/j.molcel.2014.08.016

Tehranchi AK, Myrthil M, Martin T, Hie BL, Golan D, Fraser HB. 2016. Pooled ChIP-seq links variation in transcription factor binding to complex disease risk. *Cell* **165:** 730–741. doi:10.1016/j.cell.2016.03.041

Tewari AK, Yardimci GG, Shibata Y, Sheffield NC, Song L, Taylor BS, Georgiev SG, Coetzee GA, Ohler U, Furey TS, et al. 2012. Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome Biol* **13:** R88. doi:10.1186/gb-2012-13-10-r88

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489:** 75–82. doi:10.1038/nature11232

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol* **58:** 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Vockley CM, D'Ippolito AM, McDowell IC, Majoros WH, Safi A, Song L, Crawford GE, Reddy TE. 2016. Direct GR binding sites potentiate clusters of TF binding across the human genome. *Cell* **166:** 1269–1281.e19. doi:10.1016/j.cell.2016.07.049

Voss TC, Schiltz RL, Sung M-H, Yen PM, Stamatoyannopoulos JA, Biddie SC, Johnson TA, Miranda TB, John S, Hager GL, et al. 2011. Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell* **146:** 544–554. doi:10.1016/j.cell.2011.07.006

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22:** 1798–1812. doi:10.1101/gr.139105.112

Wasson T, Hartemink AJ. 2009. An ensemble model of competitive multifactor binding of the genome. *Genome Res* **19:** 2101–2112. doi:10.1101/gr.093450.109

Yardımcı GG, Frank CL, Crawford GE, Ohler U. 2014. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res* **42:** 11865–11878. doi:10.1093/nar/gku810

Zeng H, Hashimoto T, Kang DD, Gifford DK. 2016. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* **32:** 490–496. doi:10.1093/bioinformatics/btv565

Zhong J, Wasson T, Hartemink AJ. 2014. Learning protein–DNA interaction landscapes by integrating experimental data through computational models. *Bioinformatics* **30:** 2868–2874. doi:10.1093/bioinformatics/btu408

Zhong J, Luo K, Winter PS, Crawford GE, Iversen ES, Hartemink AJ. 2016. Mapping nucleosome positions using DNase-seq. *Genome Res* **26:** 351–364. doi:10.1101/gr.195602.115

Received September 30, 2020; accepted in revised form May 6, 2022.