RoboCOP: Multivariate State Space Model Integrating Epigenomic Accessibility Data to Elucidate Genome-Wide Chromatin Occupancy

Sneha Mitra[1], Jianling Zhong[2], David M. MacAlpine[2,3,4], and Alexander J. Hartemink[1,2,4(✉)]

[1] Department of Computer Science, Duke University, Durham, NC 27708, USA
amink@cs.duke.edu
[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
[4] Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, USA

Abstract. Chromatin is the 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 level of occupancy 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. On the other hand, epigenomic accessibility data like ATAC-seq, DNase-seq, and MNase-seq provide insight into the chromatin landscape of all factors bound along the genome, but with minimal insight into the identities of those factors. Here, we present `RoboCOP`, a multivariate state space model that integrates chromatin information from epigenomic accessibility data with nucleotide sequence to compute genome-wide probabilistic scores of nucleosome and TF occupancy, for hundreds of different factors at once. `RoboCOP` can be applied to any epigenomic dataset that provides quantitative insight into chromatin accessibility in any organism, but here we apply it to MNase-seq data to elucidate the protein-binding landscape of nucleosomes and 150 TFs across the yeast genome. Using available protein-binding datasets from the literature, we show that our model more accurately predicts the binding of these factors genome-wide.

Keywords: Chromatin accessibility · MNase-seq · Hidden Markov model

R. Schwartz (Ed.): RECOMB 2020, LNBI 12074, pp. 136–151, 2020.
https://doi.org/10.1007/978-3-030-45257-5_9

1 Introduction

Chromatin is a tightly packaged structure of proteins and DNA in the nucleus of a cell. The arrangement of different proteins along the DNA determines how gene expression is regulated. Two important groups of DNA binding factors (DBFs) are transcription factors (TFs) and nucleosomes. TFs are key gene regulatory proteins that promote or suppress transcription 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 exhibit preferences for a periodic arrangement of dinucleotides that facilitate DNA wrapping. Likened to beads on a string, nucleosomes are positioned fairly regularly along the DNA, occupying about 81% of the genome in the case of *Saccharomyces cerevisiae* (yeast) [14]. In taking up their respective positions, nucleosomes allow or block TFs from occupying their putative binding sites, thereby contributing to the regulation of gene expression. Revealing the chromatin landscape—how all these DBFs are positioned along the genome—is therefore crucial to developing a more mechanistic (and eventually predictive) understanding of gene regulation.

Antibody-based methods have been used extensively to assay the binding of particular DBFs at high resolution. However, such methods are limited to assaying only one factor at a time. Chromatin accessibility datasets, on the other hand, provide information about open regions of the chromatin, indirectly telling us about the regions occupied by various proteins. Many protocols can be used to generate chromatin accessibility data, including transposon insertion (ATAC-seq), enzymatic cleavage (DNase-seq), or enzymatic digestion (MNase-seq). In the latter, the endo-exonuclease MNase is used to digest unbound DNA, leaving behind undigested fragments of bound DNA. Paired-end sequencing of these fragments reveals not only their location but also their length, yielding information about the sizes of the proteins bound in different genomic regions. MNase-seq has been widely used to study nucleosome positions [3,4], but evidence of TF binding sites has also been observed in the data [10].

Several chromatin segmentation methods use epigenomic data to infer the locations of 'states' like promoters and enhancers, particularly in human and mouse genomes [1,6,11,22], but identifying the precise binding locations of myriad individual DBFs is more difficult. The high cost of repeated deep sequencing of large genomes poses a major challenge. In comparison to the complex human and mouse genomes, the problem is a bit simpler when working with the yeast genome, because it is smaller and therefore more economical to sequence deeply.

In earlier work, we proposed `COMPETE` to compute a probabilistic occupancy landscape of DBFs along the genome [23]. `COMPETE` considers DBFs binding to the genome in the form 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 (static) and DBF concentrations (dynamic); it is entirely theoretical, in that it makes no use of experimental chromatin data to influence its predictions of the chromatin landscape. A modified version of `COMPETE` was later developed to estimate DBF

concentrations by maximizing the correlation between COMPETE's output and an MNase-seq signal, improving the reported binding landscape [25]. However, it still does not directly incorporate chromatin accessibility data into the model.

Here, we present RoboCOP, a new method that integrates epigenomic accessibility data and genomic sequence to produce accurate chromatin occupancy profiles of the genome. With nucleotide sequence and chromatin accessibility data as input, RoboCOP uses a multivariate hidden Markov model (HMM) to compute a probabilistic occupancy landscape of hundreds of DBFs genome-wide at single-nucleotide resolution. In this paper, we use paired-end MNase-seq data to predict TF binding sites and nucleosome positions throughout the *Saccharomyces cerevisiae* genome. We validate our TF binding site predictions using annotations reported by ChIP [15], ChIP-exo [19], and ORGANIC [13] experiments, and our nucleosome positioning predictions using high-precision annotations reported by a chemical cleavage method [2]. We find that RoboCOP provides valuable insight into the chromatin architecture of the genome, and can elucidate how it changes in response to different environmental conditions.

2 Results

2.1 MNase-seq Fragments of Different Lengths Are Informative About Different DNA Binding Factors

In Fig. 1a, we plot MNase-seq fragments around the transcription start sites (TSSs) of all yeast genes [16]. Fragments of length 127–187 (which we call

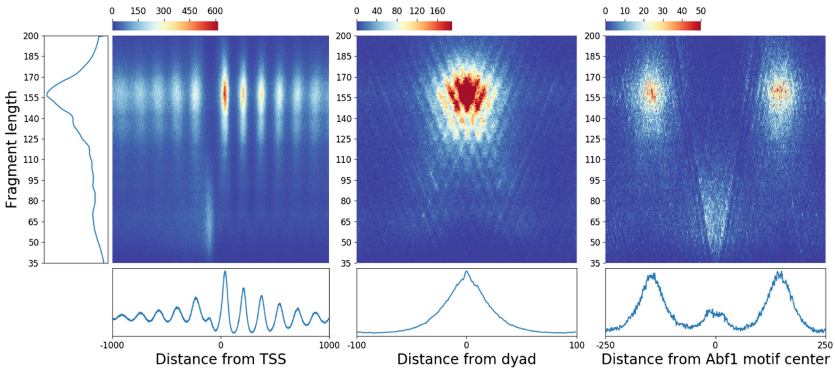


Fig. 1. (a) Heatmap of MNase-seq fragments, centered on all TSSs. Each fragment is plotted based on its length (y-axis) and the location of its midpoint (x-axis). Panels along the side and bottom show marginal densities. Heatmap reveals strong enrichment (red) of fragments corresponding to +1 nucleosomes (just downstream of TSS, lengths near 157). Upstream of TSS, in the promoter region, are many `shortFrags` (length $\leq$80). (b) Heatmap of MNase-seq fragments, centered on dyads of top 2000 well-positioned nucleosomes [2]. Fragment midpoint counts are highest at the dyad and decrease symmetrically in either direction. (c) Heatmap of MNase-seq fragments, centered on annotated Abf1 binding sites [15], showing an enrichment of `shortFrags` near Abf1 sites.

nucleosomal fragments, or `nucFrags` for short) occur in tandem arrays within gene bodies but are generally absent from promoters (Fig. 1a). Fragments are particularly concentrated at the +1 nucleosome position, 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 [2] peaks precisely at the dyad, with counts dropping nearly symmetrically in either direction (Fig. 1b). This makes sense because MNase digests linker regions, leaving behind undigested DNA fragments wrapped around histone octamers. So the midpoint 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 [10]. To verify that we see this signal in our data, Fig. 1a reveals 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 annotated TF binding sites. 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 (Fig. 1c). We denote these short fragments of length less than 80 as `shortFrags`. Unlike the midpoint counts of the `nucFrags` which have a symmetrically decreasing shape around the nucleosome dyads, the midpoint counts of `shortFrags` are more uniformly distributed within the binding site (Fig. 1c). 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 [18]. 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`. A representative snapshot of MNase-seq fragments is shown in Fig. 2a. We further simplify the two-dimensional plot in Fig. 2a to form two one-dimensional signals by separately aggregating the midpoint counts of `nucFrags` and `shortFrags`, as shown in Fig. 2b.

2.2 `RoboCOP` Computes Probabilistic Chromatin Occupancy Profiles

`RoboCOP` (**robo**tic **c**hromatin **o**ccupancy **p**rofiler) is a multivariate hidden Markov model (HMM) that jointly models the nucleotide sequence and the midpoint counts of `nucFrags` and `shortFrags` to learn the occupancy landscape of nucleosomes and TFs across a genome at single-nucleotide resolution. We apply `RoboCOP` on the *Saccharomyces cerevisiae* genome to predict nucleosome positions and the binding sites of 150 TFs (listed in Table S1). The HMM is

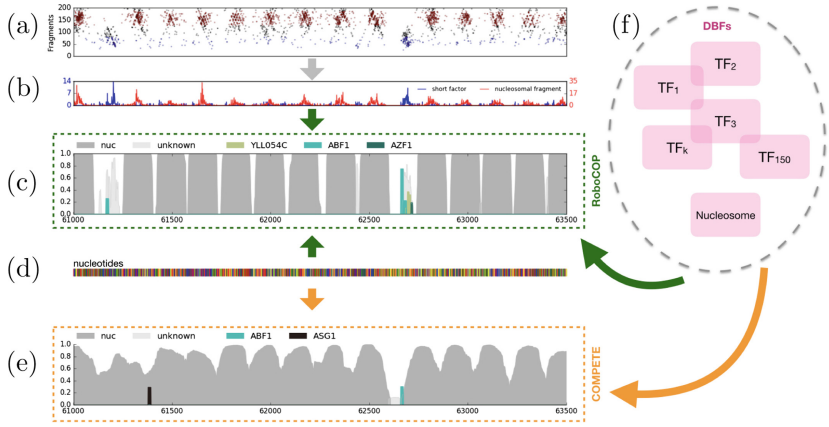


Fig. 2. (a) MNase-seq fragment midpoints in region chrI:61000–63500 of the yeast genome. Each dot at position (x, y) (red for `nucFrags`; blue for `shortFrags`) represents a fragment of length y centered on genomic coordinate x. (b) Aggregate numbers of red and blue dots. (d) Nucleotide sequence for chrI:61000–63500. (f) Set of DBFs (nucleosomes and TFs). (c) `RoboCOP` and (e) `COMPETE` outputs, with inputs depicted using green and orange arrows respectively. The score on the y-axes of (c) and (e) is the probability of that location being bound by each DBF.

structured such that each DBF corresponds to a collection of hidden states and each hidden state corresponds to a single genome coordinate. The hidden states of `RoboCOP` are inferred from a set of three observables at each coordinate: the nucleotide and the midpoint counts of `nucFrags` and `shortFrags` (Fig. S1). Based on these three observables, we estimate the posterior distribution over all hidden states. The resulting posterior probability of each DBF at each position in the genome provides a probabilistic profile of DBF occupancy at base-pair resolution (Fig. 2c). The inputs to `RoboCOP` are a set of DBFs (Fig. 2f), MNase-seq midpoint counts (Fig. 2b), and nucleotide sequence (Fig. 2d). From `RoboCOP` output in Fig. 2c, we observe that the nucleosome predictions line up well with the nucleosome signal in Figs. 2a,b.

`RoboCOP`'s emission probabilities are derived from published position weight matrices [7] and MNase-seq signals around annotated DBF-occupied regions. These emission probabilities are fixed, remaining unchanged during model optimization. The transition probabilities among the DBFs, however, are unknown, so we optimize these parameters using expectation maximization (EM).

Our model contains 150 TFs and nucleosomes, but that is not all possible DBFs. Thus, factors not in the model could be responsible for the enrichment in `nucFrags` and `shortFrags` at certain locations. We observe the midpoint counts of `shortFrags` to be noisier, likely because of the binding of other small complexes that are not a part of the model, including general TFs. We therefore add an 'unknown factor' into the model to account for such DBFs. It is plotted in light gray in Fig. 2c. Incorporating this unknown factor reduces false positives among binding site predictions for the other TFs (Fig. S2).

2.3 RoboCOP's Use of Epigenomic Accessibility Data Improves the Resulting Chromatin Occupancy Profiles

Our group's previous work, COMPETE [23], is an HMM that computes a probabilistic occupancy landscape of the DBFs in the genome using only nucleotide sequence as input. The model output is theoretical in that it does not incorporate experimental data in learning the binding landscape of the genome. Perhaps unsurprisingly, the nucleosome positions learned by COMPETE (Fig. 2e) do not line up well with the nucleosomal signal apparent in MNase-seq data (Figs. 2a, b). The nucleosome predictions of COMPETE (Fig. 2e) are more diffuse, which is understandable because it relies entirely on sequence information, and nucleosomes have only weak and periodic sequence specificity. Because of a lack of chromatin accessibility data, COMPETE fails to identify the clear nucleosome depleted regions (all throughout the genome, as seen in Figs. 3a, b), as a result of which it fails to recognize the two Abf1 binding sites known to exist in this locus (Fig. 2e) [15]. In contrast, RoboCOP utilizes the chromatin accessibility data to accurately learn the nucleosome positions and the annotated Abf1 binding sites (Fig. 2c).

2.4 Predicted Nucleosome Positions

Nucleosomes have weak sequence specificity and can adopt alternative nearby positions along the genome. 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

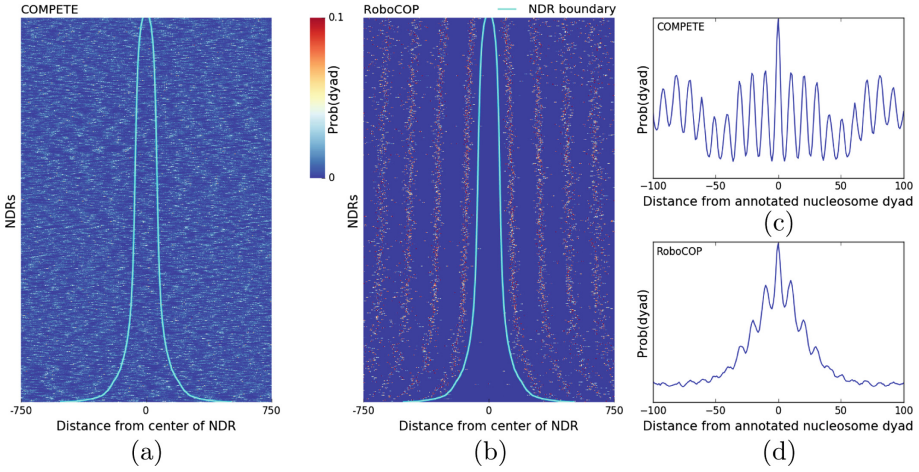


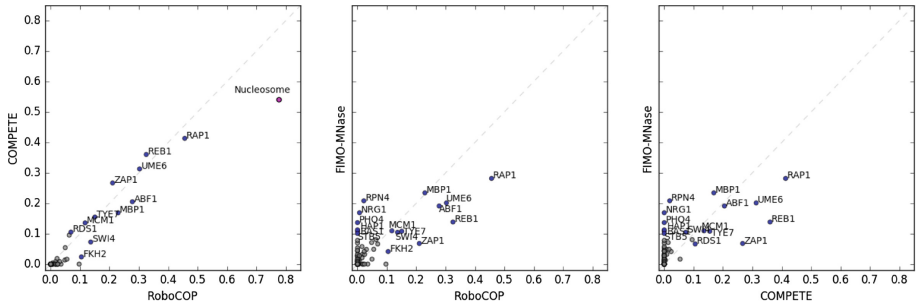
Fig. 3. Probability that positions around NDRs correspond to a nucleosome dyad, as computed by (a) COMPETE and (b) RoboCOP. Cyan lines depict experimentally determined NDR boundaries [5]. Note that Prob(dyad) computed by RoboCOP is appropriately almost always zero within NDRs, unlike COMPETE. Aggregate Prob(dyad), as computed by (c) COMPETE and (d) RoboCOP across all annotated nucleosome dyads [2]. Note that Prob(dyad) computed by RoboCOP appropriately peaks at annotated dyads.

occupancy, we can plot the probability of a nucleosome dyad, Prob(dyad), around annotated nucleosome locations [2]. We find that the RoboCOP dyad score peaks precisely at the annotated dyads (Fig. 3d), and decreases almost symmetrically in either direction. In contrast, COMPETE does not provide accurate location predictions (Fig. 3c); the oscillatory nature of the score reported by COMPETE reflects the periodic dinucleotide sequence specificity model for nucleosomes, and does not correspond well with actual nucleosome locations. When evaluated genome-wide using an F1-score (Fig. 4), the nucleosome positions called by RoboCOP are far more similar to the nucleosome annotations in [2] than are the ones called by COMPETE, which turn out to be not much better than random.

2.5 Predicted TF Binding Sites

MNase-seq is primarily used to study nucleosome positions; at present, no methods exist to predict TF binding sites from MNase-seq. It is challenging to extract TF binding sites from the noisy signal of the shortFrags generated by MNase digestion. TFs can sometimes be bound for an extremely short span of time [21] in which case the entire region could be digested by MNase, leaving behind no shortFrags signal. Nevertheless, MNase-seq data has been reported to provide evidence of TF binding [10], so we explore how well RoboCOP is able to identify TF binding sites. When we compare TF binding site predictions made by RoboCOP to predictions made by COMPETE, we see consistent but slight improvement in F1-score with RoboCOP (Fig. 4a). As a baseline, we compare these results to an approach we call FIMO-MNase, in which we simply run FIMO [8] around the peaks of midpoint counts of shortFrags. We find both RoboCOP and COMPETE are better than FIMO-MNase (Figs. 4b, c). Abf1, Reb1, and Rap1 have the most precise annotation datasets, and for these TFs in particular, both COMPETE and RoboCOP make better predictions. Overall, the highest F1-score is for Rap1 binding site predictions made by RoboCOP.

Although RoboCOP predicts the binding of a set of 150 TFs, we can only validate the binding sites of 81 of them, given available X-ChIP-chip [15], ChIP-



(a) (b) (c)

Fig. 4. Comparisons of F1-scores of TF binding site predictions by (a) RoboCOP and COMPETE, (b) RoboCOP and FIMO-MNase, and (c) COMPETE and FIMO-MNase. TFs with F1-score less than 0.1 in both methods of any given scatter plot are colored gray.

exo [19], and ORGANIC [13] datasets (Table S1). We have more precise binding sites from ChIP-exo and ORGANIC experiments for Abf1, Rap1, and Reb1. In addition, the binding sites in X-ChIP-chip data for many TFs were generated under multiple conditions [9] (Table S1) and the conditions are not specified for the reported annotations. This makes the X-ChIP-chip dataset fairly unreliable for validation purposes. In Fig. S3, we plot Venn diagrams showing the number of overlaps in the computed binding sites with the annotated binding sites for all three methods (`RoboCOP`, `COMPETE`, FIMO-MNase) and find that both `COMPETE` and `RoboCOP` have high false positives for certain motifs that are AT-rich such as Azf1 and Sfp1. Since the yeast genome is AT-rich and the `shortFrags` signal is noisy, any enrichment of the midpoint counts could be identified as a potential binding site. We believe prior knowledge about the occupancy of the TFs could yield higher accuracy.

2.6 `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 subjected to cadmium stress for 60 minutes. We run `RoboCOP` separately on two MNase-seq datasets: one for a cell population before it was treated with cadmium and another 60 minutes after treatment. Cadmium is toxic to the cells and activates stress response pathways. Stress response related genes are heavily transcribed under cadmium treatment, while ribosomal genes are repressed [12]. We use RNA-seq to identify the 100 genes most up-regulated (top 100) and the 100 most down-regulated (bottom 100). As a control, we choose 100 genes with no change in transcription under cadmium treatment (mid 100) (see Table

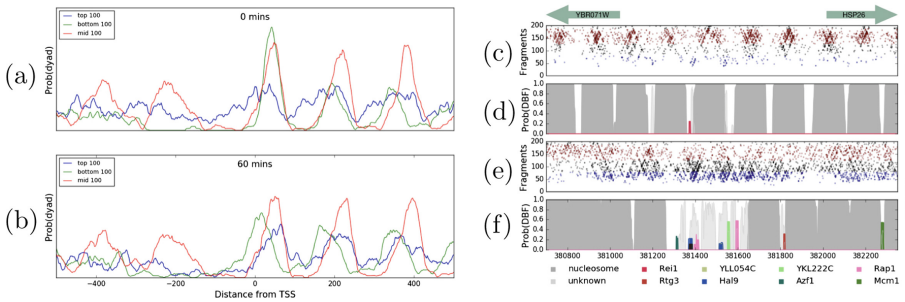


Fig. 5. Aggregate of Prob(dyad) computed by `RoboCOP` around the TSSs of genes most up-regulated (blue), most down-regulated (green), and unchanged in transcription (red), (a) before and (b) 60 min after treating cells with cadmium. After treatment, we see the +1 nucleosome closing in on the promoters of repressed genes (green) but opening up the promoters of highly transcribed genes (blue). MNase-seq fragment plot and `RoboCOP`-predicted occupancy profile of HSP26 promoter at chrII:380700–382350, (c, d) before and (e, f) after treatment with cadmium. HSP26 is highly expressed under cadmium stress, and its promoter exhibits much TF binding after treatment, most prominently by Rap1, known to bind the HSP26 promoter under stress response.

S2 for the three gene lists). Plotting the RoboCOP-predicted Prob(dyad) around the TSSs of the three gene groups, we notice that the nucleosomes in the top 100 genes are generally less well-positioned in comparison to the other groups of genes (Fig. 5a, b). Because of the uncertainty in the nucleosome positions of the top 100 genes, Prob(dyad) does not have any sharp peaks (blue curve in Fig. 5a, b). On the other hand, Prob(dyad) has sharp peaks indicating well-positioned nucleosomes for bottom 100 and mid 100 genes (green and red curves in Fig. 5a, b). This suggests RoboCOP-predicted Prob(dyad) can be used to classify 'fuzzy' or less well-positioned nucleosomes in the genome. Additionally, we see that the +1 nucleosomes of the top 100 genes move downstream after treatment with cadmium, thereby opening up the promoter region. In contrast, the +1 nucleosomes of the bottom 100 genes move upstream and close in on the promoter region to repress transcription.

HSP26, a key stress response gene, is among the top 100 most up-regulated genes. We can use RoboCOP to study how the chromatin landscape changes in the HSP26 promoter under cadmium stress. In Figs. 5c–f, 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 identifies multiple potential TF binding sites, most prominently for Rap1. Rap1 is known to relocalize to the promoter region of HSP26 during general stress response [17]; antibody-based methods could be used to validate whether Rap1 binds in the HSP26 promoter under cadmium treatment in particular.

In comparison, COMPETE fails to capture the dynamics of chromatin occupancy because it does not incorporate chromatin accessibility information into its model. We ran COMPETE with the RoboCOP-trained DBF weights for the two time points of cadmium treatment and found that COMPETE generates binding landscapes for the two time points that are nearly identical (Fig. S4). This is a key difference between RoboCOP and COMPETE: being able to incorporate experimental chromatin accessibility data allows RoboCOP to provide a more accurate binding profile for cell populations undergoing dramatic chromatin changes.

The above analysis highlights the 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. 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 (Fig. 5a), and to focus on specific genomic loci to understand their chromatin dynamics (Fig. 5b).

3 Methods

3.1 RoboCOP Model Structure and Transition Probabilities

RoboCOP is a multivariate hidden Markov model (HMM) for computing a genome-wide chromatin occupancy profile using nucleotide sequence and epigenomic accessibility data (here MNase-seq) as observables. The HMM structure

has been adapted from [23]. Let the number of TFs be K. Let $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K$ denote the models for the 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 [25]; suggestively, let $\boldsymbol{\pi}_0$ denote this model. In summary, we have a total of $K+2$ DBFs in the model. We use a central non-emitting ('silent') state to simplify state transitions among the DBFs in the model. The HMM may transition from this central non-emitting state to any one of the DBF models (including for unbound DNA); at the end of each DBF, the HMM always transitions back to the central silent state (Fig. S5). 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 $\{\alpha_0, \ldots, \alpha_{K+1}\}$.

Each hidden state represents a single genome coordinate. 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 (Fig. S6a). 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 1 between consecutive hidden states. On reaching the final hidden state of either motif, the HMM transitions back to the central silent state with probability 1. 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 (Fig. S6b). 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:

- $P(z_{g+1} = \pi_{k,l+1} | z_g = \pi_{k,l}) = 1$ 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.
- $P(z_{g+1} = \pi_{k_1,1} | z_g = \pi_{k_2,L_{k_2}}) = P(z_{g+1} = \pi_{k_1,1}) = \alpha_{k_1}$. The transition probability to the first state of a DBF is a constant, independent of which DBF the HMM visited previously.
- $P(z_{g+1} | z_g) = 0$ for all other cases.

The HMM always starts in the central non-emitting state with probability 1; this guarantees that it cannot start in the middle of a DBF.

3.2 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. In our case, these observables are modeled as independent, conditioned on the hidden state, but adding dependence would be straightforward. 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 nucFrags $\{l_1, \ldots, l_G\}$, and (iii) midpoint counts of MNase-seq shortFrags $\{m_1, \ldots, m_G\}$. For any position g in the genome, the hidden state z_g is thus responsible for emitting a nucleotide s_g, a number l_g of midpoints of nucFrags, and a number m_g of midpoints of shortFrags (Fig. S1). Since these three observations are independent of one another given the hidden state z_g, the hidden states have an emission model for each of the three observables, and the joint probability of the multivariate emission is the product of the emission probabilities of the three observables.

For 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 [7] (except for Rap1, where we use the more detailed motifs in [19]). 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 [20]. 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 [23].

As described earlier, the two-dimensional MNase-seq data is used to compute two one-dimensional signals. The midpoint counts of nucFrags are primarily used for learning nucleosome positions and the midpoint counts of shortFrags are used for learning the 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, $NB(\boldsymbol{\mu}_{nuc}, \phi_{nuc})$, 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 (Fig. 1b), we model each of the 147 positions separately. To obtain $\boldsymbol{\mu}_{nuc}$ and ϕ_{nuc}, we collect the midpoint counts of nucFrags in a window of size 147 centered on the annotated nucleosome dyads of the top 2000 well-positioned nucleosomes [2] and estimate 147 NB distributions using maximum likelihood estimate (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 (Fig. S7). To compute $NB(\mu_{l_b}, \phi_{l_b})$, we estimate an NB distribution for the midpoint counts of nucFrags at the linker regions of the same set of 2000 nucleosomes using MLE. The linker length is chosen to be 15 bases long on either side of the nucleosome [5].

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 at

the TF binding sites, while the other, $NB(\mu_{m_b}, \phi_{m_b})$, explains the counts elsewhere. To estimate the parameters of $NB(\mu_{TF}, \phi_{TF})$, we collect the midpoint counts of `shortFrags` at the annotated Abf1 and Reb1 binding sites from [15] and fit an NB distribution using MLE. A quantile-quantile plot shows the NB distribution to be a good fit (Fig. S8). 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. For parameterizing $NB(\mu_{m_b}, \phi_{m_b})$, we compute the midpoint counts of `shortFrags` at the same linker regions used earlier and estimate the NB distribution using MLE.

3.3 `RoboCOP` Parameter Updates

The transition probabilities between hidden states within a DBF can only be 0 or 1 (except for the two transition probabilities from each TF model's first, non-emitting state to either its forward or reverse motif, which are 0.5). Consequently, only the transition probabilities $\{\alpha_0, \ldots, \alpha_{K+1}\}$ from the central silent state to the first state of each DBF are unknown. We initialize these probabilities as described below, and then iteratively update them using Baum-Welch until convergence to a local optimum of the likelihood.

To initialize the probabilities, we assign weight 1 to the 'empty' DBF (representing an unbound DNA nucleotide) and 35 to the nucleosome. To each TF, we assign a weight which is that TF's dissociation constant K_D (or alternatively, a multiple thereof: $8K_D$, $16K_D$, $32K_D$, or $64K_D$). Finally, we transform these weights into a proper probability distribution to yield the initial probabilities.

To update α_k, the transition probability from the central silent state to the first hidden state $\pi_{k,1}$ of DBF k, we compute:

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

Here, $\boldsymbol{\theta^*}$ represents all the model parameters. We find the likelihood converges within 10 iterations (Fig. S9) and the optimized transition probabilities for each DBF almost always converge to the same final values regardless of how we initialize the weights (Fig. S10). We find convergence is faster for most DBFs when we initialize TF weights to K_D rather than multiples thereof (Fig. S10).

We do not use any prior information about the transition probabilities of the DBFs. We find that a few TFs such as Azf1 and Smp1 can have a large number of binding sites in the genome that are potential false positives. To curb the number of binding site predictions for such TFs, we apply a threshold on the TF transition probabilities. The threshold δ is chosen to be two standard deviations more than the mean of the initial transition probabilities of the TFs (Fig. S11). Therefore, after the Baum-Welch step in every iteration, an additional modified Baum-Welch step is computed as follows:

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

Here n is the number of TFs that have transition probability more than δ. So, all the TFs whose transition probability would be more than δ are set instead to δ, and the remaining TFs have a regular Baum-Welch update of their transition probabilities. We find that this approach reduces the number of false positives (Fig. S12). Using an informed prior might be an alternative mechanism for yielding a more accurate binding profile for such TFs.

To ensure fair comparisons between `RoboCOP` and `COMPETE`, we ran `COMPETE` using the same parameters estimated by `RoboCOP`. Therefore, the output profiles of the two methods highlight the differences in the results that occur because of the inclusion of chromatin accessibility data.

3.4 Implementation Details for 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).

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 5000, with an overlap of 1000 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 space. In addition, we use only the longest chromosome (chrIV) to train DBF transition probabilities with Baum-Welch, and then undertake posterior decoding genome-wide.

3.5 TF and Nucleosome Predictions and Validation

We use posterior probabilities of TF occupancy from `RoboCOP` and `COMPETE` output to identify binding sites, calling all sites whose probability is at least 0.1. In the case of Rap1 which has multiple PWMs, the maximum probability among the PWMs is chosen at every position. The same comparison is applied when choosing between the forward and reverse complement of the motif. 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 use the sacCer2 (June 2008) genome version in our analyses.

We called nucleosomes from `RoboCOP` and `COMPETE` output using a greedy algorithm, as described previously [24]. Briefly, nucleosome dyads with decreasing probability were iteratively selected. A window of size 101 around the selected dyad was removed from future rounds of dyad selection (this window size was chosen to allow mild overlap between adjacent nucleosome locations). The nucleosome annotations in [2] contain 67548 nucleosomes. We selected the same number of top scoring nucleosomes from the output of `RoboCOP` and `COMPETE`. A nucleosome position was considered to be a true positive (TP) if the distance between the predicted and annotated dyad was less than 50 bases.

3.6 FIMO-MNase

MNase-seq midpoint counts of `shortFrags` (length less than 80) are smoothed using a window of size 21. Peaks are detected if they have a value greater than 2 with consecutive peaks being at least 25 bases apart. Peaks for midpoint counts of `nucFrags` are detected if they have a value greater than 1 and are at least 100 bases apart. To prevent nucleosomal peaks occluding peaks of `shortFrags`, peaks of midpoint counts of `shortFrags` within 60 bases of peaks of midpoint counts of `nucFrags` are removed. After these steps, we detect 4137 peaks of `shortFrags` genome-wide. FIMO [8] is run using PWMs from [7] on 50-bp windows centered on the peak sites with a p-value cutoff of 10^{-4}.

3.7 Data Access

MNase-seq and RNA-seq of yeast cells before and after cadmium treatment is available at https://doi.org/10.7924/r4hx1b43s. Code and supplementary material may be downloaded from https://github.com/HarteminkLab/RoboCOP.

4 Discussion

`RoboCOP` is a novel method that utilizes a multivariate HMM to generate a probabilistic occupancy profile of the genome by integrating chromatin accessibility data with nucleotide sequence. We choose to apply the model to the yeast genome because of the availability of high quality MNase-seq data and the small size of the genome, which simplifies computation. Chromatin accessibility data from MNase-seq, DNase-seq, and ATAC-seq are generally noisy, so it is a challenging task to infer precise genome-wide DBF occupancy from the data, particularly for TFs. While alternative approaches using peak identification or footprint identification followed by TF-labeling with FIMO [8] can offer some insight into protein-DNA binding, we observe that `RoboCOP` performs notably better, presumably because it considers all DBFs together in a joint model that incorporates the thermodynamic competition among DBFs (including nucleosomes).

RoboCOP improves upon COMPETE in a number of ways: it slightly improves TF binding site predictions, it markedly improves nucleosome positioning predictions, and it uses experimental data to learn DBF transition probabilities in a principled way. When these same transition probabilities are provided to COMPETE, its TF binding site predictions are similar to RoboCOP's because of the generally high sequence specificity of TFs, but its nucleosome position predictions are much worse because of the weak sequence specificity of nucleosomes. In future work, it might be possible to improve the learned transition probabilities further through the use of prior information.

Finally, we note that RoboCOP can be used to study the chromatin architecture of the genome under varying conditions, an important task to which COMPETE is unsuited. Because RoboCOP uses data to model a collection of DBFs competing to bind to the genome, we can observe dynamic levels of occupancy for different DBFs under different environmental conditions. Since gene expression also varies in response to changing environmental conditions, we believe RoboCOP will help elucidate how the dynamics of chromatin occupancy and the dynamics of gene expression interrelate.

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

References

1. Benner, P., Vingron, M.: ModHMM: A modular supra-Bayesian genome segmentation method. In: Cowen, L.J. (ed.) RECOMB 2019. LNCS, vol. 11467, pp. 35–50. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17083-7_3
2. Brogaard, K., Xi, L., Wang, J.P., Widom, J.: A map of nucleosome positions in yeast at base-pair resolution. Nature **486**(7404), 496–501 (2012)
3. Chen, K., et al.: DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. **23**(2), 341–351 (2013)
4. Chen, W., Liu, Y., Zhu, S., Green, C.D., Wei, G., Han, J.D.J.: Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. Nat. Commun. **5**(1), 4909 (2014)
5. Chereji, R.V., Ramachandran, S., Bryson, T.D., Henikoff, S.: Precise genome-wide mapping of single nucleosomes and linkers in vivo. Genome Biol. **19**(1), 19 (2018)
6. Ernst, J., Kellis, M.: ChromHMM: Automating chromatin-state discovery and characterization. Nat. Methods **9**(3), 215–216 (2012)
7. Gordân, R., Murphy, K.F., McCord, R.P., Zhu, C., Vedenko, A., Bulyk, M.L.: Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol. **12**(12), R125 (2011)
8. Grant, C.E., Bailey, T.L., Noble, W.S.: FIMO: Scanning for occurrences of a given motif. Bioinformatics **27**(7), 1017–1018 (2011)
9. Harbison, C.T., et al.: Transcriptional regulatory code of a eukaryotic genome. Nature **431**(7004), 99–104 (2004)

10. Henikoff, J.G., Belsky, J.A., Krassovsky, K., MacAlpine, D.M., Henikoff, S.: Epigenome characterization at single base-pair resolution. Proc. Natl. Acad. Sci. U. S. A. **108**(45), 18318–18323 (2011)
11. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., Noble, W.S.: Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods **9**(5), 473–476 (2012)
12. Hosiner, D., Gerber, S., Lichtenberg-Fraté, H., Glaser, W., Schüller, C., Klipp, E.: Impact of acute metal stress in Saccharomyces cerevisiae. PLoS ONE **9**(1), e83330 (2014)
13. Kasinathan, S., Orsi, G.A., Zentner, G.E., Ahmad, K., Henikoff, S.: High-resolution mapping of transcription factor binding sites on native chromatin. Nat. Methods **11**(2), 203–209 (2014)
14. Lee, W., et al.: A high-resolution atlas of nucleosome occupancy in yeast. Nat. Genet. **39**(10), 1235–1244 (2007)
15. MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., Fraenkel, E.: An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinform. **7**(1), 113 (2006)
16. Park, D., Morris, A.R., Battenhouse, A., Iyer, V.R.: Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. Nucl. Acids Res. **42**(6), 3736–3749 (2014)
17. JM, Platt, et al.: Rap1 relocalization contributes to the chromatin-mediated gene expression profile and pace of cell senescence. Genes Dev. **27**(12), 1406–1420 (2013)
18. Rhee, H.S., Bataille, A.R., Zhang, L., Pugh, B.F.: Subnucleosomal structures and nucleosome asymmetry across a genome. Cell **159**(6), 1377–1388 (2014)
19. Rhee, H.S., Pugh, B.F.: Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell **147**(6), 1408–1419 (2011)
20. Segal, E., et al.: A genomic code for nucleosome positioning. Nature **442**(7104), 772–778 (2006)
21. Sung, M.H., Guertin, M.J., Baek, S., Hager, G.L.: DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol. Cell **56**(2), 275–285 (2014)
22. Tarbell, E.D., Liu, T.: HMMRATAC: A Hidden Markov ModeleR for ATAC-seq. Nucl. Acid Res. **47**, e91 (2019)
23. Wasson, T., Hartemink, A.J.: An ensemble model of competitive multi-factor binding of the genome. Genome Res. **19**(11), 2101–2112 (2009)
24. Zhong, J., Luo, K., Winter, P.S., Crawford, G.E., Iversen, E.S., Hartemink, A.J.: Mapping nucleosome positions using DNase-seq. Genome Res. **26**(3), 351–364 (2016)
25. Zhong, J., Wasson, T., Hartemink, A.J.: Learning protein-DNA interaction landscapes by integrating experimental data through computational models. Bioinformatics **30**(20), 2868–2874 (2014)