Nucleic Acids Research, 2021 **1**
https://doi.org/10.1093/nar/gkab553

RoboCOP: jointly computing chromatin occupancy profiles for numerous factors from chromatin accessibility data

Sneha Mitra [1], Jianling Zhong [2], Trung Q. Tran [1], David M. MacAlpine [2,3,4] and Alexander J. Hartemink [1,2,4,*]

[1]Department of Computer Science, Duke University, Durham, NC 27708, USA, [2]Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708, USA, [3]Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC 27710, USA and [4]Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, USA

Received December 23, 2020; Revised May 28, 2021; Editorial Decision June 08, 2021; Accepted July 08, 2021

ABSTRACT

Chromatin is a tightly packaged structure of DNA and protein within the nucleus of a cell. The arrangement of different protein complexes along the DNA modulates and is modulated by gene expression. Measuring the binding locations and occupancy levels of different transcription factors (TFs) and nucleosomes is therefore crucial to understanding gene regulation. Antibody-based methods for assaying chromatin occupancy are capable of identifying the binding sites of specific DNA binding factors, but only one factor at a time. In contrast, epigenomic accessibility data like MNase-seq, DNase-seq, and ATAC-seq provide insight into the chromatin landscape of all factors bound along the genome, but with little insight into the identities of those factors. Here, we present RoboCOP, a multivariate state space model that integrates chromatin accessibility data with nucleotide sequence to jointly compute genome-wide probabilistic scores of nucleosome and TF occupancy, for hundreds of different factors. We apply RoboCOP to MNase-seq and ATAC-seq data to elucidate the protein-binding landscape of nucleosomes and 150 TFs across the yeast genome, and show that our model makes better predictions than existing methods. We also compute a chromatin occupancy profile of the yeast genome under cadmium stress, revealing chromatin dynamics associated with transcriptional regulation.

INTRODUCTION

The chromatin of a cell consists of the genome and all the proteins and protein complexes arrayed along it. The ensemble of potential arrangements of proteins along the genome is combinatorially vast, but the specific configuration of the chromatin within each cell determines whether and to what extent its various genes are expressed. Therefore, deciphering the chromatin landscape—which proteins are bound at every position in the genome—is crucial to developing a more mechanistic and predictive understanding of gene regulation.

Two important types of DNA binding factors (DBFs) are transcription factors (TFs) and nucleosomes. TFs are gene regulatory proteins that activate or repress the transcription of genes by binding with specific sequence preferences to sites along the DNA. Nucleosomes form when 147 base pairs of DNA are wrapped around an octamer of histone proteins. They have lower sequence specificity than TFs, but still exhibit a preference for a periodic arrangement of dinucleotides that facilitates DNA wrapping (1). Likened to beads on a string, nucleosomes are positioned fairly regularly along the DNA, occupying about 81% of the genome in the case of the yeast *Saccharomyces cerevisiae* (2). In taking up their respective positions, nucleosomes contribute to the regulation of gene expression in part by allowing or blocking TFs from occupying their putative binding sites. Useful models of the chromatin landscape must therefore be able to simultaneously represent and reason about many DBFs at once, and must explicitly account for the way they compete with one another to bind the genome.

The binding locations of DBFs have been assayed extensively at high resolution with antibody-based methods (3–5). However, these methods are limited to assaying only one particular factor at a time, and require a separate antibody for each factor. Consequently, using this approach to identify the binding locations of myriad different DBFs is extremely expensive and laborious, especially if we are interested in studying how the chromatin landscape changes dynamically across time or in response to changing environ-

*To whom correspondence should be addressed. Tel: +1 919 660 6514; Email: amink@cs.duke.edu

mental conditions. In such scenarios, antibody-based methods are often used to assay a small number of important histone modifications. Computational algorithms are often employed to integrate these datasets and segment the genome into broad 'epigenomic states' that may be associated with larger regulatory loci like promoters and enhancers (6–9).

In contrast to antibody-based methods, chromatin accessibility assays probe unoccupied, or open, regions of the chromatin, thereby providing indirect information about the genomic regions occupied by bound proteins. Chromatin accessibility data can be generated in a few different ways, including enzymatic cleavage (DNase-seq), enzymatic digestion (MNase-seq), or transposon insertion (ATAC-seq). Here, we consider the latter two. In MNase-seq, the endo-exonuclease MNase is used to digest unbound DNA, leaving behind fragments of bound DNA. In ATAC-seq, the transposase Tn5 is loaded with adapters and used to cleave and tag unbound DNA, yielding fragments of DNA whose ends were unbound. In either case, paired-end sequencing of the resulting fragments reveals not only their location but also their length, providing information about the length of protein-bound sites throughout the genome. MNase-seq and ATAC-seq have primarily been used to study nucleosome positioning (2,10–14) and identify accessible regulatory regions (15), respectively. However, recent experiments have demonstrated that MNase-seq can be used to identify some TFs (16–19) and ATAC-seq to study some nucleosome positions (20). We set out to explore whether a sufficiently sophisticated computational model might be able to leverage these data to produce a chromatin occupancy profile at every position in the genome, revealing the precise binding locations for numerous different DBFs at once.

In earlier work, we developed COMPETE to compute a probabilistic landscape of DBF occupancy along the genome (21). COMPETE considers DBFs binding to the genome from the perspective of a thermodynamic ensemble, where the DBFs are in continual competition to occupy locations along the genome and their chances of binding are affected by their concentrations, akin to a repeated game of 'musical chairs'. COMPETE output depends only on genome sequence (which is static) and DBF concentrations (which may be dynamic); it does not utilize experimental data, so its predictions of the chromatin landscape are entirely theoretical. We later developed a modified version of COMPETE to estimate DBF concentrations by maximizing the correlation between the output of COMPETE's theoretical model and an MNase-seq signal, improving the reported binding landscape (22). However, this modified version still does not incorporate chromatin accessibility data directly into the underlying probabilistic model.

Here, we present RoboCOP (**robo**tic **c**hromatin **o**ccupancy **p**rofiler), a new method that takes nucleotide sequence and chromatin accessibility data as input and then uses a multivariate hidden Markov model (HMM) (23) to compute a probabilistic occupancy landscape of hundreds of DBFs genome-wide at single-nucleotide resolution. We demonstrate that RoboCOP can use chromatin accessibility data generated from paired-end MNase-seq and/or ATAC-seq experiments to compute a chromatin occupancy profile of 150 TFs and nucleosomes across the *Saccharomyces cerevisiae* genome. We validate its nucleosome positioning predictions using high-precision annotations resulting from a chemical cleavage method (24), and its TF binding site predictions using annotations reported by ChIP-chip (25), ChIP-exo (5), and ORGANIC (26) experiments. Because ATAC-seq fragments occur primarily in regions of open chromatin, the chromatin occupancy profiles learned by RoboCOP from ATAC-seq data alone are generally informative only in those regions. In contrast, MNase-seq fragments are distributed more evenly across the genome, and as a result, can be used to generate chromatin occupancy profiles genome-wide. RoboCOP is the first method for using MNase-seq data to elucidate the chromatin landscape of the entire genome, and can be used to study how chromatin responds dynamically to changing environmental conditions, as we demonstrate by revealing the genome-wide chromatin landscape of yeast under cadmium stress.

MATERIALS AND METHODS

MNase-seq fragments of different lengths provide information about different kinds of DNA binding factors

Our high-resolution paired-end MNase-seq data can be plotted in two dimensions by representing every fragment as a point, whose x-coordinate is the genomic location of its midpoint and whose y-coordinate is its length, thereby capturing both the fragment length and location distributions at single base pair resolution. As can be seen from the region in Figure 1A, gene bodies mostly contain fragments about 150 bases long, corresponding to nucleosomes. Promoter regions contain shorter fragments, often associated with TF binding sites. Each of the two promoter regions in Figure 1A has an annotated Abf1 binding site (25) that can explain the enrichment of short fragments nearby.

Since the degree of MNase digestion can influence the fragment length distribution (27,28), we plotted the MNase-seq fragments around transcription start sites (TSSs) to get an estimate of the length of the fragments corresponding to nucleosomes and TF binding sites (Figure 1C). We find that fragments of length 157 have the highest frequency (left panel of Figure 1C). Given that nucleosomes are about this size and occupy about 81% of the yeast genome (2), fragments of this length generally correspond to nucleosomes. We denote all fragments whose length is 157 ± 30 to be nucleosomal fragments, or `nucFrags` for short. The midpoints of `nucFrags` are depicted in red dots in Figure 1A. As expected, these `nucFrags` occur in tandem arrays within gene bodies but are generally absent from promoters (Figure 1A,C). Fragments are particularly concentrated at the +1 nucleosome position in Figure 1C, just downstream of the TSS, because the +1 nucleosome is usually well-positioned. Furthermore, the marginal density of the midpoints of these fragments around annotated nucleosome dyads (24) peaks precisely at the dyad (the central nucleotide position of the nucleosome, through which is an axis of symmetry for the nucleosome in 3D), with counts dropping nearly symmetrically in either direction (Figure 1D). This makes sense because MNase digests linker regions, leaving behind undigested DNA fragments wrapped around histone octamers. So the midpoint

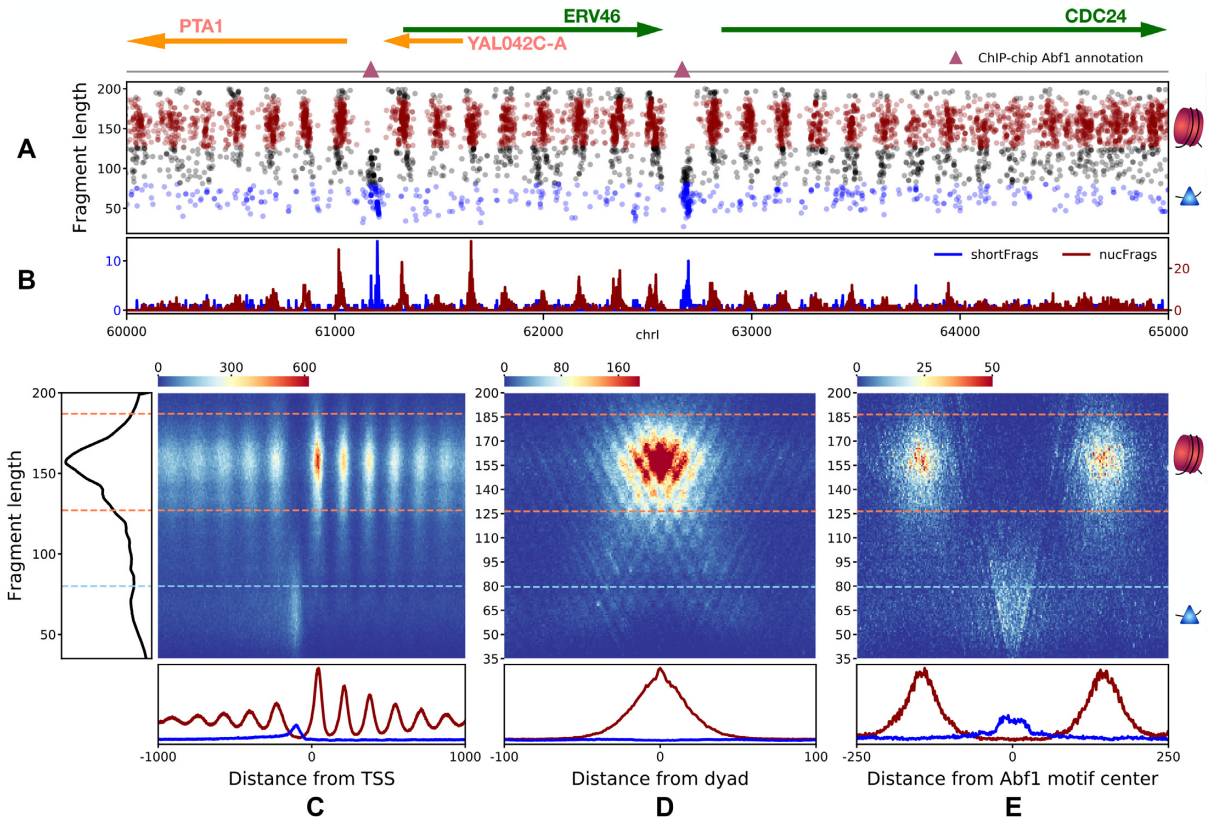

Figure 1. Paired-end MNase-seq data is informative about the binding of both nucleosomes and smaller DBFs, such as transcription factors. (**A**) Two-dimensional plot of MNase-seq fragments from positions 60,000 to 65,000 of yeast chromosome I. Each fragment is plotted based on its length (y-axis) and the genomic location of its midpoint (x-axis). Nucleosome-sized fragments (`nucFrags`, length 157±30) are colored red, while shorter fragments corresponding to smaller proteins (`shortFrags`, length ≤ 80) are colored blue. Above the plot are genomic annotations for this region, with Watson strand genes depicted as green arrows and Crick strand genes as orange. Below the gene annotations, known TF binding sites (25) are indicated using triangles. This region contains two annotated binding sites for Abf1 (pink). (**B**) Aggregate numbers of red and blue dots at each genomic position in (A), resulting in the one-dimensional `nucFrags` and `shortFrags` signals, respectively. (**C**) Composite heatmap of MNase-seq fragments around all yeast genes, centered on each gene's TSS. Panels along the side and bottom show marginal densities. The side panel shows that `nucFrags` predominate, consistent with the fact that over 80% of the yeast genome is occupied by nucleosomes (2), but the bottom panel clarifies that `nucFrags` and `shortFrags` are positioned differently with respect to genes. `nucFrags` appear in tandem arrays within gene bodies, with particularly strong enrichment (deep red) at +1 nucleosomes just downstream of the TSS. In contrast, `shortFrags` are enriched in the nucleosome-free promoter region just upstream of the TSS. (**D**) Composite heatmap of MNase-seq fragments around the 2,000 most well-positioned nucleosomes in the yeast genome (24), centered on each nucleosome's dyad. The `nucFrags` signal peaks precisely at the dyad and decreases symmetrically in either direction. (**E**) Composite heatmap of MNase-seq fragments around all annotated Abf1 binding sites (25) in the yeast genome, centered on each site's motif. Note the clear enrichment of `shortFrags` near Abf1 sites.

counts of these `nucFrags` would be highest at the annotated dyads and decrease on moving away from the dyad.

In addition, it has been shown that shorter fragments in MNase-seq provide information about TF binding sites (16). To verify that we see this signal in our data, both the composite plot in Figure 1C and the genomic region in Figure 1A reveal that promoter regions are enriched with shorter fragments. The promoter region is often bound by specific and general TFs that aid in the transcription of genes. To ensure that the MNase-seq signal in these promoter regions is not just noise, we plot the MNase-seq midpoints around a set of annotated TF binding sites (Figure 1E). We choose the well-studied TF, Abf1, because it has multiple annotated binding sites across the genome. On plotting the MNase-seq midpoint counts around these annotated binding sites, we notice a clear enrichment of short fragments at the binding sites. We denote these short fragments of length less than 80 as `shortFrags`. The midpoints of `shortFrags` are plotted as blue dots in Figure 1A. Unlike the midpoint counts of the `nucFrags` which have a symmetrically decreasing shape around the nucleosome dyads (Figure 1D), the midpoint counts of `shortFrags` are more uniformly distributed within the binding site (Figure 1E). The `shortFrags` signal at the Abf1 binding sites is noisier than the MNase signal associated with nucleosomes. One reason for this increased noise is that fragments protected from digestion by bound TFs may be quite small, and the smallest fragments (of length less than 27 in our case) are not even present in the dataset due to sequencing and alignment limitations.

We ignore fragments of intermediate length (81–126) in our analysis, though these could provide information about other kinds of complexes along the genome, like hexasomes (29). Such factors would also be important for a complete understanding of the chromatin landscape, but we limit our analysis here to studying the occupancy of nucleosomes and

TFs. For the subsequent sections of this paper, we only consider the midpoint counts of `nucFrags` and `shortFrags` as depicted in red and blue respectively in Figure 1A. We further simplify the two-dimensional plot in Figure 1A to form two one-dimensional signals by separately aggregating the midpoint counts of `nucFrags` and `shortFrags`, as shown in Figure 1B.

ATAC-seq fragments provide similar information near regions of open chromatin

To understand whether paired-end ATAC-seq provides similar information, we studied the midpoint distribution of ATAC-seq fragments (20) in the same genomic region (Supplementary Figure S1A). Because the Tn5 insertion into the genome is offset by 4 bp (15), we decreased the length of all ATAC-seq fragments by 8 bp in our analyses. Owing to the nature of the ATAC-seq protocol, most fragments map within and adjacent to nucleosome-depleted promoter regions; the signal becomes quite sparse away from these regions (Supplementary Figure S1B). We confirm this by examining the aggregate ATAC-seq signal at all yeast genes and find that even the nucleosomal signal weakens beyond the +1 nucleosome within the gene body (Supplementary Figure S1C). Aggregate ATAC-seq signal around annotated nucleosome positions (Supplementary Figure S1D) looks similar in character to MNase-seq; the signal is strongest when we consider fragments of length 135–200, so we use this range for `nucFrags` in ATAC-seq data. Aggregate ATAC-seq signal around annotated Abf1 binding sites exhibits a very strong `shortFrags` signal (Supplementary Figure S1E), although the `shortFrags` peak with ATAC-seq is a bit broader than with MNase-seq (Supplementary Figure S2), likely because the exonuclease activity of MNase allows it to digest unoccupied DNA beyond its site of initial cleavage.

RoboCOP model structure and transition probabilities

RoboCOP is a multivariate hidden Markov model (HMM) for jointly computing genome-wide chromatin occupancy profiles using nucleotide sequence and chromatin accessibility data as observables (Figure 2). The HMM structure has been adapted from (21). Let the number of TFs be K. Let $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K$ denote the models for the K TFs, and let $\boldsymbol{\pi}_{K+1}$ denote the model for nucleosomes. To simplify notation, we consider an unbound DNA nucleotide to be occupied by a special 'empty' DBF (22); suggestively, let $\boldsymbol{\pi}_0$ denote this model. Therefore, we have $K + 2$ DBF models in total, and we use a central non-emitting ('silent') state to simplify transitions among them. The HMM may transition from this central silent state to any one of the $K + 2$ DBF models; at the end of each DBF model, the HMM always transitions back to the central silent state (Figure 2B, Supplementary Figure S3). This approach assumes DBFs bind independently of their neighbors, and each DBF therefore has just a single transition probability associated with it. The transition probabilities from the central state to the various DBFs are denoted as $\{\alpha_0, \ldots, \alpha_{K+1}\}$.

Each genome coordinate is represented by one hidden state in the HMM. An unbound DNA nucleotide is length one, so its model $\boldsymbol{\pi}_0$ has just a single hidden state. The other DBFs (nucleosomes and TFs) have binding sites of greater length and are thus modeled using collections of multiple hidden states. For TF k with a binding site of length L_k, the HMM either transitions through L_k hidden states of its binding motif or L_k hidden states of the reverse complement of its binding motif. An additional non-emitting state is added as the first hidden state of the TF model $\boldsymbol{\pi}_k$, allowing the HMM to transition through the forward or reverse complement of the motif with equal probability (Supplementary Figure S4A). The complete TF model $\boldsymbol{\pi}_k$ therefore has a total of $2L_k + 1$ hidden states. Once the HMM enters the hidden states for either the forward or reverse motif, it transitions through the sequence of hidden states with probability one between consecutive hidden states. On reaching the final hidden state of either motif, the HMM transitions back to the central silent state with probability one. Likewise, once the HMM enters the nucleosome model $\boldsymbol{\pi}_{K+1}$, it transitions through a sequence of hidden states corresponding to 147 nucleotides, after which it transitions back to the central silent state (Supplementary Figure S4B). The nucleosome model differs from the TF models in that the latter are modeled with simple PWM motifs, while the former is implemented using a dinucleotide sequence specificity model.

Suppose the sequence of hidden states for the entire genome of length G is denoted as $z_1, \ldots, z_G$. Then the transition probabilities satisfy the following:

- $P(z_{g+1} = \pi_{k,l+1} \mid z_g = \pi_{k,l}) = 1$ whenever $l < L_k$. Within a DBF, the HMM only transitions to that DBF's next state and not any other state, until it reaches the end of the DBF.
- $P(z_{g+1} = \pi_{k_1,1} \mid z_g = \pi_{k_2,L_{k_2}}) = \alpha_{k_1}$ for all k_1 and k_2. The transition probability to the first state of a DBF is a constant, independent of which DBF the HMM visited previously.
- $P(z_{g+1} \mid z_g) = 0$ for all other cases.

The HMM always starts in the central silent state with probability one; this guarantees that it cannot start in the middle of a DBF.

RoboCOP emission probabilities

The HMM employed by RoboCOP is multivariate, meaning that each hidden state is responsible for emitting multiple observables per position in the genome (Figure 2C). In our case, these observables are modeled as independent, conditioned on the hidden state, but adding dependence would be straightforward. The structure of RoboCOP is flexible enough to allow any combination of chromatin accessibility data to be used as input. In this paper, we apply it to combinations of paired-end MNase-seq and paired-end ATAC-seq data. So for a genome of length G, the sequences of observables being explained by the model are: (i) nucleotide sequence $\{s_1, \ldots, s_G\}$, (ii) midpoint counts of MNase-seq and/or ATAC-seq `nucFrags` $\{l_{1,j}, \ldots, l_{G,j}\}$, and (iii) midpoint counts of MNase-seq and/or ATAC-seq `shortFrags` $\{m_{1,j}, \ldots, m_{G,j}\}$. Here, j indexes over the chromatin accessibility datasets provided as input. So if

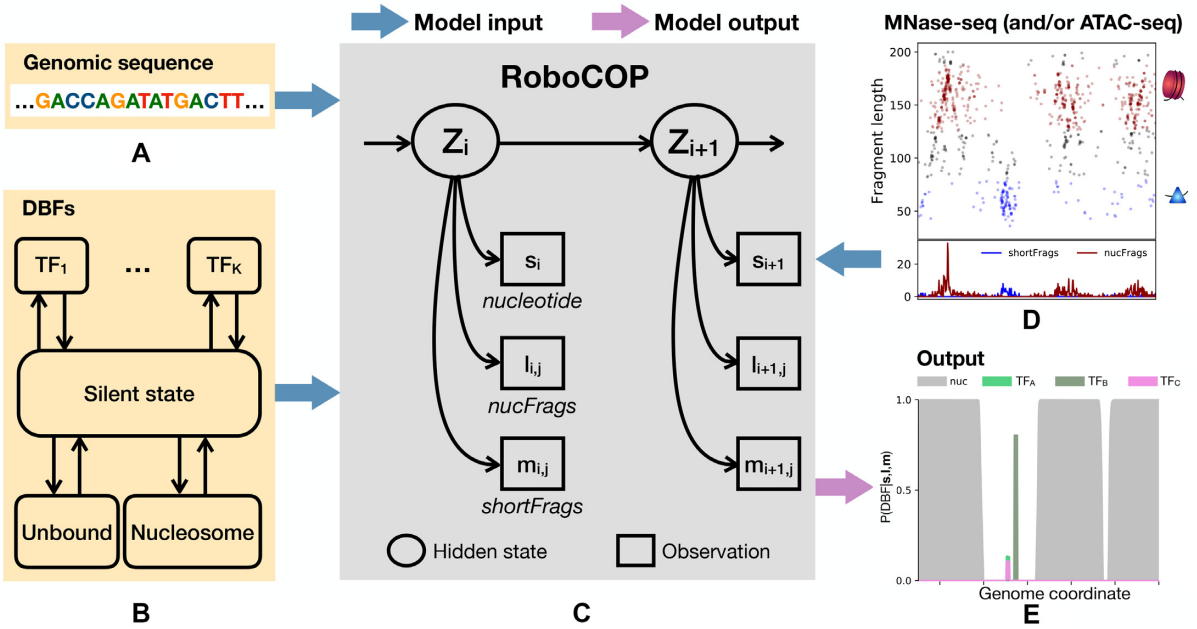

Figure 2. RoboCOP takes various inputs (blue arrows) and produces as output (pink arrow) a chromatin occupancy profile providing quantitative estimates of occupancy for the specified collection of DBFs. The underlying genomic sequence (**A**) and the collection of DBFs and their sequence specificity models (**B**) are provided as input to the RoboCOP model (**C**), along with the `nucFrags` and `shortFrags` signals that result from aggregation of MNase-seq/ATAC-seq fragment midpoint counts (**D**). (**B**) The state transition matrix for the HMM is simplified by the inclusion of a central, non-emitting silent state; from this state, the model can transition to any DBF, after which it necessarily transitions back to the central silent state, thereby removing dependencies among the DBFs. (**C**) RoboCOP is a multivariate HMM where the hidden state z_i at genomic position i emits a nucleotide (s_i), and for each chromatin accessibility dataset input, j, a `nucFrags` count ($l_{i,j}$), and a `shortFrags` count ($m_{i,j}$). (**E**) RoboCOP performs posterior decoding and yields the probability of each DBF at every position in the genome. The score on the y-axis is the probability of that location being bound by a given DBF.

RoboCOP is run with both MNase-seq and ATAC-seq, we would have two sets of `nucFrags` and two sets of `shortFrags` observed at each position in the genome.

For any position g in the genome, the hidden state z_g is thus responsible for emitting a nucleotide s_g, and for each of the chromatin accessibility datasets, j, a number $l_{g,j}$ of midpoints of `nucFrags`, and a number $m_{g,j}$ of midpoints of `shortFrags` (Figure 2C). Since these observations are independent of one another given the hidden state z_g, each hidden state has an emission model for each of the observables, and the joint probability of the multivariate emission is the product of the emission probabilities of all the observables.

For the hidden states corresponding to the TF models $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K$, emission probabilities for nucleotide sequences are represented using PWMs. For each of our 150 TFs, we use the PWM of its primary motif reported in (30) (except for Rap1, where we use the more detailed motifs in (5)). For the nucleosome model $\boldsymbol{\pi}_{K+1}$, the emission probability for a nucleotide sequence of length 147 can be represented using a position-specific dinucleotide model (31). To represent this dinucleotide model, the number of hidden states in $\boldsymbol{\pi}_{K+1}$ is roughly 4×147. We use the same dinucleotide model that was used earlier in COMPETE (21).

To simplify the model description in what follows, without loss of generality, let us suppose that the model is run with MNase-seq alone. As described earlier, the two-dimensional MNase-seq data are transformed into two one-dimensional signals (Figure 2D); the midpoint counts of `nucFrags` primarily influence the learned nucleosome positions and the midpoint counts of `shortFrags` primarily influence the learned TF binding sites. In both cases, a negative binomial (NB) distribution is used to model the emission probabilities. We use two sets of NB distributions to model the midpoint counts of `nucFrags`. One distribution, $NB(\mu_{nuc}, \phi_{nuc})$, explains the counts of `nucFrags` at the nucleosome positions and another distribution, $NB(\mu_{l_b}, \phi_{l_b})$, explains the counts of `nucFrags` elsewhere in the genome. Since the midpoint counts of `nucFrags` within a nucleosome are not uniform (Figure 1B), we model each of the 147 positions separately. To obtain $\boldsymbol{\mu}_{nuc}$ and $\boldsymbol{\phi}_{nuc}$, we collect the midpoint counts of `nucFrags` in a window of size 147 centered on the annotated nucleosome dyads of the 2,000 most well-positioned nucleosomes (24) and estimate 147 NB distributions using maximum likelihood estimation (MLE). The 147 estimated values of μ are denoted as $\boldsymbol{\mu}_{nuc}$. The mean of the 147 estimated values of ϕ is denoted as ϕ_{nuc} (shared across all 147 positions). Quantile-quantile plots show the resulting NB distributions to be a good fit (Supplementary Figure S5). As for $NB(\mu_{l_b}, \phi_{l_b})$, we use MLE to estimate its parameters from the midpoint counts of `nucFrags` within the linker regions on both sides of the same set of 2,000 nucleosomes. For this purpose, we considered linkers to be 15 bases long (32).

Similarly, we model the midpoint counts of `shortFrags` using two distributions where one of them, $NB(\mu_{TF}, \phi_{TF})$, explains the counts of `shortFrags` within TF binding sites, while the other, $NB(\mu_{m_b}, \phi_{m_b})$, explains counts elsewhere. To estimate the parameters of

$NB(\mu_{TF}, \phi_{TF})$, we collect the midpoint counts of `shortFrags` within annotated Abf1 and Reb1 binding sites (25) and fit the NB distribution using MLE. A quantile-quantile plot again shows the NB distribution provides a good fit (Supplementary Figure S6). We chose Abf1 and Reb1 for fitting the distribution because these TFs have many binding sites in the genome and the binding sites are often less noisy compared to those of other TFs. For parameterizing $NB(\mu_{m_b}, \phi_{m_b})$, we collect the midpoint counts of `shortFrags` within the same linker regions used earlier and estimate the NB distribution using MLE.

When we use multiple chromatin accessibility datasets as input, we fit separate negative binomial distributions for the `nucFrags` and `shortFrags` signals of each. We denote the model run with MNase-seq alone as RoboCOP$_{\text{MNase}}$, the model run with ATAC-seq alone as RoboCOP$_{\text{ATAC}}$, and the model run with both MNase-seq and ATAC-seq as RoboCOP$_{\text{MNase+ATAC}}$.

RoboCOP transition probability updates

Within each single DBF model, the transition probabilities between hidden states can only be zero or one (except for the two transition probabilities from each TF model's first, non-emitting state to the first state of either its forward or reverse motif; these are fixed at 0.5). Consequently, the only transition probabilities we need to learn are $\{\alpha_0, \ldots, \alpha_{K+1}\}$, those from the central silent state to the first state of each DBF (Supplementary Figure S3). Our approach is to initialize these to sensible values, and then optimize them using Baum-Welch, which is guaranteed to converge to a local maximum of the model's likelihood.

To initialize the transition probabilities $\{\alpha_0, \ldots, \alpha_{K+1}\}$, we first assign a non-negative concentration or 'weight' to each DBF. Let the weight for DBF i be denoted w_i. Following previous work (21,22), we assign weight $w_0 = 1$ to the 'empty' DBF (representing an unbound DNA nucleotide) and $w_{K+1} = 35$ to the nucleosome. To each TF $k \in \{1, \ldots, K\}$, we assign a weight w_k which is that TF's estimated dissociation constant K_d (or alternatively, a multiple thereof: $8K_d$, $16K_d$, $32K_d$, or $64K_d$), which can be calculated from its PWM (21,33).

To convert these weights into transition probabilities $\{\alpha_0, \ldots, \alpha_{K+1}\}$, we need to account for the fact that the weights are contributing to a Boltzmann distribution that is being normalized by a partition function. Because of the partition function, preserving an identical Boltzmann distribution while rescaling the weights requires that the weight for each DBF be rescaled by an amount that accounts for its length. Specifically, for any choice of positive constant c, the weight of DBF k must be rescaled by c^{L_k}, where L_k is its length, from as little as 1 for an unbound nucleotide ($k = 0$) to 147 for a nucleosome ($k = K + 1$). Maintaining this property for all DBFs preserves the Boltzmann distribution.

Since $L_0 = 1$ and $w_0 = 1$, it follows from the above discussion that $\alpha_k = w_k \alpha_0^{L_k}$. Since $\{\alpha_0, \ldots, \alpha_{K+1}\}$ are a set of probabilities, it must also be the case that they sum to 1:

$$1 = \sum_{k=0}^{K+1} \alpha_k = \sum_{k=0}^{K+1} w_k \cdot \alpha_0^{L_k}$$

Finally, because we know all the values of w_k and L_k, we are left with an expression in just one unknown, α_0. We can easily solve for α_0, and then use it and the relationship above to compute the transition probabilities of all the other DBFs.

After initializing the transition probabilities as described above, we iteratively update them using Baum-Welch until convergence to a local optimum of the likelihood. To update α_k, we compute:

$$\alpha_k = \frac{\sum_{g=1}^{G} P(\pi_{k,1} \mid \boldsymbol{\theta^*}, \boldsymbol{s}, \boldsymbol{l}, \boldsymbol{m})}{\sum_{k'=0}^{K+1} \sum_{g=1}^{G} P(\pi_{k',1} \mid \boldsymbol{\theta^*}, \boldsymbol{s}, \boldsymbol{l}, \boldsymbol{m})}$$

Here, $\boldsymbol{\theta^*}$ represents all the model parameters. We find the likelihood converges within ten iterations (Supplementary Figure S7) and the optimized transition probabilities for each DBF almost always converge to the same final values regardless of how we initialize the weights (Supplementary Figure S8). We find convergence is faster for most DBFs when we initialize TF weights to K_d rather than multiples thereof (Supplementary Figure S8).

We find that transition probabilities for a few TFs with AT-rich motifs like Azf1 and Smp1 can grow quite large, resulting in a large number of binding sites in the genome, most of which are potential false positives. To curb the number of binding site predictions for such TFs, we apply a threshold on TF transition probabilities. The threshold δ is chosen to be two standard deviations above the mean of the initial transition probabilities of all the TFs (Supplementary Figure S9). Therefore, after the Baum-Welch step in every iteration, an additional modified Baum-Welch step is computed as follows:

$$\alpha_k = \begin{cases} (1 - n\delta) \dfrac{\sum_{g=1}^{G} P(\pi_{k,1} \mid \boldsymbol{\theta^*}, \boldsymbol{s}, \boldsymbol{l}, \boldsymbol{m})}{\sum_{k'=0, \alpha_{k'} < \delta}^{K+1} \sum_{g=1}^{G} P(\pi_{k',1} \mid \boldsymbol{\theta^*}, \boldsymbol{s}, \boldsymbol{l}, \boldsymbol{m})}, & \text{if } \alpha_k < \delta \\ \delta, & \text{otherwise} \end{cases}$$

where n is the number of TFs that have a transition probability more than δ. So, for all the TFs whose transition probabilities would be more than δ, they are instead set to δ, and the remaining DBFs (including the nucleosome and unbound state) have a regular Baum-Welch update of their transition probabilities. We find that this approach reduces the number of false positives (Supplementary Figure S10). An alternative mechanism might be to use an informed prior, in situations where prior information is available.

Implementation of posterior decoding

RoboCOP employs posterior decoding to infer probabilistic occupancy profiles of protein-DNA binding. The motivation behind posterior decoding is that it represents the thermodynamic ensemble of potential binding configurations; the resulting probability distribution sheds light on the many different ways proteins may be bound to the genome across a cell population (applying Viterbi decoding would not provide a probabilistic landscape, but only a single, most likely chromatin configuration). The resulting posterior probability of each DBF at each position in the genome provides a probabilistic profile of DBF occupancy at base-pair resolution (Figure 2E).

As a multivariate HMM, RoboCOP has a time complexity of $O(GN^2)$ and a space complexity of $O(GN)$ (for a genome of length G and where N denotes the total number of hidden states). The high complexity makes it difficult to decode the entire genome at once. To reduce the computational complexity of RoboCOP, we perform posterior decoding separately on blocks of the genome of length 5,000, with an overlap of 1,000 bases, and stitch results together. This ensures that the model has a sufficiently long sequence to learn an accurate chromatin landscape, but not so long that we run out of memory. In addition, we use only the longest chromosome (chrIV in yeast) to train DBF transition probabilities with Baum-Welch, and then undertake posterior decoding genome-wide.

Validation of TF and nucleosome predictions

Using the MNase-seq and ATAC-seq datasets, we ran RoboCOP three times, each time with three different set of inputs. $\text{RoboCOP}_{\text{MNase}}$ was run with MNase-seq alone, $\text{RoboCOP}_{\text{ATAC}}$ was run with ATAC-seq alone, and $\text{RoboCOP}_{\text{MNase+ATAC}}$ was run with both MNase-seq and ATAC-seq data as input. Unlike RoboCOP, COMPETE is not able to learn DBF weights from the data using Baum-Welch; rather, the weights need to be provided as input. We ran COMPETE with the TF weights set to different multiples of the TF dissociation constant K_d, $8K_d$, and $16K_d$ (21) and these models are referred to as $\text{COMPETE}_{\text{Kd}}$, $\text{COMPETE}_{\text{8Kd}}$, and $\text{COMPETE}_{\text{16Kd}}$, respectively. To isolate the differences in the learned chromatin occupancy profiles that arise from the inclusion of chromatin accessibility data as input to $\text{RoboCOP}_{\text{MNase}}$, we also ran COMPETE with the DBF weights learned by $\text{RoboCOP}_{\text{MNase}}$ after Baum-Welch training. We refer to this model as $\text{COMPETE}_{\text{RoboCOP}}$.

We use posterior probabilities of TF occupancy from RoboCOP and COMPETE outputs to identify binding sites. The starting probability of a motif is computed by adding the starting probability of the forward and reverse complement of the motif for every position in the genome. In the case of Rap1 which has multiple PWMs, the maximum starting probability among the PWMs is chosen at every position. For validation, a site is considered a true positive (TP) if it overlaps with an annotated binding site for that TF, and a false positive (FP) otherwise. If an annotated TF binding site does not overlap any of our predictions, it is a false negative (FN). We plotted precision-recall curves and calculated area under precision-recall curve for validating TF binding sites. We consider the TF binding sites in (25) as ground truth. We have more precise binding site predictions from ORGANIC (26) for Abf1 and Reb1, and from ChIP-exo for Reb1 and Rap1 (5). We combined these datasets with (25) and considered the combined dataset to be ground truth for these three TFs. For each pair of models, a pairwise Mann-Whitney U test was performed on the AUPR values of its TF predictions to establish statistical significance.

We call nucleosomes from RoboCOP and COMPETE outputs using the probability of the nucleosome dyad as computed by the two methods. We employ a greedy approach as described in (34). Briefly, we iteratively call nucleosome dyads with the highest probability and remove a genomic window of length 117 bases centered on the called nucleosome dyad from future inclusion. This allows nucleosomes to partially overlap. For validation, a nucleosome position is considered a true positive (TP) if the distance between the predicted and annotated dyad is 50 bases or less. We consider the nucleosome dyads reported in (24) to be our ground truth, and used precision-recall curves to validate all other predicted nucleosome positions against this reference set.

Comparison to DANPOS2 and NucleoATAC nucleosomes

To assess whether alternative methods might identify nucleosome positions more simply or accurately, we compared the nucleosome position predictions of RoboCOP against two established methods, one for calling nucleosome predictions from MNase-seq data (DANPOS2), and the other for calling nucleosome predictions from ATAC-seq data (NucleoATAC).

We provided our paired-end MNase-seq data to DANPOS version 2.2.2 (13), and allowed it to call nucleosome positions using the command: `danpos.py dpos -m 1`. The resulting occupancy attribute was used to sort predictions when evaluating precision and recall.

We used the nucleosome positions reported by the authors of NucleoATAC (20). Because NucleoATAC reports only a small fraction of all genomic nucleosomes (around 20%), we also evaluated how well different models predicted the limited set of 'NucleoATAC nucleosomes'. When doing so, we retained only nucleosomes within 55 bp of those predicted by NucleoATAC.

Comparison to FIMO-MNase and FIMO-ATAC TFs

To calibrate the accuracy of the TF binding site predictions of RoboCOP and COMPETE, we developed a baseline by running FIMO (35) on peaks of `shortFrags` for MNase-seq and ATAC-seq as follows. We first filtered `shortFrags` into a separate BAM file and then ran MACS2 (36) to call peaks with parameters `-f BAMPE -p 1e-20 --nolambda --nomodel`. We called peaks separately on MNase-seq and ATAC-seq data. Then, to scan for matches to any of our PWMs, we ran FIMO (35) within 50-bp windows centered on the reported peaks.

ORC mutant analysis

We used the most prevalent ORC binding motif within ORC ChIP-seq peaks, as previously reported (37). Annotated ACS sites in the yeast genome were obtained from a previous study (38).

Data sources

We generated RNA-seq and paired-end MNase-seq data from *S. cerevisiae* before and after cadmium treatment (19). These are available for download at the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) under accession number GSE153609. The MNase-seq data from the 0 min timepoint, before cadmium treatment

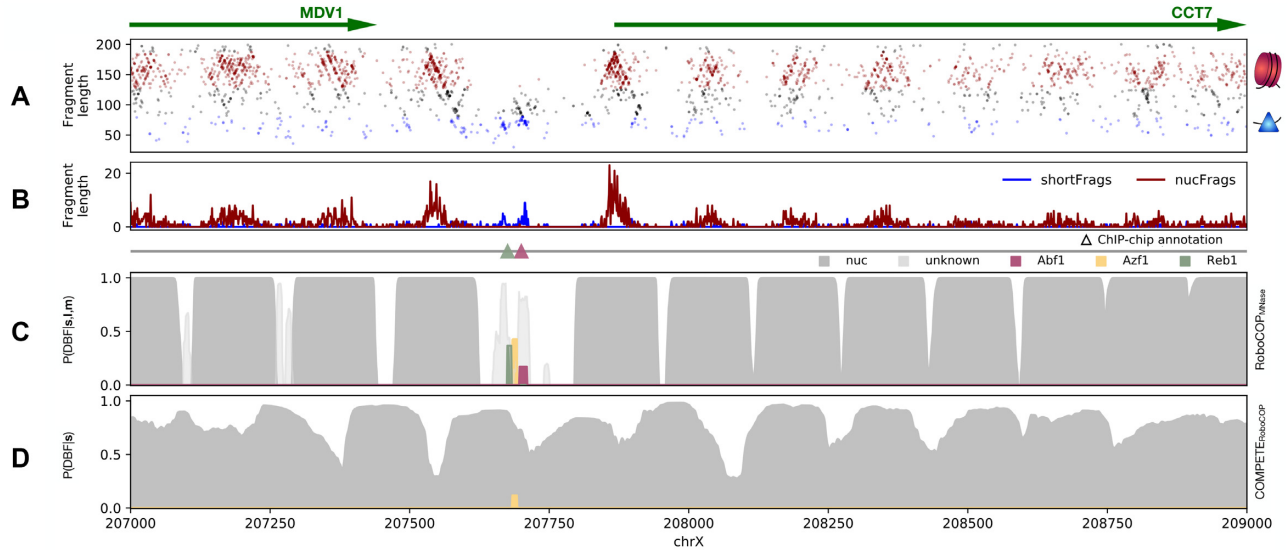

Figure 3. Representative chromatin occupancy profile produced by RoboCOP$_{MNase}$, in comparison with that of COMPETE$_{RoboCOP}$, an existing method. (**A**) Two-dimensional plot of MNase-seq fragments from positions 207,000 to 209,000 of yeast chromosome X, with `nucFrags` in red and `shortFrags` in blue. Gene annotations depicted with arrows at the top. (**B**) The `nucFrags` and `shortFrags` signals that result from aggregation of MNase-seq fragment midpoint counts in the region. (**C**) RoboCOP$_{MNase}$ and (**D**) COMPETE$_{RoboCOP}$ outputs for the region, with known TF binding sites indicated with triangles above. Because RoboCOP$_{MNase}$ makes use of MNase-seq data in generating its chromatin occupancy profile, it, unlike COMPETE$_{RoboCOP}$, positions nucleosomes more precisely and successfully identifies not only the nucleosome-depleted region, but also the known Abf1 and Reb1 binding sites therein (25).

(DM504), was the basis of most of our MNase-seq analysis. When exploring chromatin dynamics under cadmium stress, we used the 60 min timepoint (DM508).

Paired-end ATAC-seq data from *S. cerevisiae* were downloaded from GEO accession number GSE66386. Data from the 11 non-osmotic stress datasets were merged. Nucleosome calls from NucleoATAC applied to that data were downloaded from the same accession number.

We previously generated paired-end MNase-seq data from *S. cerevisiae* under G_2 arrest, in both wild-type and *orc1-161* mutant cells (17). These are available for download at the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under SRA study accession number SRP041314. Data from the two replicates in each case were merged.

The sacCer3 (April 2011) version of the yeast genome was used for all analyses.

RESULTS

RoboCOP computes probabilistic chromatin occupancy profiles

We use RoboCOP to predict the nucleosome positions and binding sites of 150 different TFs across the *Saccharomyces cerevisiae* genome using MNase-seq and/or ATAC-seq data. Even though we include 150 different TFs in our model (listed in Supplementary Table S1), this does not exhaust what binds the genome: We are missing replication factors and general transcription factors, as well as sequence-specific TFs for which we have no binding preference information. To address this, we add a 10-bp DBF labeled 'unknown' that we use to capture any extra `shortFrags` signal not captured by our 150 known TFs (this also has the salutary effect of reducing false positive predictions for the known TFs; see Supplementary Figure S10 for a comparison).

Beyond the genome sequence and the collection of DBFs and their binding preferences, RoboCOP$_{MNase}$, RoboCOP$_{ATAC}$, and RoboCOP$_{MNase+ATAC}$ take as input the `nucFrags` and `shortFrags` signals derived from paired-end MNase-seq, ATAC-seq, and both MNase-seq and ATAC-seq data respectively. Figure 3 shows the input MNase-seq data and the resulting RoboCOP$_{MNase}$ output for a representative segment of the genome. The nucleosome predictions in RoboCOP's output (Figure 3C) line up well with the nucleosomal fragments in the data (Figure 3A,B). In addition, RoboCOP$_{MNase}$ predicts one Abf1 and one Reb1 binding site, which align with the short fragments in the data and match annotated binding sites in this locus (25).

RoboCOP's use of chromatin accessibility data improves chromatin occupancy profiles

Our group's earlier work, COMPETE (21) uses only nucleotide sequence as input to an HMM in order to compute a probabilistic occupancy landscape of DBFs across a genome. COMPETE's output is theoretical in that it does not incorporate experimental data in learning the binding landscape of the genome. In order to make a fair comparison, COMPETE was run with DBF weights learned using RoboCOP$_{MNase}$ which we refer to as COMPETE$_{RoboCOP}$. Unsurprisingly, due to the lack of chromatin accessibility information in COMPETE$_{RoboCOP}$, the nucleosome positions learned by the model (Figure 3D) do not line up well with the nucleosomal signal apparent in the MNase-seq and ATAC-seq data (Figure 3A,B, Supplementary Figure

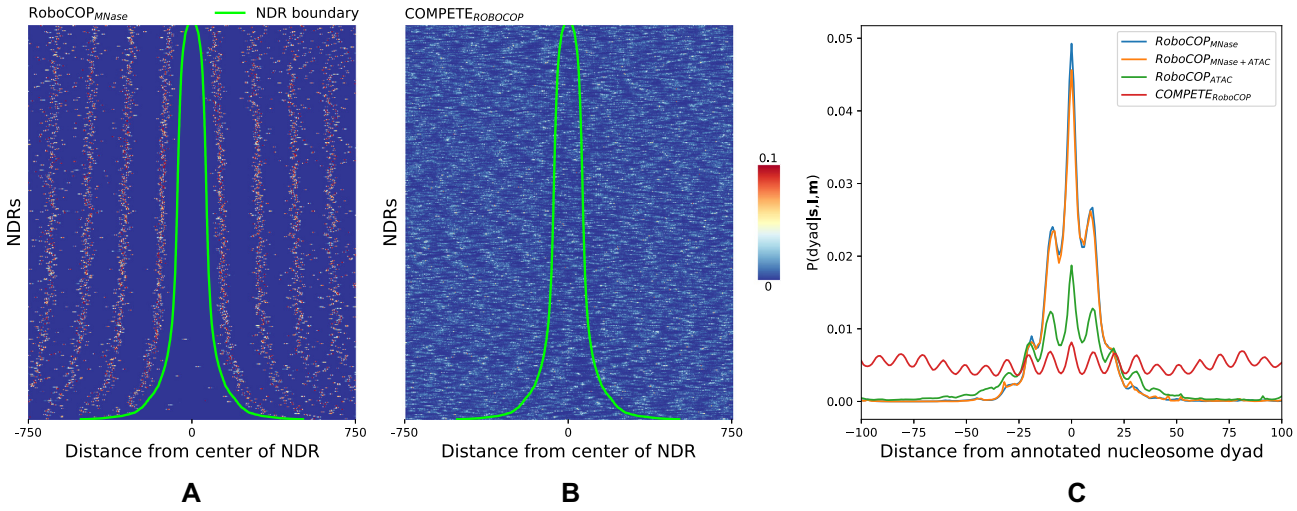

Figure 4. RoboCOP$_{MNase}$ positions nucleosomes with precision and accuracy, including avoiding their placement within nucleosome-depleted regions (NDRs). (**A**,**B**) Heatmaps depict the posterior probability of a nucleosome dyad as calculated by (**A**) RoboCOP$_{MNase}$ and (**B**) COMPETE$_{RoboCOP}$, at each position around experimentally determined NDRs genome-wide (32). Each row is a distinct NDR, sorted by NDR size. Lime green lines depict the experimentally determined NDR boundaries. Note that P(dyad | **s**,**l**,**m**) computed by RoboCOP$_{MNase}$ is appropriately almost always zero within NDRs, unlike COMPETE$_{RoboCOP}$, and the signal is well-phased in both directions. (**C**) Curves depict aggregate values of the posterior probability of nucleosome dyad across the 2,000 most well-positioned nucleosomes (24), as computed by RoboCOP$_{MNase}$, RoboCOP$_{MNase+ATAC}$, RoboCOP$_{ATAC}$, and COMPETE$_{RoboCOP}$. Aggregate signals all exhibit an expected ∼10 bp periodicity that arises from the periodic nature of the weak sequence specificity of nucleosomes. Note that RoboCOP-based predictions, especially those that include MNase data, appropriately peak at annotated dyads and fall off rapidly in both directions, indicating that learned positions are both more precise and more accurate than those of COMPETE$_{RoboCOP}$.

S11). The nucleosome predictions of COMPETE$_{RoboCOP}$ (Figure 3D) are more diffuse, which is understandable because it relies entirely on sequence information, and nucleosomes have only weak and periodic sequence specificity (1). Because of a lack of chromatin accessibility data, COMPETE$_{RoboCOP}$ fails to identify the clear nucleosome-depleted region in this locus (and does so all throughout the genome, as seen in Figure 4A,B), as a result of which it fails to recognize the Abf1 and Reb1 binding sites known to reside in the locus in (25). In contrast, RoboCOP$_{MNase}$ utilizes the chromatin accessibility data to accurately learn the nucleosome positions and the annotated Abf1 and Reb1 binding sites (Figure 3C).

Genome-wide prediction of nucleosome positioning

Nucleosomes have weak sequence specificity and can adopt alternative nearby positions along the genome (32,39). It is therefore likely that the nucleosome positions reported by one method do not exactly match those reported by another. However, since RoboCOP generates genome-wide probabilistic scores of nucleosome occupancy, we can plot the probability of a nucleosome dyad given the nucleotide sequence and MNase-seq signals, P(dyad | **s**,**l**,**m**), around annotated nucleosome locations (24). We find that the RoboCOP$_{MNase}$ dyad score peaks precisely at the annotated dyads (Figure 4C), and decreases almost symmetrically in either direction. In contrast, COMPETE$_{RoboCOP}$ does not provide accurate location predictions (Figure 4C); the oscillatory nature of the score reported by COMPETE$_{RoboCOP}$ reflects the periodic dinucleotide sequence specificity model for nucleosomes, and does not correspond well with actual nucleosome locations. When evaluated genome-wide using precision-recall curves (Supplementary Figure S12A), the nucleosome positions called by RoboCOP$_{MNase}$ are far more similar to the nucleosome annotations of Brogaard and colleagues (24) than are the ones called by COMPETE$_{RoboCOP}$, which are only slightly better than random (Supplementary Table S2).

Similarly, we find that when RoboCOP is run with ATAC-seq alone or with both MNase-seq and ATAC-seq, it can efficiently infer the nucleosome occupancy profile (Figure 4C) and also infer the NDR boundaries in the genome (Supplementary Figure S13,S14). Since RoboCOP uses a combination of `nucFrags`, `shortFrags`, and nucleotide sequence to infer DBF occupancy profile, it can successfully differentiate between nucleosome-dense regions and NDRs from both MNase-seq and ATAC-seq data. However, because MNase-seq generates a cleaner nucleosomal signal, the nucleosome predictions are more accurate when MNase-seq is provided as input to RoboCOP in comparison to ATAC-seq and the performance is intermediate when using both MNase-seq and ATAC-seq (Figure 4C, Supplementary Figure S15).

Likewise, precision-recall curves in Supplementary Figure S12A further confirm that the nucleosome positions predicted by the RoboCOP models incorporating MNase-seq have higher AUPR compared to RoboCOP$_{ATAC}$ that uses ATAC-seq alone. These models outperform all the COMPETE models that do not utilize chromatin accessibility data. We also find that COMPETE$_{RoboCOP}$ which has its parameters initialized with the parameters learned by RoboCOP$_{MNase}$, performs better than other models of COMPETE. On comparing to the nucleosome positions predicted by DANPOS2 (13) to MNase-seq, the RoboCOP models incorporating MNase-seq have higher AUPR (Supplementary Figure S12A). The AUPR of RoboCOP$_{ATAC}$

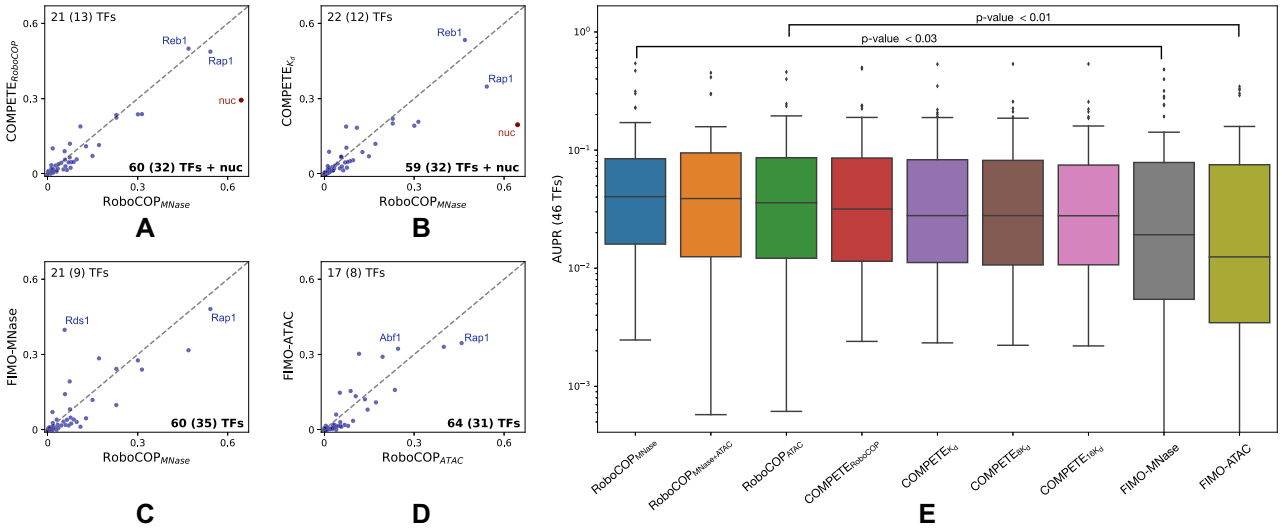

Figure 5. Scatter plots comparing AUPR curves of 81 TFs (blue dots) for (**A**) RoboCOP$_{MNase}$ and COMPETE$_{RoboCOP}$, (**B**) RoboCOP$_{MNase}$ and COMPETE$_{Kd}$, (**C**) RoboCOP$_{MNase}$ and FIMO-MNase, and (**D**) RoboCOP$_{ATAC}$ and FIMO-ATAC. (**A**) and (**B**) also compare the AUPR for nucleosome positions for the three methods (red dots). FIMO-MNase and FIMO-ATAC do not predict nucleosome positions. Hence, (**C**) and (**D**) do not have AUPR value for nucleosome. The numbers on the top-left and bottom-right corners of (**A**–**D**) indicate the number of DBFs for which the model on the given axis performs better than the model on the other axis. The numbers are in bold if they are greater than the numbers on the other axis. Because the AUPR values are close to 0 for many DBFs, the numbers within parentheses indicate the number of factors for which the model on the given axis has higher AUPR value and is at least 0.01. (**E**) Box plots comparing the AUPR values across all methods for all 46 TFs that have AUPR $\geq$ 0.01 for at least one method. The box plots are sorted in decreasing order of median AUPR per method. *P*-value significance is shown only for two of the four comparisons made in (**A**–**D**). Both RoboCOP$_{MNase}$ and RoboCOP$_{ATAC}$ outperform the baseline methods FIMO-MNase and FIMO-ATAC. All RoboCOP models have higher accuracy than COMPETE models or baseline methods. Remaining *p*-values are reported in Supplementary Table S3.

is lower compared to DANPOS2 because of the sparsity of ATAC-seq data within the gene bodies. However, RoboCOP$_{ATAC}$ finds more nucleosomes than NucleoATAC, another method that incorporates ATAC-seq to find nucleosome positions (Supplementary Figure S12A). We notice that even though NucleoATAC finds fewer nucleosomes, it finds the nucleosomes with high precision (Supplementary Figure S12A). Since NucleoATAC identifies nucleosomes that are around open chromatin (20), we compared the precision recall curves of the nucleosomes identified by other models near the NucleoATAC nucleosomes. We find that the RoboCOP models and DANPOS2 have similar performance as NucleoATAC (Supplementary Figure S12B). Moreover, we notice in Supplementary Figure S12B that the COMPETE models also have high precision indicating that the nucleosomes around the accessible regions of the chromatin are overall easier to predict solely based on nucleotide sequence specificity, even without chromatin accessibility data.

Genome-wide prediction of TF binding sites

RoboCOP seeks to learn a full chromatin occupancy profile of multiple DBFs at once, with the hope that the binding sites of many TFs can be predicted simultaneously at least as well as they are by existing methods and assays that identify them one at a time. This is a significant challenge because while MNase-seq and ATAC-seq data have been reported to provide evidence of binding for at least some TFs and DNA replication initiation factors (15–18,40), the `shortFrags` signal of chromatin accessibility datasets is quite noisy. For instance, TFs can sometimes be bound transiently (41), allowing the entire region to be digested by MNase or cleaved by Tn5 transposase and leaving behind no `shortFrags` signal.

Although RoboCOP predicts the genome-wide occupancy of a set of 150 TFs, we can only validate the binding sites of 81 of them, given available ChIP-chip (25), ChIP-exo (5), and ORGANIC (26) datasets (Supplementary Table S1). Making things more complicated, available yeast ChIP-chip data assay binding at the genomic resolution of whole intergenic regions, with computational algorithms being used to refine those into specific binding sites, making the ChIP-chip dataset somewhat less reliable for validation purposes. Compounding the problem, data for many of the TFs were generated under multiple conditions (3) (Supplementary Table S1), but the annotations are not generally condition-specific.

With those caveats in place, we compare TF binding site predictions made by the three runs of RoboCOP (RoboCOP$_{MNase}$, RoboCOP$_{ATAC}$, and RoboCOP$_{MNase+ATAC}$) to predictions made by the different runs of COMPETE (COMPETE$_{RoboCOP}$, COMPETE$_{Kd}$, COMPETE$_{8Kd}$, and COMPETE$_{16Kd}$) and FIMO-MNase and FIMO-ATAC. We observe that RoboCOP$_{MNase}$ has higher AUPR values for more DBFs compared to COMPETE$_{RoboCOP}$, both of which have the same DBF weights with the exception that RoboCOP$_{MNase}$ has access to MNase-seq data whereas COMPETE$_{RoboCOP}$ builds the model entirely based on the nucleotide sequence of the genome (Figure 5A). Both RoboCOP$_{MNase}$ and COMPETE$_{RoboCOP}$ have higher AUPR values compared to other COMPETE models that were assigned DBF weights using multiples of dissociation constants of the

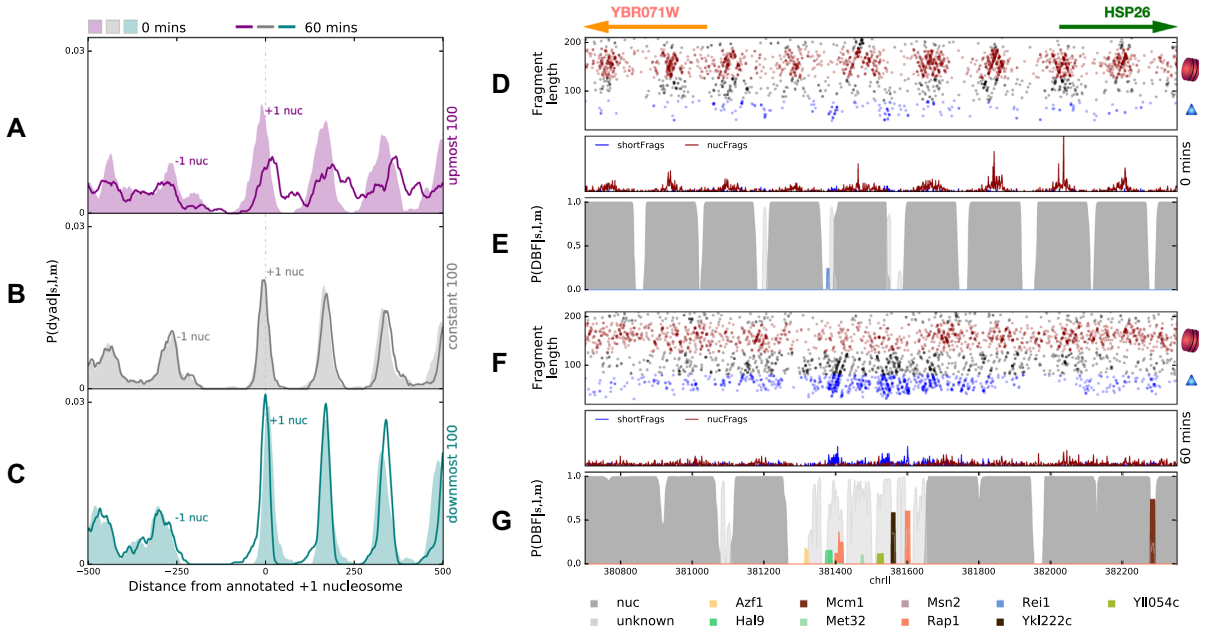

Figure 6. (**A**–**C**) Aggregate nucleosome dyad probability, as computed by RoboCOP$_{MNase}$, around annotated +1 nucleosomes (32) of (**A**) the 100 most up-regulated genes (purple), (**B**) the 100 genes least changed in transcription (gray), and (**C**) the 100 most down-regulated genes (teal), before and 60 min after treating cells with cadmium. After treatment, we see the +1 nucleosome closing in on the promoters of the most down-regulated genes (teal) but opening up the promoters of the most up-regulated genes (purple). (**D**) Two-dimensional plot of MNase-seq fragments near the HSP26 promoter (positions 380,700 to 382,350 of yeast chromosome II are shown) before treatment with cadmium (`nucFrags` in red; `shortFrags` in blue), along with the `nucFrags` and `shortFrags` signals that result from aggregating those midpoint counts. Gene annotations depicted with arrows at the top (Watson strand in green; Crick strand in orange). (**E**) RoboCOP$_{MNase}$-predicted occupancy profile of this region before treatment with cadmium. (**F,G**) The same as (**D**,**E**), respectively, but 60 min after cadmium treatment. HSP26 transcription is highly up-regulated under cadmium stress, and we observe here that its promoter exhibits marked TF binding after treatment, most prominently by Rap1, known to bind this promoter during stress response. Nucleosome positions also shift notably.

TFs (Figure 5B, Supplementary Figure S16). We also observe that RoboCOP$_{MNase}$ and RoboCOP$_{ATAC}$ perform significantly better than the two baseline methods that we defined as FIMO-MNase and FIMO-ATAC (Figure 5A–E). We notice that overall the RoboCOP models fare better than the COMPETE models which are further significantly better than the baseline models of FIMO-MNase and FIMO-ATAC (Figure 5E, Supplementary Figure S16, Supplementary Table S3). The precision-recall curves in Supplementary Figure S17 show the AUPR values for the individual TFs.

RoboCOP reveals chromatin dynamics under cadmium stress

One of the most powerful uses of RoboCOP is that it can elucidate the dynamics of chromatin occupancy, generating profiles under changing environmental conditions. As an example, we explore the occupancy profiles of yeast cells before and after being subjected to cadmium stress for 60 min. We run RoboCOP$_{MNase}$ separately on two MNase-seq datasets: one for a cell population before treatment and another 60 min after treatment with 1 mM of $CdCl_2$. Cadmium is toxic to the cells and activates stress response pathways. Stress response genes are heavily transcribed under cadmium treatment, while ribosomal genes are repressed (42). We use RNA-seq to identify the 100 genes most up-regulated ('upmost 100', for short) and the 100 genes most down-regulated ('downmost 100'). As a control, we choose the 100 genes with the least change in transcription under treatment ('constant 100') (see Supplementary Table S4 for the three gene lists). Separately for each group of genes, we plot the composite RoboCOP$_{MNase}$-predicted nucleosome dyad probability in a 1000-bp window centered on established +1 nucleosome annotations (32). Prior to cadmium treatment, the composite +1 nucleosome peaks for all three groups align closely with the annotations (filled curves in Figure 6A–C). Upon treatment with cadmium, the +1 nucleosomes of the upmost 100 genes shift downstream, expanding the NDR (solid curve in Figure 6A). Owing to high variability in the new positions of the +1 nucleosomes of the upmost 100 genes, the composite +1 nucleosome peak for these genes becomes shorter and broader. Furthermore, the position of the −1 nucleosome also becomes more uncertain with the expansion of the NDR. In contrast, the +1 nucleosomes of the downmost 100 genes shift upstream, closing in on the NDR (solid curve in Figure 6C). Interestingly, the shift is precise, resulting in the composite +1 nucleosome peak remaining narrow and sharp. Unlike the upmost 100 genes, we do not see changes in the position of the −1 nucleosomes of the downmost 100 genes. As expected from a control, we observe no changes in the position of the +1 nucleosome for the constant 100 genes (Figure 6B).

We can also use RoboCOP$_{MNase}$ to study detailed changes in the chromatin landscape under cadmium stress within a specific locus, for example that of HSP26, a key stress response gene in the upmost 100 genes. In Figure 6D–G, we notice the HSP26 promoter opening up under stress, with shifts in nucleosomes leading to more TF binding in

the promoter. From the `shortFrags` midpoint counts, RoboCOP$_{MNase}$ identifies multiple potential TF binding sites, most prominently for Rap1, which has already been shown to re-localize to the promoter region of HSP26 during general stress response (43).

In comparison, COMPETE$_{RoboCOP}$ fails to capture the dynamics of chromatin occupancy under cadmium stress because it does not incorporate chromatin accessibility information into its model. We ran COMPETE$_{RoboCOP}$ with the RoboCOP$_{MNase}$-trained DBF weights for the two time points of cadmium treatment and found that COMPETE$_{RoboCOP}$ generates binding landscapes for the two time points that are nearly identical (Supplementary Figure S18). This is a key difference between RoboCOP$_{MNase}$ and COMPETE$_{RoboCOP}$: Being able to incorporate experimental chromatin accessibility data allows RoboCOP$_{MNase}$ to provide a more accurate binding profile for cell populations undergoing dramatic chromatin changes.

The preceding analysis highlights the broad utility of RoboCOP. Because RoboCOP models DBFs competing to bind the genome, it produces a probabilistic prediction of the occupancy level of each DBF at single-nucleotide resolution. Moreover, as the chromatin architecture changes under different environmental conditions, RoboCOP is able to elucidate the dynamics of chromatin occupancy. The cadmium treatment experiment shows that the predictions made by RoboCOP can be used both to study overall changes for groups of genes (Figure 6A–C), as well as to focus on specific genomic loci in order to understand their detailed chromatin dynamics (Figure 6D–G).

DISCUSSION

RoboCOP is a new computational method that utilizes a multivariate HMM to generate a probabilistic occupancy profile of the genome by integrating chromatin accessibility data with nucleotide sequence. The integration of experimental data leads to a number of improvements over COMPETE: It increases the accuracy of TF binding site predictions, it markedly increases the accuracy of nucleosome positioning predictions, and it provides a principled mechanism for learning DBF transition probabilities.

RoboCOP can learn chromatin occupancy profiles from different kinds of accessibility data; in this paper, we have demonstrated its application to MNase- and ATAC-seq data, separately or in combination. However, we observe markedly better performance when using MNase-seq than when using ATAC-seq, for two primary reasons. First, ATAC-seq fragments are highly enriched for regions of open chromatin, whereas MNase-seq fragments are distributed more evenly across the entirety of the genome, allowing for a more comprehensive view of the chromatin genome-wide. Second, MNase has an exonuclease activity which allows fragments to be digested to more accurately reflect the sizes of the DBFs that protect them. If one were designing an experiment to generate data for RoboCOP, we would recommend MNase-seq, but in many cases, ATAC-seq (or DNase-seq) data is already available, and it is reassuring that RoboCOP is applicable to this kind of data also, though its insights will likely be concentrated near regions of open chromatin.

The chromatin occupancy profiles produced by RoboCOP are very effective at positioning nucleosomes (especially with MNase-seq data, but RoboCOP$_{ATAC}$ also positions many more nucleosomes than NucleoATAC using the same data). However, inferring TF occupancy from chromatin accessibility data remains a challenge. Nevertheless, we observe that RoboCOP with MNase- or ATAC-seq data performs notably better than alternative approaches that use peak or footprint identification with MACS2 (36) followed by TF-labeling with FIMO (35). Presumably this is because RoboCOP considers all DBFs together within a single joint model that explicitly accounts for the thermodynamic competition among DBFs, including nucleosomes. In future work, it might be possible to improve TF binding site predictions through the incorporation of prior information about DBF transition probabilities. Regardless, the accuracy and comprehensiveness of chromatin occupancy profiles will improve with deeper sequencing. Sufficient read depth is already commonplace in compact eukaryotic genomes like that of yeast (as we demonstrate here), attainable in medium-sized genomes like worm or fly, and hopefully feasible in the near future in larger mammalian genomes.

One of the most important applications of RoboCOP is that it enables comparison of the chromatin landscape across time or varying conditions, facilitating the study of chromatin dynamics (a task to which COMPETE is unsuited). We have demonstrated that RoboCOP can reveal chromatin changes in response to an environmental stimulus like cadmium exposure, but it can be applied in other contexts as well. As one example, we applied RoboCOP to MNase-seq data from two different populations of cells under G_2 arrest, one wild-type and the other an *orc1-161* temperature-sensitive mutant that prevents binding of the origin recognition complex (ORC) at origins of replication. In Supplementary Figure S19, we show that RoboCOP detects strong ORC binding during G_2 at origins in wild-type cells, but not in mutant cells, and also reveals that flanking nucleosomes are positioned farther apart when ORC is bound. This example additionally highlights that RoboCOP can be applied to any site-specific DBF, not just nucleosomes and TFs.

Importantly, since changes in transcription can likewise be measured across time or varying conditions, in conjunction with such measurements, RoboCOP can help elucidate how the dynamics of chromatin occupancy and the dynamics of gene expression interrelate.

DATA AVAILABILITY

RoboCOP source code is available at: https://github.com/HarteminkLab/RoboCOP. MNase-seq and RNA-seq data used for various analyses (including cadmium stress) can be downloaded from GEO under accession number GSE153609. ATAC-seq data can be downloaded from GEO under accession number GSE66386. MNase-seq data used to compare the *orc1-161* mutant and wild type under G_2 arrest can be downloaded from SRA under accession number SRP041314.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Heather MacAlpine and Vinay Tripuraneni for generating the MNase-seq data, and Greg Crawford, Raluca Gordân, Ed Iversen, Yulong Li, and Albert Xue for helpful comments and feedback during the development of RoboCOP.

FUNDING

National Institute of General Medical Sciences [R35 GM127062 to D.M.M., R01 GM118551 to A.J.H.]. Funding for open access charge: National Institute of General Medical Sciences [follow-up grant R35 GM141795 to A.J.H.]
Conflict of interest statement. None declared.

REFERENCES

1. Kaplan,N., Moore,I.K., Fondufe-Mittendorf,Y., Gossett,A.J., Tillo,D., Field,Y., LeProust,E.M., Hughes,T.R., Lieb,J.D., Widom,J. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
2. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
3. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., MacIsaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.-B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
4. Park,P.J. (2009) ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
5. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
6. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
7. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
8. Benner,P. and Vingron,M. (2019) ModHMM: a modular supra-Bayesian genome segmentation method. *J. Comput. Biol.*, **27**, 442–457.
9. Tarbell,E.D. and Liu,T. (2019) HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Res.*, **47**, e91.
10. Bernstein,B.E., Liu,C.L., Humphrey,E.L., Perlstein,E.O. and Schreiber,S.L. (2004) Global nucleosome occupancy in yeast. *Genome Biol.*, **5**, R62.
11. Yuan,G.-C., Liu,Y.-J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, **309**, 626–630.
12. Mavrich,T.N., Ioshikhes,I.P., Venters,B.J., Jiang,C., Tomsho,L.P., Qi,J., Schuster,S.C., Albert,I. and Pugh,B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.
13. Chen,K., Xi,Y., Pan,X., Li,Z., Kaestner,K., Tyler,J., Dent,S., He,X. and Li,W. (2013) DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.*, **23**, 341–351.
14. Chen,W., Liu,Y., Zhu,S., Green,C.D., Wei,G. and Han,J.-D.J. (2014) Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nat. Commun.*, **5**, 4909.
15. 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.
16. Henikoff,J.G., Belsky,J.A., Krassovsky,K., MacAlpine,D.M. and Henikoff,S. (2011) Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 18318–18323.
17. Belsky,J.A., MacAlpine,H.K., Lubelsky,Y., Hartemink,A.J. and MacAlpine,D.M. (2015) Genome-wide chromatin footprinting reveals changes in replication origin architecture induced by pre-RC assembly. *Gene. Dev.*, **29**, 212–224.
18. Ramachandran,S. and Henikoff,S. (2016) Transcriptional regulators compete with nucleosomes post-replication. *Cell*, **165**, 580–592.
19. Tran,T.Q., MacAlpine,H.K., Tripuraneni,V., Mitra,S., MacAlpine,D.M. and Hartemink,A.J. (2021) Linking the dynamics of chromatin occupancy and transcription with predictive models. *Genome Res.*, **31**, 1035–1046.
20. Schep,A.N., Buenrostro,J.D., Denny,S.K., Schwartz,K., Sherlock,G. and Greenleaf,W.J. (2015) Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.*, **25**, 1757–1770.
21. Wasson,T. and Hartemink,A.J. (2009) An ensemble model of competitive multi-factor binding of the genome. *Genome Res.*, **19**, 2101–2112.
22. Zhong,J., Wasson,T. and Hartemink,A.J. (2014) Learning protein-DNA interaction landscapes by integrating experimental data through computational models. *Bioinformatics*, **30**, 2868–2874.
23. Zhong,J. (2015) In: *Computational inference of genome-wide protein-DNA interactions using high-throughput genomic data*. PhD dissertation, Duke University.
24. Brogaard,K., Xi,L., Wang,J.-P. and Widom,J. (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature*, **486**, 496–501.
25. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
26. Kasinathan,S., Orsi,G.A., Zentner,G.E., Ahmad,K. and Henikoff,S. (2014) High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods*, **11**, 203–209.
27. Zhang,Z. and Pugh,B.F. (2011) High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, **144**, 175–186.
28. Mieczkowski,J., Cook,A., Bowman,S.K., Mueller,B., Alver,B.H., Kundu,S., Deaton,A.M., Urban,J.A., Larschan,E., Park,P.J. *et al.* (2016) MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.*, **7**, 11485.
29. Rhee,H.S., Bataille,A.R., Zhang,L. and Pugh,B.F. (2014) Subnucleosomal structures and nucleosome asymmetry across a genome. *Cell*, **159**, 1377–1388.
30. Gordân,R., Murphy,K.F., McCord,R.P., Zhu,C., Vedenko,A. and Bulyk,M.L. (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol.*, **12**, R125.
31. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thåström,A., Field,Y., Moore,I.K., Wang,J.-P.Z. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
32. Chereji,R.V., Ramachandran,S., Bryson,T.D. and Henikoff,S. (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol.*, **19**, 19.
33. Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.*, **6**, R87.
34. Zhong,J., Luo,K., Winter,P.S., Crawford,G.E., Iversen,E.S. and Hartemink,A.J. (2016) Mapping nucleosome positions using DNase-seq. *Genome Res.*, **26**, 351–364.
35. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
36. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. and Liu,X.S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
37. Lee,C. S.K., Cheung,M.F., Li,J., Zhao,Y., Lam,W.H., Ho,V., Rohs,R., Zhai,Y., Leung,D. and Tye,B.-K. (2021) Humanizing the yeast origin recognition complex. *Nat. Commun.*, **12**, 33.
38. Eaton,M.L., Galani,K., Kang,S., Bell,S.P. and MacAlpine,D.M. (2010) Conserved nucleosome positioning defines replication origins. *Gene. Dev.*, **24**, 748–753.

39. Fragoso,G., John,S., Roberts,M.S. and Hager,G.L. (1995) Nucleosome positioning on the MMTV LTR results from the frequency-biased occupancy of multiple frames. *Gene. Dev.*, **9**, 1933–1947.
40. Li,Z., Schulz,M.H., Look,T., Begemann,M., Zenke,M. and Costa,I.G. (2019) Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.*, **20**, 45.
41. 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.
42. Hosiner,D., Gerber,S., Lichtenberg-Fraté,H., Glaser,W., Schüller,C. and Klipp,E. (2014) Impact of acute metal stress in *Saccharomyces cerevisiae*. *PLOS One*, **9**, e83330.
43. Platt,J.M., Ryvkin,P., Wanat,J.J., Donahue,G., Ricketts,M.D., Barrett,S.P., Waters,H.J., Song,S., Chavez,A., Abdallah,K.O. *et al.* (2013) Rap1 relocalization contributes to the chromatin-mediated gene expression profile and pace of cell senescence. *Gene. Dev.*, **27**, 1406–1420.