ChromWave: Deciphering the DNA-encoded competition between transcription factors and nucleosomes with deep neural networks

Sera Aylin Cakiroglu[1,5], Sebastian Steinhauser[1], Jon Smith[2], Wei Xing[2], Nicholas M. Luscombe[1,3,4]

[1]Bioinformatics and Computational Biology Laboratory, The Francis Crick Institute, London, NW1 1AT, UK.

[2]Scientific Computing Science Technology Platform, The Francis Crick Institute, London, NW1 1AT, UK.

[3]UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK.

[4]Okinawa Institute of Science & Technology Graduate University, Okinawa 904-0495, Japan

[5]Correspondence to: aylin.cakiroglu@crick.ac.uk or sacakiroglu@gmail.com

Summary

Transcription factors (TFs) regulate gene expression by recognising and binding specific DNA sequences. At times, these regulatory elements may be occluded by nucleosomes, making them inaccessible for TF-binding. The competition for DNA occupancy between TFs and nucleosomes, and associated gene regulatory outputs, are important consequences of the cis-regulatory information encoded in the genome. However, these sequence patterns are subtle and remain difficult to interpret. Here, we introduce ChromWave, a deep-learning model that, for the first time, predicts the competing profiles for TF and nucleosomes occupancies with remarkable accuracy. Models trained using short- and long-fragment MNase-Seq data successfully learn the sequence preferences underlying TF and nucleosome occupancies across the entire yeast genome. They recapitulate nucleosome evictions from regions containing “strong” TF binding sites and knock-out simulations show nucleosomes gaining occupancy in the absence of these TFs, accompanied by lateral rearrangement of adjacent nucleosomes. At a local level, models anticipate with high accuracy the outcomes of detailed experimental analysis of partially unwrapped nucleosomes at the GAL4 UAS locus. Finally, we trained a ChromWave model that successfully predicts nucleosome positions at promoters in the human genome. We find that human promoters generally contain few sites at which simple sequence changes can alter nucleosome occupancies and that these positions align well with causal variants linked to DNase hypersensitivity. ChromWave is readily combined with diverse genomic datasets and can be trained to predict any output that is linked to the underlying genomic sequence. ChromWave's application is limited only by the user's imagination and availability of training data.

Key words

Chromatin-accessibility, deep learning, regulatory code, convolutional neural networks

1 Introduction

Transcription factors (TFs)and nucleosomes play critical roles in regulating gene expression. TFs recognise and bind short, specific DNA sequences and occupy thousands of discrete sites in the genome. Nucleosomes on the other hand - comprising a 147bp section DNA wound around a histone octamer - cover most of the genome with a preference for GC-rich sequence regions. With some exceptions, binding of a stretch of DNA by TFs and nucleosomes is generally considered to be mutually exclusive (Hayes and Wolffe, 1992; Zhu et al., 2018). Histones are evicted or repositioned to occlude or expose TF-binding sites. Thus, the competition between TF and nucleosome occupancies plays an important role in effecting gene regulatory outcomes (Almouzni and Wolffe, 1995; He et al., 2013; John et al., 2008, 2011; Kaplan et al., 2011; Kim and O'Shea, 2008; Lam et al., 2008; Mirny, 2009; Neph et al., 2012; Raveh-Sadka et al., 2012; Svaren et al., 1994). However, genome-wide understanding of DNA sequence information that determines the competition between these two sets of molecules remains elusive.

Existing sequence-based computational models successfully predicted either TF-binding (e.g. via position weight matrices; (Jayaram et al., 2016; Mathelier and Wasserman, 2013; Wasserman and Sandelin, 2004)) or nucleosome occupancies individually (e.g. using Hidden Markov Models, HMM; (Gupta et al., 2008; Kaplan et al., 2009; Peckham et al., 2007; Segal et al., 2006)). HMMs have also been applied to predict competitive multi-factor binding in yeast (Ozonov and van Nimwegen, 2013; Wasson and Hartemink, 2009; Zhong et al., 2014) as well as logistic regression (He et al., 2013). However, key underlying assumptions of HMMs are violated, notably that (i) all DNA-binding proteins and their binding motifs are known, (ii) binding sites and probability of binding are fully described by the motif alone, and (iii) the binding strength for each protein and each motif is known. HMMs do not allow for easy integration of various data types measuring different influences on TF and nucleosome occupancies (e.g. DNA-sequence, 3D conformation of the genome, histone marks, DNA methylation). Moreover, the computational effort required to compute the right concentrations for even just a few TFs is substantial. Such drawbacks limit the accuracy and application of these models.

Recent advances in the application of deep learning in genomics have shown great promise to model several signals (often simultaneously) from DNA sequence alone (Alipanahi et al., 2015; Angermueller et al., 2017; Avsec et al., 2021; Kelley et al., 2018; Wang et al., 2018; Zeng and Gifford, 2017). These results make a unified model that can integrate all these different data to predict gene regulation genome-wide conceivable. However, most existing approaches apply predictions to bins of 100 nucleotides or more, thus limiting their applications to interpreting the effects of local sequence changes such as single nucleotide polymorphisms.

Here, we present ChromWave, a flexible deep learning model to understand the effect of genetic variation on DNA binding at nucleotide-resolution. Our approach substantially improves the current gold standard of models and alleviates some of their constraining assumptions, such as the knowledge of all DNA-binding proteins together with their sequence specificity. Although we present ChromWave predicting TF and nucleosome occupancies jointly from DNA sequence, the framework is easily extendable to predict other data modalities (e.g. histone marks, DNA methylation and so on), or dynamically changing DNA binding profiles.

2 Results

2.1 The ChromWave deep-learning architecture

ChromWave is a deep-learning algorithm to predict competing TF and nucleosome-occupancies at nucleotide resolution on a genome-wide scale. It is based on Google Deepmind’s WaveNet model (van den Oord et al., 2016a) that interprets textual information from many adjacent positions to infer an audio output. The input to ChromWave is a genomic sequence and its outputs are one or more discretised binding profiles of the same length. The number and types of outputs can be varied (for example, TFs v nucleosomes, nucleosomes in different conditions). Predictions are conditioned on several kilobases of sequence spanning the region of interest, meaning that nucleotides surrounding a site are taken into account. Models are readily trained using data from experimental measurements of genomic occupancies (such as ChIP-seq, CUT&RUN, MNase-seq or ATAC-seq): data can be presented as processed log-ratio values or raw count data with minimal preprocessing (Methods).

ChromWave accepts large sequence regions as input (Figure 1A; “Input”); in the example below, training was performed using 5 Kb sequences and predictions were made across whole chromosomes. The model applies one layer of convolutions to transform the DNA sequence into a series of vectors of the same length (Figure 1B; ”Convolution Layers”). This first layer of convolution learns patterns in the DNA sequence that can be considered to be motifs that are important for model predictions; we refer to these learned patterns as motif filters (Alipanahi et al., 2015). Transfer-learning can be applied at this stage - for instance, to incorporate models trained on another data set or on data from another organism - by combining the first convolutional layer with a pre-trained layer and concatenating their outputs.

Next, to share information across neighbouring regions, we apply multiple layers of dilated convolution by stacking residual blocks (Figure 1C; “Residual Blocks”); these can be considered the main building blocks of ChromWave. A residual block consists of two dilated convolutional layers with exponentially dilating filters, followed by element-wise multiplication. The first residual block accepts the concatenated output of the first convolutional layer(s) as input and subsequent residual blocks accept the output of the previous one as input. In this way, local features in lower dilated convolutional layers are accumulated sequentially to capture dependencies between distal sequence locations up- and downstream.

Each residual block produces two outputs: (i) a feature map from the previous residual block to be used as input to the next one and (ii) a skip connection that is used to calculate the loss function (for the input batch) once all the residual blocks have been processed. Within a residual block, the input passes through a gated activation z=sigmoid(x) ⊙ tanh(x), where ⊙ denotes the element-wise multiplication (similar to the LSTM gated activation units used in PixelCNN (van den Oord et al., 2016b)). The gated activation output is finally summed element-wise with the block-input to form the residual. Residuals and skip connections allow faster convergence and stable training for deeper networks.

Finally, for each output profile, we apply a width-one convolutional layer with a Rectified Linear Unit (ReLU) activation (Figure 1D; “Output”). ChromWave outputs a categorical distribution over the discretized binding profile values with a softmax layer, optimized to maximize the negative log-likelihood (for example, multi-class cross-entropy) of the data with respect to the parameters. In the case of multiple output profiles, the loss function is the sum of the individual cross-entropies for each profile. We tune hyperparameters on a validation set and we can easily gauge if the model is over- or underfitting, while we use a held-out test set for final validation after training. ChromWave predicts binned read counts. However, the distribution of bins tends to be highly imbalanced, with most representing the genomic average. To deter the model from defaulting to the most frequent bins instead of attempting to predict rarer ones with very high or very low read counts, we apply a weighted multi-class cross-entropy using the median/frequency per class as weight (Methods).

2.2 ChromWave performs a nucleotide-resolution prediction of TF and nucleosome occupancies across the whole yeast genome

We trained a model of TF and nucleosome occupancies for the yeast genome using paired-end MNase-seq data from (Henikoff et al., 2011). The dataset contains multiple fragment lengths (1-200bp) reflecting the diverse footprint sizes of the bound proteins and protein complexes. We separated fragments into long (140–200 bp) and short (1-80bp) groups, corresponding to nucleosome and TF occupancies respectively; the short fragments do not distinguish between TF types. We used the numbers of read fragments that cover each nucleotide position as the measure of nucleosome and TF occupancies (Zhong et al., 2014).

Among numerous alternative transfer-learning strategies tested (Methods), we selected a binary classification convolutional neural network that applies 256 convolutional filters of kernel size 6 to learn nucleosomal sequence patterns around dyads mapped by chemical cleavage (Brogaard et al., 2012). The best ChromWave model includes these 256 pre-trained convolutional filters and another 256 trainable convolutional filters with kernel size 16. 56 filters were initialised with the weights from the position weight matrices of 28 known nucleosome-displacing transcription factors and their reverse complements: Abf1, Cbf1, McM1, Rap1, Reb1, Orc1, Asg1, Azf1, Bas1, Ecm22, Ino4, Leu3, Rfx1, Rgm1, Rgt1, Rsc3, Sfp1, Stb4, Stb5, Stp1, Sum1, Sut1, Tbf1, Tbs1, Tea1, Uga3, Ume6, and Urc2 (Yan et al., 2018).

Figure 2A provides the first view of ChromWave's prediction in two example regions. The model has produced two output profiles: one for nucleosome occupancies and another for TF-binding. The two examples illustrate the excellent agreement of the predicted occupancies with the MNase-seq signal. To measure the model's performance across the entire yeast genome, we compute the geometric mean of Pearson's correlation coefficients between the predicted occupancies and smoothed MNase-seq read counts for nucleosomes and TFs (Figures 2B and S1A). It is immediately apparent that ChromWave achieves an excellent fit across all chromosomes in both the TF and nucleosome profiles (genome-wide Pearson correlation of 0.65 and 0.49 for TFs and nucleosomes respectively) with no apparent biases towards specific chromosomal regions such as centromeres and telomeres (Figures 2B and S1A). Black regions in Figures 2B and S1A indicate regions where no MNase-seq reads were mapped and thus no observed profile could be compared against. However, ChromWave imputes these regions for both TFs and nucleosomes with highly plausible binding profiles (Figure S1B).

2.3 ChromWave successfully predicts the competition in DNA occupancies between TFs and nucleosomes

Figure 2A provides encouraging examples that ChromWave is able to model the competition for DNA-binding between nucleosomes and TFs. The tracks show several nucleosome-free regions (5' NFRs) just upstream of transcriptional start sites (TSSs; indicated by stippled boxes). These regions are known distinguishing features between *in vitro* and *in vivo* nucleosome occupancy maps that arise through binding of TFs and transcription initiation factors. To assess this more generally, we examined occupancy patterns in the promoters of protein-coding genes: a meta-profile of TF and nucleosome occupancies demonstrates that ChromWave accurately captures their profiles around TSSs (Figure 2C).

To assess whether ChromWave has learned a general promoter architecture rather than an implicit competition between TFs and nucleosomes, we tested ChromWave's predictions against genome-wide *in vivo* nucleosome dyads identified in a independent chemical cleavage (Chereji et al., 2018) and ATAC-Seq studies (Schep et al., 2015). Whereas ATAC-Seq lacks coverage in heterochromatin (and thus, defined dyads are located in accessible chromatin only), chemical cleavage maps dyads genome-wide. ChromWave correctly predicts nucleosome occupancy when comparing its predictions with derived dyads from both datasets, including the phasing of nucleosomes on either side (Figure 2D). Notably, there is a depletion of TF signal directly on the dyad, indicating that ChromWave has learned that either TFs or nucleosomes can be bound at a DNA-site at a time.

2.4 ChromWave learns known and unknown regulatory motifs associated with DNA occupancy competition in S.cerevisiae

ChromWave's first convolutional layer has learned the DNA sequence patterns that are important in predicting nucleosome and TF occupancies. Since the short fragments in the input MNase-seq data do not distinguish between TF types, we decided to inspect the learned sequence motifs to see if we could match the predicted occupancy positions to specific TFs. We examined the first convolutional layers of ChromWave that take the one-hot encoded DNA-sequences as their input. Each filter can be summarised as a conventional position frequency matrix and visualised as a seqlogo plot as previously described (Alipanahi et al., 2015; Angermueller et al., 2017). We searched these matrices against known yeast TF motifs (Fornes et al., 2020; Gupta et al., 2007; Hume et al., 2015; MacIsaac et al., 2006; Pachkov et al., 2013; Zhu and Zhang, 1999). In total, 36 of 256 filters matched known motifs (q-value < 0.05); hence, not all 56 filters that were initialised with TF motifs and additional filters were augmented during training to capture information that is useful for the prediction task . As a measure of importance, we set each filter matching a TF motif o to zero and computed the maximal increase ('gains') and decrease ('loss') in predicted TF occupancies within 300bp windows compared with predictions of the full model. A high importance value indicates a large contribution to predictions: a 'loss' value corresponds to a loss of a predicted TF peak upon deletion of the filter, whereas a 'gain' represents an increase in predicted occupancy.

Many of the matched and high-scoring motifs correspond to those of sequence-specific TFs that regulate nucleosome occupancy in 5' NFRs such as Reb1 and Abf1, as well as poly-(d:A,d:T) stretches (Figure 3A). These were also identified as some of the best-learned motifs of the convolutional layer (measured as the information content in the derived motif) together with several unknown motifs capturing lower order sequence composition such as GC content. We also found additional motifs associated with TF binding that were deemed unmatched including several degenerated versions of the Reb1, the Rap1 motif, poly-(d:A,d:T) stretches, and TATA-box motifs.

2.5 ChromWave uses regulatory motifs to model the competition between transcription factors and nucleosomes

In addition to chromatin remodelers, a class of abundant and essential TFs including Abf1, Rap1, Reb1 and Cbf1 drive promoter nucleosome exclusion around TSSs (Badis et al., 2008; Ganapathi et al., 2011; Hartley and Madhani, 2009; Kent et al., 1994; Tsankov et al., 2011, 2010; Yarragudi et al., 2004). ChromWave's learned sequence patterns include motifs for these TFs and these filters have strong influence on ChromWave's predictions (Figure 3A; Cbf1 did not have a good match).

To assess ChromWave's abilities to model the competition between nucleosomes and specific TFs, we extracted previously published binding sites for Abf1, Rap1, Reb1 and Cbf1 derived from ChIP-exo assays (Rossi et al., 2018) and aligned ChromWave's outputs (Figure 3B). ChromWave predicts remarkably accurate nucleosome exclusion and TF occupancies at these sites with the largest effects seen for Rap1 and Reb1(Figure 3B). To understand how ChromWave relies on the learned TF motifs for its predictions, we set the weights of filters that had learned the motifs for Abf1, Rap1, Reb1, or Cbf1 individually to zero to mimic knock-out mutants. Figure 3C displays the difference in predictions between the mutant and wild-type conditions. For all four TFs, ChromWave predicts a loss of TF-occupancies and a corresponding increase of nucleosome occupancies. Taken together, these results indicate that ChromWave has learned that the competition between nucleosomes and TFs is encoded in the underlying DNA-sequence.

2.6 ChromWave computationally recapitulates an experimental study of an unwrapped nucleosome between the Gal1 and Gal10 genes

Given these encouraging genome-wide results, we set out to investigate ChromWave's performance at a detailed resolution on remodelled nucleosomes. One of the most abundant remodelers in yeast is Remodeling the Structure of Chromatin (RSC). In the Henikoff training data, a RSC-nucleosome complex is identified by shorter reads than for nucleosomes alone, as the DNA is unwound and becomes more susceptible to MNase digestion. Predicted occupancies for these remodelled nucleosomes thus are likely to be present in our TF profiles rather than the nucleosome profiles. It is therefore not straightforward for ChromWave to identify these pile-ups of shorter reads as nucleosome peaks.

We focused on the GAL4 UAS in the GAL1-GAL10 region described as containing a partially unwrapped nucleosome which is remodelled by RSC (Floer et al., 2010). This nucleosome is apparent in the Henikoff MNase-seq dataset obtained at shorter digestion times (Henikoff et al., 2011). This peak has nearly disappeared in the 20-min digestion sample which we used for training ChromWave, whereas a peak presumably representing RSC or the RCS-nucleosome complex appears in the TF profile. As expected, ChromWave's predictions on the yeast regulatory locus GAL4 UAS follow closely the training data: it predicts a broad TF peak representing the RSC-nucleosome complex directly in the UAS with strongly positioned nucleosomes either side (Figure 4A).

The locus encodes three Gal4-binding sites as well as three putative binding sites of the RSC complex (Floer et al., 2010). To quantify how the underlying motifs and the surrounding sequence influences ChromWave's prediction, we examined the changes in prediction that arose upon perturbing the input sequence through *in silico* mutagenesis. At every nucleotide position in the GAL4 UAS, we considered mutations to each of the three other bases and computed changes in the predicted TF and nucleosome occupancies. For each position and each possible base, we then summed across all changes in the GAL4 UAS (Figure 4B).

The largest effect is seen in a section region harbouring the 5' most Rsc3 motif indicated in Figure 4B. Mutating the bases of the Rcs3 consensus motif contributed to the most extreme prediction changes in the TF profile: targeted disruption of the RSC motif by reducing the GC-content of the locus decreased the TF prediction in the shaded region dramatically. This putative binding site is also where RSC shows its maximal cross-linked ChIP signal (Figure 4D) (Floer et al., 2010).

Next, we simulated the deletion of a 61bp sequence containing this RSC3 binding motif performed by Floer (Floer et al., 2010). ChromWave correctly predicts a loss of TF occupancy at the deleted site, and introduces a new narrower peak covering the three GAL4 motifs. There is a dramatic increase in a narrow and strongly positioned nucleosome on the UAS, as well as nucleosomes encroaching from either side of the UAS (Figure 4C). These nucleosomes now cover hypersensitive sites in wild-type that were reported to be increasingly occluded by nucleosomes upon RSC inactivation (Floer et al., 2010). Comparison with the molecular measurements by Floer highlights the remarkable accuracy of the ChromWave predictions (compare Figure 4D showing results from (Floer et al., 2010)).

2.7 Using ChromWave to quantify sequence patterns that differentiate between *in vitro* and *in vivo* nucleosome maps

Often only nucleosome-sized MNase-seq fragments are retained in publicly available datasets and it will not always be possible to train a ChromWave model on both TF and nucleosome profiles. Therefore we set out to assess the predictions that can be made using models trained on different MNase-seq datasets.

Kaplan et al. published *in vitro* and *in vivo* nucleosome maps for the yeast genome (Kaplan et al., 2009), the former gauging the effects of DNA sequence alone on nucleosome-positioning and the latter including the impact of trans-acting factors. Nucleosome occupancy was measured as the log-ratio between the number of MNase-seq reads (of length ~ 147 bp) at a nucleotide position and the genome-wide average. In the same study, a Hidden Markov Model (HMM) trained on *in vitro* data was shown to predict *in vivo* nucleosome organisation with good accuracy, albeit with differences at transcription start sites (TSSs).

We first trained an *in vitro* ChromWave model using the *in vitro* MNase-seq dataset. We then trained an *in vivo* ChromWave model using the *in vivo* dataset, but applying the pre-trained convolutional filters of the *in vitro* model as fixed filters (i.e. non-trainable), together with a further 256 trained convolutional layers. This approach gives us a direct way to detect differences in the learned sequence patterns between the *in vitro* and *in vivo* models.

Figure 5A shows examples of predictions from ChromWave's *in vitro* and *in vivo* models and the Kaplan HMM (Kaplan et al., 2009), each producing a single output profile representing nucleosome occupancies. Overall, there is excellent agreement between all three model predictions and the MNase-seq data, with expected differences between the *in vivo* and *in vitro* profiles around TSSs (stippled boxes). Figure 5B gives an overview of ChromWave's predictions across the entire genome: the predictions show exceptionally good agreement with the observed data across all chromosomes with no apparent biases towards specific regions such as centromeres and telomeres (average Pearson correlation with observed data of 0.9 and 0.89 for *in vitro*- and *in vivo* nucleosomes respectively). The lower correlation of the *in vitro* prediction on chromosome 10 is likely caused by artefacts in the original MNase-seq data (Figure S2B) and has been reported previously for other nucleosome positions models (Liu et al., 2016; Zhang et al., 2009). We also compared the predicted profiles with independently generated *in vivo* ATAC-seq datasets (Schep et al., 2015)(Brogaard et al., 2012; Chereji et al., 2018)(Schep et al., 2015). Both ChromWave models perform better than the HMM, and the *in vivo* ChromWave model is far superior to the *in vitro* models (Figure S2A).

Finally, to test the accuracy of the models in predicting nucleosome positions, we aligned the ChromWave predictions to the nucleosome dyad positions defined by chemical cleavage (Brogaard et al., 2012; Chereji et al., 2018) or ATAC-Seq (Schep et al., 2015)). Both models learned the strongly positioned nucleosome at the dyad: the *in vivo* model produced a phasing pattern either side of the central nucleosome that is similar to the TF+nucleosome ChromWave model (Figures 5C and S2C). However, the *in vitro* model failed to capture the phasing pattern. This is in line with the original MNase-seq data (Figure S2D) previous experimental observations that phasing is lost if certain factors such as remodelers RSC and SWI/SNF and general regulatory factors Abf1 and Reb1 are absent (Gkikopoulos et al., 2011; Krietenstein et al., 2016; Zhang et al., 2011). This is also reflected in the accuracy of the NFR predictions around TSSs: this characteristic depletion in average nucleosome occupancy *in vivo* is correctly predicted by the *in vivo* ChromWave model, together with the phasing of nucleosome both up- and downstream of the NFR in contrast to both the HMM and the *in vitro* ChromWave model (Figure S2B).

2.8 Determinants of *in vivo* nucleosome occupancy encoded in the DNA sequence

The *in vivo* ChromWave model uses the pre-trained filters of the *in vitro* model as fixed convolutional filters; thus the trainable parameters have been used to learn subtle differences between the *in vivo* and *in vitro* nucleosome maps. To understand what additional information the *in vivo* model has learned, we inspected the convolutional filters that are unique to this model compared with the *in vitro* one.

We ranked the 256 *in vivo*-specific motif filters according to their importance, computed using their maximum output values (maximal activation) for each 2Kb DNA-sequence input (see *'Convolutional Layers'* in Figure 1 where the output of the filters when applied to a one-hot encoded DNA-sequence is annotated as a blue square before being concatenated). The maximal activation provides a position-independent measure of the occurrence of the learned motifs (or highly similar motifs) for each DNA input sequence. To interpret the impact of motifs on ChromWave's nucleosome predictions, we calculated correlations between the vectors of filter activation values for an input sequence and the predicted nucleosome occupancy. A positive correlation indicates that the presence of a sequence pattern matching the motif is favourable to nucleosome occupancy, whereas a negative correlation indicates that it is unfavourable. We used the mean of these correlations across all input sequences to interpret the effect of each filter across the whole genome.

In Figure 5D, we show a PCA plot of the maximal activations of the *in vivo*-specific filters across all DNA-sequence windows, with each point coloured according to their correlations. We can see that motifs with similar GC-content tend to co-occur in the same input sequences with similar maximal activations. There are two large clusters: AT-rich motifs which are negatively correlated with nucleosome occupancy, and GC-rich motifs that positively correlate with occupancy. Among the AT-rich group are known motifs for sequence-specific chromatin remodelers such as Stb3, Sfp1, Dig1, Rlm1, and several TATA-box motifs. Poly-(d:A,d:T) stretches have been reported to deter nucleosome formation due to the rigidity of the DNA and thus are predictive of nucleosome binding (Segal and Widom, 2009). GC-rich filters include motifs for TFs whose DNA-binding correlate with nucleosome occupancy (for example Hal9, Cbf1 and YLL054C; (Charoensawan et al., 2012) and other motifs predominantly found in the 5' NFRs, echoing many of the patterns learned by the TF+nucleosome ChromWave model (for example Gal80, Cbf1 and Pho2) (Figure 5D).

To compare the learned motifs across the different ChromWave models more systematically, we computed similarity scores and clustered the convolutional filters (excluding the pre-trained filters of each model) as previously described by Vierstra (Vierstra et al., 2020); Supplementary Table 2). 63 of the *in vivo*-specific 256 filters comprise similar motifs to those in the TF+nucleosome ChromWave model. Among the other 183 *in vivo*-specific filters, only 2 - both reflecting RSC motifs - clustered with similar motifs found in the *in vitro* filters. The other *in vivo*-specific filters included an additional RSC motif as well as sub-motifs of PHO4/RTG and MCM1/PDR1 (Figure S3).

Given the disparities in learned motifs, we hypothesized that the *in vitro* ChromWave model would not be able to correctly model nucleosome eviction at binding sites of general regulatory factors, whereas the *in vivo* model would be able to model this competition correctly. Indeed, when we aligned the predictions again around known binding sites of the Abf1, Rap1, Reb1 and Cbf1(Rossi et al., 2018) we observed a stark difference between the predictions of the two ChromWave models: the *in vitro* model predicts a nucleosome directly on the binding sites of the TFs, while the *in vivo* model was able to learn nucleosome depletion on the binding sites, albeit not as well as the ChromWave model trained jointly on TF and nucleosome data (Figure 5E).

In summary, although ChromWave when trained on *in vivo* nucleosome occupancy maps can recover *in vivo* nucleosome positions to high accuracy, the missing information on TF binding hinders modelling the competition of nucleosomes with other sequence-driven binding events.

2.9 Predicting nucleosome occupancy in higher organisms with ChromWave

Nucleosome occupancy models trained on yeast data have been reported to produce predictions that correlate well with nucleosome maps from other organisms (Awazu, 2017; Gupta et al., 2008). However, we found that applying the *in vitro* and *in vivo* ChromWave models to 23,156 human promoter regions produced terrible agreement with MNase-seq datasets (Gaffney et al., 2012; Zhao et al., 2018), with average Pearson correlation coefficients of 0.002 and -0.0003, respectively.

We therefore trained a new ChromWave model to predict *in vivo* nucleosome occupancies around human promoter sequences (+/-1KB around TSSs; Methods). We used transfer learning across organisms: the first convolutional layers of a yeast *in vivo* nucleosome occupancy ChromWave model as a set of pretrained layers. The final model had only 32 additional trained convolutional layers and achieved excellent predictions across all promoters with similar performance across the training, test and validation sets (Figures 6A and S4A).

2.10 ChromWave learns known and unknown motifs associated with nucleosome binding and DNA-methylation in human promoter regions

Again we visualised the sequence motifs learned by the first convolutional layer and compared them to known motifs (JASPAR CORE vertebrates 2016, JASPAR POLII and HOCOMOCO v.10). We then repeated the PCA and correlation analysis described above (Section 2.8) to elucidate the co-occurrence of motifs and their impact on nucleosome occupancy predictions. As for the yeast *in vivo* model, ChromWave learned AT-rich motifs, including the TATA-box binding motif, that anti-correlate with nucleosome occupancies in human promoters, and GC-rich motifs that are positively correlated (Figure 6B).

A big difference between yeast and human genomes is the potential for cytosine methylation, which at promoters causes transcriptional repression. It is associated with certain sequence patterns; for instance, CG dinucleotides repeats (CpGs) and other motifs have been shown to be predictive of DNA methylation states (Angermueller et al., 2017). Methylation states affect TF and nucleosome occupancies, high density dinucleotide stretches such as CpG islands become densely occupied by nucleosomes upon methylation (Collings and Anderson, 2017; Collings et al., 2013). CpG-methylation has been shown *in vitro* to affect the binding of most human TFs to DNA, for some increasing and for some TFs decreasing binding (Yin et al., 2017). Although ChromWave was not given any information on methylation, the model has successfully learned the interplay between CpG patterns, methylation and nucleosome organisation (Figure 6B). For instance, learned motifs include a CpG pattern (CG) which is associated with DNA methylation in humans (Schübeler, 2015), and the motifs of TFDP1, SP3 and MNT, three proteins that are excluded upon methylation of open chromatin (Bartke et al., 2010).

We hypothesized that a model trained only on promoters might lack the information necessary to predict nucleosome occupancies in other types of genomic regions. For instance, CTCF-binding sites in intergenic regions are known to be depleted of nucleosomes (Kim et al., 2007). We asked ChromWave to make predictions around experimentally identified CTCF sites (Davis et al., 2018; Zhang et al., 2020), and in stark contrast to the MNase-seq data, the model placed nucleosomes directly over the CTCF sites irrespective of the binding strength of CTCF (Figure S4B).

2.11 Genetic components of promoter architecture in the human genome

To assess the impact of sequence changes at human promoters, we used ChromWave to prioritise single nucleotide polymorphisms (SNPs) by their predicted impact on nucleosome positioning.

First, we predicted the effects of sequence changes at *DNase I sensitivity quantitative trait loci* (dsQTLs; 1366 of 6057 fell within +/-1kb around TSS; Degner et al., 2012). These are genetic variants that modify chromatin accessibility as measured by DNase I sequencing. Although many dsQTLs did not affect predictions, we identified 30 promoters where ChromWave's chromatin accessibility predictions based on MNase-seq data were highly affected by dsQTLs (Figure 7A and Supplementary Table 3). We used the annotation of DNA-Hypersensitive Sites (DHS) from the same study (Degner et al., 2012) to annotate SNPs inside and outside of DHSs. Of the 30 dsQTLs that we identified to change chromatin accessibility, 7 were inside an annotated DHS, whereas the rest acted distally.

We focused on dsQTLs that changed ChromWave's prediction locally at the dsQTL itself (Δ prediction SNP-WT > 0 or Δ prediction SNP-WT < 0), as well as a dramatic effect on the overall nucleosome occupancy profile measured by the Pearson correlation coefficient (PCC) between the predicted WT and SNP-perturbed profiles (1-PCC>0.05). We identified 17 dsQTLs leading to gains in nucleosomes thus decreasing chromatin accessibility (1-PCC > 0.05 and Δ prediction SNP-WT > 0) and 13 dsQTLs causing losses (1-PCC > 0.05 and Δ prediction SNP-WT < 0).

The promoter of the ZBTB11-AS1 gene displays one of the largest changes in both the local prediction and the overall binding profile (1-PCC > 0.15, Δ prediction [C>T] - WT > 4). The predictions of the promoter sequence with different bases substituted at the dsQTL are shown in Figure 7B. The reference genotype (C) and a C>G change at the dsQTL yield the lowest predicted local nucleosome occupancy at the dsQTL, whereas C>A and C>T changes cause dramatic increases in predicted occupancies, with a strongly positioned nucleosomes directly on the dsQTL (Figure 7B, top panel). The PLEKHM3 promoter is an example of a loss in nucleosome occupancy following a single nucleotide change (Figure 7C): alteration from the reference genotype (A) to C or G at the dsQTL removes a weakly positioned nucleosome (1-PCC >0.09, Δ prediction [A>G] - WT < -3). An A>T change does not alter the profile (Figure 7C, top panel).

2.12 Changes in nucleosome-binding in human promoters induced by simple sequence changes are restricted to ‘hot spots’ which cover previously identified dsQTLs

To assess the impact of sequence changes in the whole promoter region on nucleosome occupancy, we computed the maximal gain and loss in prediction at each position given all possible base changes along the sequence for all promoter regions in the dataset.

The bottom panels in Figures 7B and 7C show these maximal gains and losses for the two example promoters of ZBTB11-AS1 and PLEKHM3. In both promoter regions, not all nucleosome position predictions are dynamic under sequence changes. Instead, the predicted decrease and increase in prediction is restricted to small regions in the promoters (some of which coincide with the position of the dsQTL), while most nucleosome occupancy predictions are robust to individual base changes.

These observations suggest that most nucleosome positions in human promoter regions are robust to sequence changes while some can be altered considerably through simple base-changes. Figure 7D shows individual and average profiles of maximal gains and losses of nucleosome occupancy prediction given all possible base-changes in all promoter regions in the dataset. The nucleosome occupancy of around two thirds of human promoters are affected by single nucleotide changes in 1000bp around the TSSs in either gain or loss (Figure 7D). These changes of occupancy are indeed not randomly distributed in the promoters. Instead, given the right single nucleotide changes (compared with wild-type), more nucleosomes bind close to and directly on the TSS (Figure 7D left). However, given the right single nucleotide changes (compared to wild-type) nucleosomes are lost just up- and downstream of the TSS, while nucleosomes directly on the TSS are more stable (Figure 7D right).

3 Discussion

We presented a deep learning framework called ChromWave to model DNA-binding profiles of nucleosomes and TFs simultaneously at nucleotide resolution to high precision and deployed it genome-wide in yeast across different data sets and experimental methods as well as in human promoter regions. ChromWave learned intricate sequence grammar of DNA accessibility and competition between DNA-binding proteins. ChromWave surpasses with its precision, scope and flexibility current state of the art methods to predict TF and nucleosome binding from DNA-sequence alone. We demonstrated and visualised how ChromWave identifies and uses specific sequence features that encode information for DNA-binding and the competition between nucleosomes and TFs.

We showed that ChromWave recapitulates computationally results from a previous study that was based on lab experiments (Floer et al., 2010) to determine TF and nucleosomes binding in the well-studied GAL1-GAL10 locus. Taken together, our results match the experimentally derived interpretation in (Floer et al., 2010): a small (partially unwound) nucleosome is positioned at the locus UASg by the RSC complex thereby facilitating Gal4 binding to its side. Upon deletion of the RSC binding site, nucleosomes encroach over the UAS and compete with Gal4 for binding. These results illustrate how ChromWave can be used to generate and test hypotheses *in silico* before testing these in the lab.

In the human genome, ChromWave lets us annotate dynamic hotspots where nucleosome occupancy is highly affected by simple base changes in the DNA-sequence compared with the rest of the promoter region. By comparing models that were trained on different types of data, we have qualified sequence patterns that allow ChromWave to correct *in vitro* nucleosome positions to *in vivo.* In addition, these experiments have shown that lacking TF-binding information severely impedes the capability of the model to predict nucleosome positions to a high precision and the models are unable to capture the competition between different DNA-binding factors.

Previous work to model TF- and nucleosome-binding at nucleotide-resolution has been constrained to the usage of HMMs in the yeast genome (Ozonov and van Nimwegen, 2013; Wasson and Hartemink, 2009; Zhong et al., 2014). An advantage of these models over ChromWave is that they can resolve which TF is predicted to be bound at a certain position based on the assumption that the TF has a well annotated motif. However, due to the computational complexity of determining concentrations for all used TFs in the model, these studies had to constrain their models to only a small set of TFs (158 TFs were used in (Ozonov and van Nimwegen, 2013), 89 in (Wasson and Hartemink, 2009), and only 42 in (Zhong et al., 2014)). Nevertheless, our work was inspired by these early studies; in fact we used the same training data and similar pre-processing methods as in (Zhong et al., 2014) albeit the former study was constrained to promoter regions. At the expense of the ability to distinguish different TF binding events, ChromWave can predict nucleosome- and TF-binding genome-wide, and can resolve binding events not directly explainable by an annotated motif.

Our work also builds on the success of recent deep learning methods that model DNA-binding and gene expression, e.g. (Alipanahi et al., 2015; Kelley et al., 2015, 2018; Quang and Xie, 2016). Most of these models perform binary classification tasks and predict whether a relatively short DNA-sequence (30bp-1000bp) will be bound by a TF or RNA-binding protein (Alipanahi et al., 2015), a nucleosome (Quang and Xie, 2016), or surrounds a DNase I hypersensitive site in different cell types (Kelley et al., 2015). However, these models are constrained to one short sequence at time and can not provide genome-wide profiles as they ignore competition between any DNA-binding factors. In contrast, (Kelley et al., 2018) developed the first deep learning model, called Basenji, that can model distal regulatory interactions to predict quantitative (as opposed to binary) genomic profiles such as read counts of DNase I seq, histone modification ChIP-seq, and expression data. ChromWave takes a similar approach by modeling read counts of MNase-seq data. Basenji can take 131kb regions as input which is much larger than the 2000bp-5000bp we chose for ChromWave, however this scale comes at the cost of resolution: while Basenji predicts one value for every 128bp bins (which already was a refinement over the 200bp bins of ChromHMM (Ernst and Kellis, 2017)), ChromWave makes a prediction for each individual nucleotide. This is an important improvement as most of our results on the intricate competition of TFs and nucleosomes display changes on just a few base-pairs with most TF binding sites predicted to be shorter than 15bp (Stewart et al., 2012). This resolution has only been achieved by a recently published model (BPNet) which shares many similarities in its architecture with that of ChromWave (Avsec et al., 2021). While BPNet predicts several TF binding profiles simultaneously and the study focused on TF binding motifs, nucleosome occupancies and their competition was largely ignored.

A main limitation of any model relying on training data restricted to certain regions in the genome is illustrated by our human promoter ChromWave model predicting nucleosomes at CTCF-sites: as we only trained in promoter regions, its genome-wide predictions cannot be trusted as it was not exposed to binding factors outside of promoter regions. To remedy this bias, ChromWave could be further trained on more DNA-sequences that span the (mappable) genome. Another approach may be using an ensemble of several ‘expert’ ChromWaves models that have learned specific sequence context (e.g ‘promoter regions’, ‘insulator-sites’,..). Either of these approaches can be realised with the ChromWave framework, however this observation highlights the importance of understanding the data and training details whenever models such as ChromWave (and their predictions) are being used.

Although we presented ChromWave as a model of different chromatin accessibility profiles, ChromWave is a flexible framework to model genome-wide chromatin-modification and readily extendable to other data such as 3D interactions, histone modifications and DNA-methylation, as well as transferable to mammalian genomes. As such, ChromWave is an unprecedented framework to study how genetic variability impacts DNA-binding, gene expression, histone marks, DNA-methylation, and 3D interactions (or any combinations thereof) simultaneously *in silico* at nucleotide-resolution. Other applications of ChromWave may be modelling time-dependent chromatin profiles changes, where we expect the environmental changes to have an effect on DNA-binding. As a proof of concept, we have successfully used ChromWave to model nucleosome profiles of young and old yeast (data not shown, but code and data are available on github https://github.com/luslab/ChromWave).Other applications of ChromWave may be the modelling of binding profiles during differentiation, or hormone changes.

In summary, ChromWave allows the prediction of TF and nucleosome binding simultaneously genome-wide at nucleotide resolution with high precision. The sequence-based approach allows the annotation of every mutation in the genome that influences chromatin accessibility through TF or nucleosome binding. ChromWave has the potential to provide important insights into the effect of genetic variation on transcriptional changes in complex diseases, at a certain time-point, but also during changes over time. As more and more non-coding variants are discovered in the human genome, these insights can provide us with important understanding how SNPs affect biological mechanisms and phenotypes.

Acknowledgements

This research was funded in whole, or in part, by the Wellcome Trust (FC010110; 215593/Z/19/Z). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC010110), the UK Medical Research Council (FC010110), and the Wellcome Trust (FC010110). NML is a Winton Group Leader in recognition of the Winton Charitable Foundation’s support towards the establishment of the Francis Crick Institute. NML is additionally funded by a Wellcome Trust Joint Investigator Award (215593/Z/19/Z) and core funding from the Okinawa Institute of Science & Technology Graduate University.

We thank Joe Brock (The Francis Crick Institute) for helping with illustrations, and Drs. Dean Plumbley (BenevolentAI), Joe Ledsam (Google DeepMind, Google Japan), Bernadino Romera-Paredes (Google DeepMind), Erik Pfeiffenberger (The Francis Crick Institute, BMW Group) and Prof. David Jones (The Francis Crick Institute, UCL) for many interesting discussions and their valuable advice and helpful comments regarding deep learning, neural networks and visualisation techniques.

Author Contributions

Conceptualization, S.A.C. and N.M.L.; Methodology, S.A.C. and S.S.; Software, S.A.C, S.S, J.S and W.X.; Validation, S.A.C and S.S; Formal Analysis, S.A.C., S.S. and N.M.L.; Investigation, S.A.C. and S.S.; Resources, W.X. and N.M.L; Writing – Original Draft, S.A.C. and N.M.L.; Writing – Review & Editing, S.A.C. and N.M.L; Visualization, S.A.C. and S.S.; Supervision, S.A.C. and N.M.L.

Declaration of Interests

The authors declare no competing interests.

References

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Almouzni, G., and Wolffe, A.P. (1995). Constraints on transcriptional activator function contribute to transcriptional quiescence during early Xenopus embryogenesis. EMBO J. *14*, 1752–1765.

Angermueller, C., Lee, H.J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. *18*, 67.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366.

Awazu, A. (2017). Prediction of nucleosome positioning by the incorporation of frequencies and distributions of three different nucleotide segment lengths into a general pseudo k-tuple nucleotide composition. Bioinformatics *33*, 42–48.

Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D., Gossett, A.J., Hasinoff, M.J., Warren, C.L., et al. (2008). A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol. Cell *32*, 878–887.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell *143*, 470–484.

Bergstra, J., Yamins, D., and Cox, D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In International Conference on Machine Learning, pp. 115–123.

Bergstra, J.S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, eds. (Curran Associates, Inc.), pp. 2546–2554.

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.

Charoensawan, V., Janga, S.C., Bulyk, M.L., Babu, M.M., and Teichmann, S.A. (2012). DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes. Mol. Cell *47*, 183–192.

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.

Chollet, F., and Others (2015). Keras.

Collings, C.K., and Anderson, J.N. (2017). Links between DNA methylation and nucleosome occupancy in the human genome. Epigenetics Chromatin *10*, 18.

Collings, C.K., Waddell, P.J., and Anderson, J.N. (2013). Effects of DNA methylation on nucleosome stability. Nucleic Acids Res. *41*, 2918–2931.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C., Dwight, S.S., Hitz, B.C., Karra, K., Nash, R.S., et al. (2014). The reference genome sequence of Saccharomyces cerevisiae: then and now. G3 *4*, 389–398.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492.

Floer, M., Wang, X., Prabhu, V., Berrozpe, G., Narayan, S., Spagna, D., Alvarez, D., Kendall, J., Krasnitz, A., Stepansky, A., et al. (2010). A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. Cell *141*, 407–418.

Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *48*, D87–D92.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of nucleosome positioning in the human genome. PLoS Genet. *8*, e1003036.

Ganapathi, M., Palumbo, M.J., Ansari, S.A., He, Q., Tsui, K., Nislow, C., and Morse, R.H. (2011). Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. Nucleic Acids Res. *39*, 2032–2044.

Gkikopoulos, T., Schofield, P., Singh, V., Pinskaya, M., Mellor, J., Smolle, M., Workman, J.L., Barton, G.J., and Owen-Hughes, T. (2011). A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. Science *333*, 1758–1760.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

Gupta, S., Dennis, J., Thurman, R.E., Kingston, R., Stamatoyannopoulos, J.A., and Noble, W.S. (2008). Predicting human nucleosome occupancy from primary sequence. PLoS Comput. Biol. *4*, e1000134.

Hartley, P.D., and Madhani, H.D. (2009). Mechanisms that specify promoter nucleosome location and identity. Cell *137*, 445–458.

Hayes, J.J., and Wolffe, A.P. (1992). The interaction of transcription factors with nucleosomal DNA. Bioessays *14*, 597–603.

He, X., Chatterjee, R., John, S., Bravo, H., Sathyanarayana, B.K., Biddie, S.C., FitzGerald, P.C., Stamatoyannopoulos, J.A., Hager, G.L., and Vinson, C. (2013). Contribution of nucleosome binding preferences and co-occurring DNA sequences to transcription factor binding. BMC Genomics *14*, 428.

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.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. *43*, D117–D122.

Jayaram, N., Usvyat, D., and R Martin, A.C. (2016). Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics *17*, 547.

John, S., Sabo, P.J., Johnson, T.A., Sung, M.-H., Biddie, S.C., Lightman, S.L., Voss, T.C., Davis, S.R., Meltzer, P.S., Stamatoyannopoulos, J.A., et al. (2008). Interaction of the Glucocorticoid Receptor with the Chromatin Landscape. Molecular Cell *29*, 611–624.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264–268.

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.

Kaplan, T., Li, X.-Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet. *7*, e1001290.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. (2004). EnsMart: a generic system for fast and flexible access to biological data. Genome Res. *14*, 160–169.

Kelley, D.R., Snoek, J., and Rinn, J. (2015). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks.

Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. *28*, 739–750.

Kent, N.A., Tsang, J.S., Crowther, D.J., and Mellor, J. (1994). Chromatin structure modulation in Saccharomyces cerevisiae by centromere and promoter factor 1. Mol. Cell. Biol. *14*, 5229–5241.

Kim, H.D., and O'Shea, E.K. (2008). A quantitative model of transcription factor-activated gene expression. Nat. Struct. Mol. Biol. *15*, 1192–1198.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Krietenstein, N., Wal, M., Watanabe, S., Park, B., Peterson, C.L., Pugh, B.F., and Korber, P. (2016). Genomic Nucleosome Organization Reconstituted with Pure Proteins. Cell *167*, 709–721.e12.

Lam, F.H., Steger, D.J., and O'Shea, E.K. (2008). Chromatin decouples promoter threshold from dynamic range. Nature *453*, 246–250.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Liu, G., Xing, Y., Zhao, H., Wang, J., Shang, Y., and Cai, L. (2016). A deformation energy-based model for predicting nucleosome dyads and occupancy. Sci. Rep. *6*, 24133.

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.

Mathelier, A., and Wasserman, W.W. (2013). The next generation of transcription factor binding site prediction. PLoS Comput. Biol. *9*, e1003214.

Mirny, L. (2009). Nucleosome-mediated cooperativity between transcription factors. Nature Precedings.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016a). WaveNet: A Generative Model for Raw Audio.

van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., and Kavukcuoglu, K. (2016b). Conditional Image Generation with PixelCNN Decoders.

Ozonov, E.A., and van Nimwegen, E. (2013). Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers. PLoS Comput. Biol. *9*, e1003181.

Pachkov, M., Balwierz, P.J., Arnold, P., Ozonov, E., and van Nimwegen, E. (2013).

SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res. *41*, D214–D220.

Peckham, H.E., Thurman, R.E., Fu, Y., Stamatoyannopoulos, J.A., Noble, W.S., Struhl, K., and Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. Genome Res. *17*, 1170–1177.

Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. *44*, e107.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A., and Segal, E. (2012). Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat. Genet. *44*, 743–750.

Rossi, M.J., Lai, W.K.M., and Pugh, B.F. (2018). Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. Genome Res. *28*, 497–508.

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.

Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326.

Segal, E., and Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr. Opin. Struct. Biol. *19*, 65–71.

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.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. *43*, W589–W598.

Stewart, A.J., Hannenhalli, S., and Plotkin, J.B. (2012). Why transcription factor binding sites are ten nucleotides long. Genetics *192*, 973–985.

Svaren, J., Klebanow, E., Sealy, L., and Chalkley, R. (1994). Analysis of the competition between nucleosome formation and transcription factor binding. J. Biol. Chem. *269*, 9335–9344.

Team BC, Maintainer BP (2019). TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s). R Package Version 3.4.6.

Tsankov, A., Yanagisawa, Y., Rhind, N., Regev, A., and Rando, O.J. (2011). Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. Genome Res. *21*, 1851–1862.

Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A., and Rando, O.J. (2010). The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol. *8*, e1000414.

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736.

Wang, M., Tai, C., E, W., and Wei, L. (2018). DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic Acids Res. *46*, e69.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. *5*, 276–287.

Wasson, T., and Hartemink, A.J. (2009). An ensemble model of competitive multi-factor binding of the genome. Genome Res. *19*, 2101–2112.

Yan, C., Chen, H., and Bai, L. (2018). Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. Mol. Cell *71*, 294–305.e4.

Yarragudi, A., Miyake, T., Li, R., and Morse, R.H. (2004). Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast Saccharomyces cerevisiae. Mol. Cell. Biol. *24*, 9152–9164.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*.

Zeng, H., and Gifford, D.K. (2017). Predicting the impact of non-coding variants on DNA methylation. Nucleic Acids Res.

Zerbino, D.R., Achuthan, P., Akanni, W., Ridwan Amode, M., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Research *46*, D754–D761.

Zhang, J., Lee, D., Dhiman, V., Jiang, P., Xu, J., McGillivray, P., Yang, H., Liu, J., Meyerson, W., Clarke, D., et al. (2020). An integrative ENCODE resource for cancer genomics. Nat. Commun. *11*, 3696.

Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., and Struhl, K. (2009). Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat. Struct. Mol. Biol. *16*, 847–852.

Zhang, Z., Wippo, C.J., Wal, M., Ward, E., Korber, P., and Pugh, B.F. (2011). A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science *332*, 977–980.

Zhao, Y., Wang, J., Liang, F., Liu, Y., Wang, Q., Zhang, H., Jiang, M., Zhang, Z., Zhao, W., Bao, Y., et al. (2018). NucMap: a database of genome-wide nucleosome positioning map across species. Nucleic Acids Res.

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.

Zhu, J., and Zhang, M.Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics *15*, 607–611.

Zhu, C., Byers, K.J.R.P., McCord, R.P., Shi, Z., Berger, M.F., Newburger, D.E., Saulrieta, K., Smith, Z., Shah, M.V., Radhakrishnan, M., et al. (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. *19*, 556–566.

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S.O., Nitta, K.R., Morgunova, E., Taipale, M., et al. (2018). The interaction landscape between transcription factors and the nucleosome. Nature.

Figure Legends

Figure 1. The ChromWave deep learning architecture to predict chromatin accessibility profiles.

Input. DNA sequences are one-hot coded to four rows representing A, C, G, and T before entering the model. The annotations of the transcription factor binding sites and transcriptional start site are examples indicated to help convey the reasons for the various elements of the architecture. **Convolutional Layers.** We first apply a convolutional layer followed by batch normalisation and ReLU layers several layers. If we used transfer-learning we also use a convolutional layer with pre-trained weights followed by a max-pooling layer before concatenating these with the trainable convolutional layer. **Residual Blocks.** To share information across large distances, we apply several layers of dilated convolutions in at least nine residual blocks. Each residual block contains two dilated convolutions with exponentially dilating filters (one gate, one filter) followed by element-wise multiplication. The gated activation output is (i) outputted as skip connections which are aggregated after all residual blocks, and (ii) as residual summed with the block input and used as input to the next residual block. **Output.** After the final residual block and aggregating the skip connections, for each output profile a 1x1 convolutional layer with ReLU activation is applied and passed to the final prediction layer where another 1x1 convolution layer makes predictions across the sequence. A prediction is made for each input position by the last layer through a softmax activation function and represents a bin of the discretised binding profiles. We compare these predictions to the experimentally observed profiles (the target) via the mutual information between observed and predicted classes, as well as the Pearson correlation coefficient (the latter is not used during the fitting of the model parameters with stochastic gradient descent with back propagation, but only for the selection of the best model during the hyperparameter optimization).The predicted (discretised) profiles are finally turned into continuous profiles by reversing the discretisation process used on the input data.

Figure 2. Nucleotide-resolution prediction of TF and nucleosome occupancies in the yeast genome.

A. Two representative examples of the predicted TF and nucleosome occupancy profiles from the training set (chromosome 16) and the unseen test set (chromosome 10). The profiles predicted by ChromWave are shown in black for comparison with the observed MNase-seq data shown in grey. **B.** Geometric means of the Pearson Correlation coefficients between predicted and observed TF and nucleosome occupancy genome-wide. Indicated are which chromosomes formed the validation and test sets (all other chromosomes formed the training set). Correlations were computed in sliding windows of size 300bp and then averaged over 300bp bins**.** **C.** Average predicted and observed TF and nucleosome occupancy in yeast promoters. Metaprofiles +/- 1KB of ChromWave's predictions and MNase-seq data (in Fragments Per Million, FPM) around all TSSs for predicted and observed TF and nucleosome occupancy. Each track was normalised by subtracting the genomic average before plotting. **D.** ChromWave's predictions of TF and nucleosome occupancy around experimentally defined nucleosome dydas. Heatmaps showing +/-1kb regions around nucleosome dyads called either on ATAC-seq (*top*, (Schep et al., 2015)) or defined by chemical cleavage (*bottom*, (Chereji et al., 2018)) data. The ATAC-seq heatmap was split according to four biological replicates (A, B, C, D). Predicted nucleosome occupancy (NUC) and TF-binding (TF) in FPM are shown as separate heatmaps next to each other. Average predictions are shown as metaprofiles on top of each heatmap. Each track was normalised by subtracting the genomic average before plotting.

Figure 3. ChromWave uses regulatory motifs to model the completion between transcription factors, chromatin remodelers and nucleosomes.

A. Importance of motif detectors associated with nucleosome-TF competition. Motif importance (setting associated filter to zero) was plotted against the sum of information content (IC) for each motif detector. Shapes indicate if the motif was annotated by TOMTOM (triangle, q<0.05) or de novo discovered (point,q>0.05). The color to gain or loss of TF binding prediction by setting the motif to zero. Selected motifs were annotated with their associated sequence logo. **B.** Nucleosome and TF occupancy per base-pair centered on TF binding sites defined by ChIP-exo. Heatmaps showing +/-1kb regions around ChIP-exo peak. Each heatmap panel is based on a different TF peak set from top to bottom: Abf1, Cbf1, Rap1 and Reb1. Average predictions are shown as metaprofiles on top of each heatmap. Heatmaps showing ChromWave's predictions (*left)* and MNase-seq data (*right)* of nucleosome (*NUC*) and TF (*TF*) occupancy as indicated (in Fragments Per Million, FPM). Each prediction track was normalised by subtracting the genomic average before plotting. **C.** Log2-fold change in nucleosome and TF occupancy predictions (in FPM) when ChromWave's fillters are perturbed per base-pair centered on TF binding sites defined by ChIP-exo. Meta-profiles showing average log2-fold changes in +/-1kb regions around ChIP-exo peak. Each panel is based on a different TF peak set from top to bottom: Abf1, Cbf1, Rap1 and Reb1. Log2-fold change was computed between the predictions of the full ChromWave model and a perturbed model where the convolutional filters corresponding to the respective TF-binding motifs (defined by TOMTOM annotation) were set to zero.

Figure 4. ChromWave recovers computationally an experimental study of an unwrapped nucleosome bound by RSC between the Gal1 and Gal10 genes

A. MNase-seq profiles and ChromWave's predictions (in Fragments Per Million, FPM) of TF and nucleosome occupancy in the intergenic region on chromosome 2 between the GAL10 and GAL1 genes. The light shaded region indicates a yeast regulatory locus GAL4 UAS. The dark shaded region indicates a 61bp region deleted in C. Inside, a partially unwrapped nucleosome bound by RSC is shown as a broad peak in the TF profile as the nucleosome is represented by shorter reads. Annotations of Gal4 and RSC motifs are indicated by stippled lines. **B.** *In silico* mutagenesis with respect to the prediction in the Gal4 UAS (shaded region). Each column of the heatmap corresponds to a position in the sequence, each row represents a mutation to the corresponding nucleotide. The quantities in the heatmap display the difference "Δ pred" in ChromWave's predictions (summed across the shaded region) after substituting the row's specified nucleotide into the sequence. Annotation of the GAL4 and RSC motifs show that point mutations in a RSC binding site (which coincides with the first RSC motif) lead to the biggest loss in prediction in the TF occupancy profile in the GAL4 UAS. **C.** ChromWave's predictions of TF and nucleosome occupancy after deleting a 61bp sequence containing the RSC binding site (indicated by the stippled lines and the dark shaded region in A) in the GAL4 UAS. TF prediction changes to a narrow peak towards the 5' end of GAL4 UAS, and a strongly positioned nucleosome is predicted in the middle of the GAL4 UAS. Previously hypersensitive sites on the borders of the shaded region are now predicted to be occluded by nucleosomes. These predictions recover experimental results in the original study on this unwrapped nucleosome (Floer et al., 2010). **D.** Original Figure 1F from (Floer et al., 2010) for comparison. RSC binding to a WT and a truncated *UASg*. Cells bearing TAP-tagged RSC and grown in raffinose. The four Gal4 sites within the *UASg* are in cyan.

Figure 5. Nucleosome-only ChromWave models to decipher determinants of *in vivo* nucleosome occupancy in the yeast genome.

A. A representative example of the predicted *in vitro* and *in vivo* nucleosome occupancy profiles from the training set (chromosome 14). Occupancy is measured and predicted as log2-fold change of number of reads at each position over genome average, in Reads Per Million (logfc-RPM). Predictions by ChromWave are shown in black for comparison with the observed MNase-seq data shown in grey and the predictions of the Hidden Markov Model (HMM) trained on the *in vitro* data (Kaplan et al., 2009) shown in blue. Stippled boxes indicate observable differences in nucleosome occupancy *in vitro* and *in vivo.* **B** Genome-wide comparison of predicted and observed nucleosome occupancy. Pearson Correlation coefficients were computed between predicted and observed nucleosome occupancy (left: *in vitro,* right: *in vivo*). The chromosomes forming the validation and test sets as indicated (all other chromosomes formed the training set). Correlations were computed in sliding windows of size 300bp and then averaged over 300bp bins for visualisation**. C.** Predictions per base-pair centered on nucleosome dyads defined by chemical cleavage (Chereji et al., 2018). Heatmaps showing +/-1kb regions around annotated nucleosome dyads. Average predictions are shown as metaprofiles on top of each heatmap. Plots from *left* to *right* display predictions of the following models: *in vitro* nucleosome ChromWave, *in vivo* nucleosome Chromwave, and nucleosome prediction and TF predictions from the joint TF-nucleosome ChromWave model. Each track was normalised by subtracting the genomic average before plotting. **D.** Motif detectors associated with *in vivo* nucleosome occupancy. Each point in the scatter plot represents a motif that is PCA embedded using maximum activations of motif detectors per sequence as input. The shape of each point indicates if the motif was annotated by TOMTOM (triangle) or de novo discovered (point). The size of each point was scaled with the mean activation of the motif detector. The colour scale corresponds to the mean Pearson correlation between activation and prediction per base-pair. Selected motifs were annotated with their associated sequence logo. **E.** *In vivo* and *in vitro* nucleosome occupancy prediction per base-pair centered on TF binding sites defined by ChIP-exo. Heatmaps showing +/-1kb regions around the ChIP-exo peaks. Each heatmap panel is based on a different TF peak set from top to bottom: Abf1, Cbf1, Rap1 and Reb1. Heatmaps from *left* to *right* display predictions of the following models: *in vitro* nucleosome ChromWave, *in vivo* nucleosome Chromwave, and nucleosome prediction and TF predictions from the joint TF-nucleosome ChromWave model. Average predictions (logfc-RPM for the nucleosome-only model, FPM for TF-nucleosome model) are shown as metaprofiles on top of each heatmap. Each track was normalised by subtracting the genomic average before plotting.

Figure 6. Predicting nucleosome occupancy in human promoters with ChromWave.

A. Comparison of MNase-seq and predicted nucleosome occupancy in +/- 1kb around 23,156 annotated human TSSs. Occupancy is measured and predicted as log2-fold change of number of reads at each position over genome average, in Reads Per Million (logfc-RPM). *Top:* Metaprofiles of average nucleosome occupancy measured by MNase-seq (left) and predicted by ChromWave (right) are centered around TSSs. Heatmaps are showing the data and predictions of each TSS in the rows. TSS are sorted by the Pearson Correlation coefficients between observed and predicted profiles from top to bottom (as indicated by the arrow) and these are shown on the left as barplot. **B.** Motif detectors associated with human promoter nucleosome occupancy. Each point in the scatter plot represents a motif that is PCA embedded using maximum activations of motif detectors per sequence as input. The shape of each point indicates if the motif was annotated by TOMTOM (triangle) or de novo discovered (point). The size of each point was scaled with the mean activation of the motif detector. The colour scale corresponds to the Pearson correlation between activation and prediction per base-pair. Selected motifs were annotated with their associated sequence logo.

Figure 7. ChromWave learns known and unknown motifs associated with nucleosome binding and DNA-methylation in human promoter regions

A. Volcano plot to identify dsQTLs that change nucleosome occupancy in promoter regions. Each point represents a dsQTL variant, some annotated with a promoter name. Along the x-axis, we plot the difference (∆ prediction (MT-WT)) between the predicted nucleosome occupancy given dsQTL variant and the wild-type input sequence. Along the y-axis, we show (1-the Pearson correlation coefficient) between the predicted nucleosome occupancy given the dsQTL variant and the wild-type input sequence. The shape of each point indicates if dsQTL falls within or outside of the previously associated DNAse-Hypersensitive Site (DHS). The colours represent the genotype of the variant as indicated. **B.** Example of maximal disruptive dsQTLs on nucleosome occupancy predictions. *Top Panel:* Nucleosome occupancy for different genotypes at the dsQTL (indicated by the dotted red line) compared with observed data. *Bottom Panel:* Line plots show the maximal increase ('gain') and maximal decrease ('loss') in the predictions at each position given all possible single base-pair changes along the sequence. **C.** Example of maximal disruptive dsQTLs on nucleosome occupancy predictions. As B. **D.** Hotspots in human promoters susceptible to changes in chromatin accessibility due to single nucleotide polymorphisms. Maximal increase ('gain') and maximal decrease ('loss') in the predictions at each position given all possible single base-pair changes along the sequence in +/- 1kb around 23,156 annotated human TSSs. *Top:* Meta-profiles of maximal increase ('gain') (*left*) and maximal decrease ('loss') (*right*) in ChromWave's predictions centered around the TSSs. *Bottom:* Heatmaps are showing the maximal increase and decrease for each TSS in the rows.

Supplementary Figure 1.

A. Genome-wide comparison ChromWave's predictions of TF and nucleosome occupancy and observed data. Shown are Pearson Correlation coefficients between predicted and observed TF (left) and nucleosome (right) occupancies (in Fragments Per Million, FPM). The chromosomes forming the validation and test sets are indicated (all other chromosomes formed the training set). Correlations were computed in sliding windows of size 300bp and averaged across 300bp bins for visualisation**. B.** A representative example imputed TF and nucleosome occupancies from the test set (chromosome 10). Shown are ChromWave's predictions of TF and nucleosome occupancy profiles (in FPM) in black in an unmapped region for comparison with the observed MNase-seq data (in FPM) shown in grey.

Supplementary Figure 2.

A. Genome-wide comparison of predicted and observed nucleosome occupancy across different datasets. Shown are Pearson's correlation coefficients of the predicted occupancy of the *in vitro* and *in vivo* nucleosome ChromWave models and the HMM from (Kaplan et al., 2009) with ATAC-Seq (Schep et al., 2015) and the *in vitro* and *in vivo* MNase-seq data both genome-wide and separately on training, test and validation sets, as indicated. **B.** Comparison of nucleosome occupancy across models and across in vivo and in vitro datasets in yeast promoter. Metaprofiles +/- 1bp around all TSS on chromosome 10 of average predicted and observed nucleosome occupancy (in logfc-RPM)(upper panel). Metaprofiles +/- 1bp around all TSS of average predicted and observed nucleosome occupancy (in logfc-RPM) *in vitro* (middle panel) and *in vivo* (lower panel). Each track was normalised by subtracting the genomic average before plotting. **C.** Model predictions per base-pair centered on nucleosome dyads called on

ATAC-seq. Occupancy is measured and predicted as log2-fold change of number of reads at each position over genome average, in Reads Per Million (logfc-RPM). Heatmaps showing +/-1kb regions around ATAC-seq nucleosome dyads for four biological replicates (A, B, C, D). Average predictions (in logfc-RPM) are shown as metaprofiles on top of each heatmap. Plots are displaying nucleosome occupancy prediction using following models: *in vitro* nucleosome ChromWave ("*in vitro")*, *in vivo* nucleosome Chromwave ("*in vivo")*, and nucleosome prediction ("NUC") and TF predictions ("TF") from the joint TF-nucleosome ChromWave model. Each track was normalised by subtracting the genomic average before plotting. **D.** MNase-seq occupancy profiles centered on nucleosome dyads defined by chemical cleavage (Brogaard et al., 2012). Plots are displaying nucleosome and TF occupancy using following data: *in vitro* nucleosome ("*in vitro")* and *in vivo* nucleosome ("*in vivo")* from (Kaplan et al., 2009) (measured as log2-fold change of number of reads at each position over genome average, in Reads Per Million (logfc-RPM)), and nucleosome ("NUC") and TF ("TF") in Fragments Per Million (FPM) from (Henikoff et al., 2011). Heatmaps showing +/-1kb regions around nucleosome dyads. Dyads not covered by the nucleosome MNase-seq data from (Kaplan et al., 2009) show as no signal in the two left heatmaps. Average occupancies are shown as metaprofiles on top of each heatmap.

Supplementary Figure 3.

A. Motif clustering on the convolutional filter from all yeast models. Briefly, pairwise motif similarities were computed between the filters from all yeast models and used as input for hierarchical clustering visualised here as heatmap. Clusters were subsequently defined by cutting the dendrogram at height h=0.9 indicated in the barplot on top of the heatmap. The clusters were manually annotated using previous TOMTOM results as shown on the right.

Supplementary Figure 4.

A. Comparison of MNase-seq and predicted nucleosome occupancy +/- 1kb around all human TSSs grouped by training, test and validation sets as indicated. Occupancy is measured and predicted as log2-fold change of number of reads at each position over genome average, in Reads Per Million (logfc-RPM). *Top:* Meta-profiles of average nucleosome occupancy measured by MNase-seq (left) and predicted by ChromWave (right) (in logfc-RPM) are centered around 23,156 annotated TSSs. *Bottom:* Heatmaps are showing the data and predictions of each TSSs in the rows. TSS are sorted by the Pearson Correlation between observed and predicted profiles from top to bottom as indicated by the arrow. Individual Pearson correlation coefficients are shown on the left of each heatmap as barplot. **B.** MNase-seq and predicted nucleosome occupancy around ChIP-seq CTCF binding sites. Metaprofiles are centered on CTCF peak summits, and these were split into quintiles according to the CTCF peak scores. *Top*: Metaprofile of the average nucleosome occupancy, as measured by MNase-seq (in logfc-RPM), for +/- 1kb around CTCF binding sites. *Bottom:* Metaprofile of ChromWave's predicted average nucleosome occupancy (in logfc-RPM). Colors indicating CTCF quintiles, where quintile "five" corresponds to highest, and quintile "one" to the lowest ChIP-seq scores.

Star Methods

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Dr Sera Aylin Cakiroglu (aylin.cakiroglu@crick.ac.uk).

Materials Availability
This study did not generate new unique reagents.

Data and Code Availability
The code, trained models, preprocessed input data and model predictions as BigWig files are publicly available on github (https://github.com/luslab/ChromWave). R-scripts implementing several visualisation methods can also be found on github (https://github.com/luslab/chromWaveR). The hyperparameter search space as well as the optimal parameters for all models can be found in Supplementary Table1. Motif clusters of the yeast models can be found in Supplementary Table 2. Human promoters with variable nucleosome occupancy at different base-changes in dsQTLs are listed in Supplementary Table 3.

METHOD DETAILS

Data Preparation

Yeast data sets

We obtained the reference genome sequences of sacCer1 and sacCer3 (Engel et al., 2014). We used the *biomaRt* package (Durinck et al., 2005, 2009) to retrieve gene annotation information from Ensembl release 91 (Zerbino et al., 2018) using BioMart web services (Kasprzyk et al., 2004; Smedley et al., 2015). We downloaded the nucleosome occupancy maps from (Kaplan et al., 2009) available as log-ratios of paired-end *in vivo* and *in vitro* MNase data aligned to sacCer1. Missing values were replaced by the chromosome mean and values exceeding 3 times the genome-wide median were clipped to this maximal value. We smoothed the profiles using a 1D Gaussian filter with standard deviation (std) σ = 5 truncated at 3 stds. The genome-wide minimum was added to all values before scaling and discretizing the continuous values y as $x = floor(2.5 * y)$. This resulted in 22 discrete classes for both the *in vitro* and *in vivo* data. We obtained the Hidden-Markov model (HMM) from the same publication

(Kaplan et al., 2009) and generated its predictions for both sacCer1 and sacCer3. We annotated the unique nucleosome map in *S.cerevisiae* from chemical cleavage sequencing of (Brogaard et al., 2012) to the yeast genome sacCer3 and obtained newer chemical cleavage nucleosome occupancy data from the Henikoff lab (Chereji et al., 2018). Nucleosome dyad positions derived from ATAC-seq samples were downloaded from (Schep et al., 2015). We obtained raw MNase-seq data from (Henikoff et al., 2011) and mapped these to the sacCer3 genome with Bowtie2 (parameters: `-X 2000`), version 2.3.4.3
(Langmead and Salzberg, 2012). Resulting alignments were filtered to remove unmapped, multi-mapping and unpaired reads and duplicated fragments were removed. Next, we computed separate signal tracks for TF fragments (max length <= 80) and mono nucleosome fragments (140 <= length <= 200). Both signaling tracks were scaled to fragments per million (FPM) where the scaling factor was computed as 10^6 divided by the number of fragments of the TF and mononuc fragments separately. Finally, signal track computation was performed using deepTools2's "bamCoverage" function (parameters: `-bs 1 --extendReads  --center`) (Ramírez et al., 2016). Given the significant difference in sequencing depth of the two replicates (~191 million reads vs ~31 million) we only used the deeper sequenced first replicate for all subsequent analysis. Instead of normalising the read coverage by taking the log ratio between coverage and genomic mean, we are modelling here directly the number of reads to preserve the sparsity of the TF-fragment signal. Missing values were replaced by the chromosome mean, read counts exceeding 3 times the genome-wide median were clipped to this maximal value and normalised by subtracting the genome-wide mean after which negative values were set to zero. We then smoothed the two profiles separately using a 1D Gaussian filter with $\sigma = 5$ truncated at 3 stds. We discretized these continuous values y as $x = floor(Z * y)$ where Z=10 and Z=5 for the transcription factor and nucleosome profiles, respectively. This resulted in 5 and 9 discrete classes for the transcription factor and nucleosome profiles respectively.
Publicly available preprocessed high-resolution binding sites for Abf1, Cbf1, Rap1 and Reb1 as measured by ChIP-exo were downloaded from (Rossi et al., 2018).

Human data sets

We obtained gene and TSS annotations as well as the reference sequences in hg38 from UCSC via the R Package TxDb.Hsapiens.UCSC.hg38.knownGene (Team BC, Maintainer BP, 2019). We downloaded nucleosome occupancy data from (Gaffney et al., 2012) preprocessed and aligned to the hg38 reference genome as bigWig file (in Reads Per Million, RPM) from (Zhao et al., 2018)(sample *hsNuc0340101*). The read coverage per base-pair around human TSSs was computed with the function "computeMatrix" in the deepTools2 package (Ramírez et al., 2016) (parameters: `reference-point --referencePoint TSS -a 1000 -b 1000 --binSize 1 --sortRegions keep --nanAfterEnd  --missingDataAsZero`). Read counts exceeding 3 times the median of the whole dataset were clipped to this maximal value and subsequently normalised by taking the log2 ratio between read counts and genome-wide mean. We then smoothed the profiles with a 1D Gaussian filter std $\sigma = 5$ truncated at 3 stds and discretized the continuous signal y as $x = floor(2.0 * y)$. This resulted in 17 discrete classes.

Publicly available CTCF binding sites for the cell line GM12878 were downloaded from the ENCODE portal (Davis et al., 2018; Zhang et al., 2020).
Previously determined DNAse hypersensitivity quantitative trait loci (dhsQTL) were downloaded from the supplementary information of (Degner et al., 2012). Downloaded dhsQTLs were overlapped with promoter regions and we only retained those dsQTLs that were overlapping promoter regions (+/-1kb around TSSs).

Model architecture, implementation and training

Binary classification CNN to model nucleosome specificity

Nucleosomal DNA-sequences of 201bp were derived from the set of unique nucleosome dyads as measured by chemical cleavage (Brogaard et al., 2012) by extending 100bp either side from the dyad position. Dyads that were closer than 100bp to chromosome ends were excluded. This set of nucleosome sequences was then divided into training (50%), validation (30%) and test (20%) set. Reverse complements of the sequences were added to each of these datasets and then one-hot encoded. One-hot encoded means in this context that for example adenine is encoded as (1, 0, 0, 0), cytosine as (0, 1, 0, 0) and so forth. We generated a set of ‘negative’ examples of the same size by randomly shuffling each of the input sequences preserving the distribution of dinucleotides. One hot encoded nucleosomal and background sequences were used as input to train convolutional neural networks (CNN) implemented to predict labels y=1 (nucleosomal sequence) and y=0 (shuffled sequence) respectively.

CNNs were implemented with the python package Keras2.0 (Chollet and Others, 2015) with tensorflow backend. We used dropout and early stopping (patience = 20). We trained each CNN for 100 epochs with the Adam optimizer and reduced the learning rate by a factor of 0.1 if the loss had not improved for 10 epochs. In addition to the loss function, we monitored the accuracy as area under the receiver-operator curve (auROC) and Matthew’s correlation coefficient (MCC). To determine the optimal architecture of a CNN capturing sequence features of nucleosome dyad preferences *in vivo,* we searched the hyperparameter search space in Supplementary Table 1 to determine the optimal set of hyperparameters using the hyperparameter optimization algorithm “Tree of Parzen Estimators” (TPE) (Bergstra et al., 2011) as implemented in hyperopt package (Bergstra et al., 2013). We performed 51 independent hyperparameter searches. In each of the hyperparameter searches, we performed 30 trials (e.g. iterations) and returned the model maximising the MCC on the validation dataset across those trials. As the final model, we chose the model which maximised this score across all 51 searches.

ChromWave models

We divided the yeast genome into windows of fixed length which formed as one-hot encoded DNA sequences the input to the ChromWave models. We split the chromosomes randomly into training, validation and test sets (excluding the mitochondrial chromosome): chromosomes 9 and 16 of sacCer1 (chromosomes 2 and 10 of sacCer3) were used for test, chromosomes 4, 5 and 7 of sacCer1 (chromosomes 3, 5, and 11 of sacCer3) for validation and all other chromosomes of sacCer1 (of sacCer3) used for training for nucleosome (TF-nucleosome) models. As prediction target for each sequence, the respective discretised *in vitro*, *in vivo* nucleosome occupancy profile or the TF-nucleosome profiles were used. For each sequence, the reverse complement together with the reversed occupancy profile(s) were added to the respective data set. Sequence and target pairs thus allocated to training, validation and test set were randomly shuffled within each set.

At each base-pair position the model is asked to perform a classification task to predict the class of the discretised nucleosome occupancy at that position. To combat the bias towards the overrepresented genomic mean, we use a weighted multi-class cross-entropy as loss function where the class weights were computed as follows: the class weight of class *i* was computed as the median of all class frequencies divided by the frequency of class *i.* Here, the class frequency was computed as the total number of class occurrences divided by the fragment length. Finally, to avoid very large weights for rarely occurring classes the minimum between the class weight and 100 was chosen. In the case of the TF-nucleosome model, the class weights were independently computed for each of the two profiles. Code implementing the models and hyperparameter search can be found in the ChromWave repository on github (https://github.com/luslab/ChromWave/). Models were implemented using the python package Keras 2.3.1 (Chollet and Others, 2015) with tensorflow backend.

In addition to the loss function, we monitored the class accuracy and Pearson’s correlation coefficient between observed and predicted profiles on the validation set. To determine the optimal architecture of a ChromWave model, we searched the hyperparameter search spaces in Supplementary Table 1 to determine the optimal sets of hyperparameters of the model maximising the sum of class accuracy and Pearson’s correlation coefficient between observed and predicted profiles in the validation set. The hyperparameter optimisation algorithm TPE (Bergstra et al., 2011) was used as implemented in the hyperopt package (Bergstra et al., 2013). In each of the hyperparameter searches, we performed 100 trials where we trained each model for 50 epochs and then trained 6 additional neural networks on the full training set using this set of hyperparameters for 100 epochs.

For the human promoter model, we used one-hot encoded DNA-sequences of the human genome (*hg38*) corresponding to +/-1000bp around annotated TSSs as input to the model. We split the chromosomes randomly into training, test and validation data (excluding the sex and mitochondrial chromosomes). As prediction target for each sequence, the respective discretised nucleosome occupancy profile was used. For each sequence, the reverse complement together with the reversed occupancy profile was added to the respective data set. Sequence and target pairs allocated to training, validation and test set were randomly shuffled within each set. The class frequency was computed as the total number of class occurrences divided by the fragment length. Finally, to avoid very large weights for rarely occurring classes the minimum between the class weight and 40 was chosen. We employed the same training strategy as for the yeast ChromWave models. The search spaces for the hyperparameters and optimal values for each of the models are shown in Supplementary Table 1.

Transfer learning

In addition to training the weights of the neural networks from a random initialisation, we also compared different transfer learning approaches.

For the *in vivo* nucleosome occupancy, we trained (i) a network was initialised with the weights of the *in vitro* nucleosome Chromwave (keeping the best architecture constant) and was trained for a further 100 epochs, (ii) networks during a hyperparameter search with (a) a set of convolutional layers are initialized with the weights of the filters of the *in vitro* nucleosome ChromWave model and subsequently frozen (ie. not further trained), or (b) trained further, and (c) a set of convolutional layers are initialized with the weights of the filters of CNNs trained on protein-binding-microarray data of 89 yeast transcription factors (Zhu et al., 2009) in addition to the nucleosome *in vitro* nucleosome ChromWave model filters, as described in (Alipanahi et al., 2015) - here all filters were further trained.

For the TF-nucleosome competition model, we compared (i) initialising all convolutional layers at random, (ii) adding the pre-trained convolutional layers of the nucleosome specificity model trained on the chemical cleavage data (but not allowing these to be trained further), and (iii) initialising a set of trainable convolutional layers with position weight matrices (and their reverse complement) of a set of yeast transcription factors.

The final models were trained using the following transfer learning approaches. For the *in vitro* nucleosome occupancy we initialise the first set of non-trainable convolutional layers of the ChromWave model with the weights from the convolutional layers of the CNN trained on the chemical cleavage data. For the *in vivo* nucleosome occupancy, we initialised a set of convolutional layers with the weights of the filters of the *in vitro* nucleosome ChromWave model which were not trained further. For the TF-nucleosome competition model, we added the convolutional layers of the CNN trained on the chemical cleavage data (but not allowing these to be trained further), and initialised a set of trainable convolutional layers with position weight matrices (and their reverse complement) of a set of yeast transcription factors. As set of transcription factors, we chose the following previously described nucleosome-deplacing factors: Abf1, Cbf1, McM1, Rap1, Reb1, Orc1, Asg1, Azf1, Bas1, Ecm22, Ino4, Leu3, Rfx1, Rgm1, Rgt1, Rsc3, Sfp1, Stb4, Stb5, Stp1, Sum1, Sut1, Tbf1, Tbs1, Tea1, Uga3, Ume6, Urc2 (Groups 1&2 in (Yan et al., 2018)). PFMs and convolutional layers of the pretrained CNNs are also available on github (https://github.com/luslab/ChromWave/).

For the human promoter nucleosome model, we trained networks during a hyperparameter search from a random initialisation as well as with a set of convolutional layers initialized with the weights of the filters of an intermediate *in vivo* nucleosome occupancy trained on the yeast data (which was not used as the final yeast model). The final human promoter model was derived employing the described transfer learning approach.

QUANTIFICATION AND STATISTICAL ANALYSIS

Correlations between observed and predicted signals genome-wide

We computed the Pearson's correlation coefficients in sliding windows of 300bp (with step size 1) between observed and predicted profiles. To evaluate the performance of the TF-Nucleosome competition model we computed the geometric means between the TF and nucleosomes profiles as previously published in (Zhong et al., 2014): if r_{TF} and $r_{NUC}$are the Pearson correlation coefficients of predicted and observed TF and nucleosome signals in a 300bp window, we assign the score $1/2\sqrt{(1+r_{TF})\times(1+r_{NUC})}$ to the first base in that window. Note that this score ranges from 0 to 1. In case of the nucleosome-only data, we report genome-wide measures (e.g Pearson's correlation coefficient) using all regions omitting bases with missing values in the MNase-seq data. This approach is not used for the TF-nucleosome model where the TF signal is very sparse and excluding regions with zero reads would highly influence this correlation and at the same time not penalise for predicted signal in unmapped regions.

Motif filter visualisations and annotations

The first convolutional layer of a CNN can be summarised as a conventional position frequency matrix (PFM) and visualised with a seqlogo plot as previously described (Alipanahi et al., 2015; Angermueller et al., 2017). We matched these PFMs against available yeast motif databases (Fornes et al., 2020; Hume et al., 2015; MacIsaac et al., 2006; Pachkov et al., 2013; Zhu and Zhang, 1999) with TOMTOM (Gupta et al., 2007). Motif filters were defined as 'known' if they matched against a known motif with q-value < 0.05.

PCA embeddings of learned patterns/motifs

As a measure of importance of the learned motifs, we computed the maximum output value (maximal activation) of the respective ChromWave filter for each DNA-sequence window. To understand the impact of these motifs on ChromWave's TF predictions better, we computed the correlation between the vector of filter activations along the sequence and the predicted TF occupancy profile. A positive correlation implies the existence of the motif in a sequence is increasing TF binding, while a negative correlation implies the motif being unfavourable for TF binding. We then applied PCA to the maximal activations of all filters across all DNA-sequence windows.

Clustering of learned patterns/motifs across models

Motif clustering was performed as previously described in (Vierstra et al., 2020) with code obtained from https://github.com/jvierstra/motif-clustering. Briefly, motif similarities were computed between filters from all yeast models (*invitro, invivo, TF-NUC,* excluding the pre-trained filters of each model) using TOMTOM (Gupta et al., 2007). These pairwise similarities were used as input for hierarchical clustering followed by cluster definition by cutting the dendrogram on a given height (h=0.9). These clusters were manually annotated using previous TOMTOM results and visualized. Results can be found in Supplementary Table 2.

In silico mutagenesis scores and visualisation

To visualise the changes in prediction given changes in input, we computed the prediction profile for each of the possible base changes at each position along the input sequence as previously described (Alipanahi et al., 2015). To visualise the effect of each base change in an area of interest, we summed over all changes within that area of interest for each base. We visualize the result as a heatmap where each column is a position in the sequence, each row a substituted base pair and the colour reflecting this summed change across the sequence if that base pair is substituted in the sequence. To visualise global changes as line plots, we computed the maximum (‘gain’) and minimum (‘loss’) difference of the prediction given of all possible changes at each position.

DHS accessibility under *in silico* mutagenesis

We used the human-promoter trained ChromWave model to elucidate the impact of different SNPs in previously annotated *DNase I sensitivity quantitative trait loci* (dsQTLs) (Degner et al., 2012). Here we substituted each possible base-change in the position of the dsQTL and computed the difference in prediction on the dsQTL of the model. To account for overall-changes to the binding profile we also computed the Pearson’s correlation coefficient between the predicted WT profiles and the prediction given the perturbed input sequence. dsQTLs were differentiated into “within DHS” and “outside DHS” depending on their position relative to the DHS which is influenced by the respective SNP (defined in the same publication (Degner et al., 2012)). Results of this analysis can be found in Supplementary Table 3.

Other visualisations

Heatmaps and average meta-profiles were produced using the “computeMatrix” and subsequent “plotHeatmap” or “plotProfile” functions in the in the deepTools2 package (Ramírez et al., 2016).

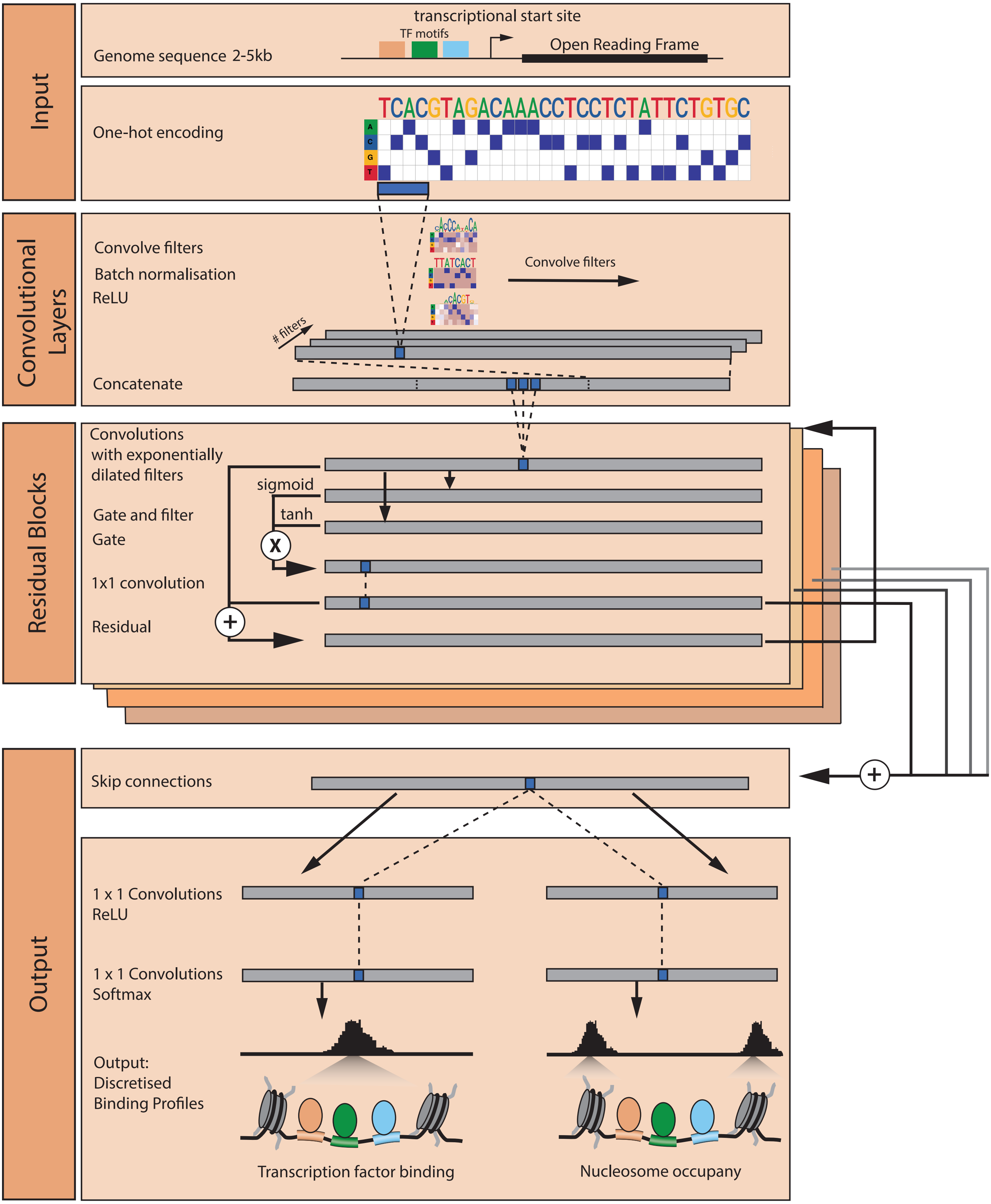

Input
transcriptional start site
TF motifs
Open Reading Frame
Genome sequence 2-5kb
One-hot encoding
TCACGTAGACAAACCTCCTCTATTCTGTGC
A
C
G
T
Convolutional Layers
Convolve filters
Batch normalisation
ReLU
TTATCACT
Convolve filters
# filters
Concatenate
Residual Blocks
Convolutions with exponentially dilated filters
sigmoid
tanh
Gate and filter
Gate
X
1x1 convolution
Residual
+
Output
Skip connections
+
1 x 1 Convolutions
ReLU
1 x 1 Convolutions
Softmax
Output:
Discretised
Binding Profiles
Transcription factor binding
Nucleosome occupany

Figure 1

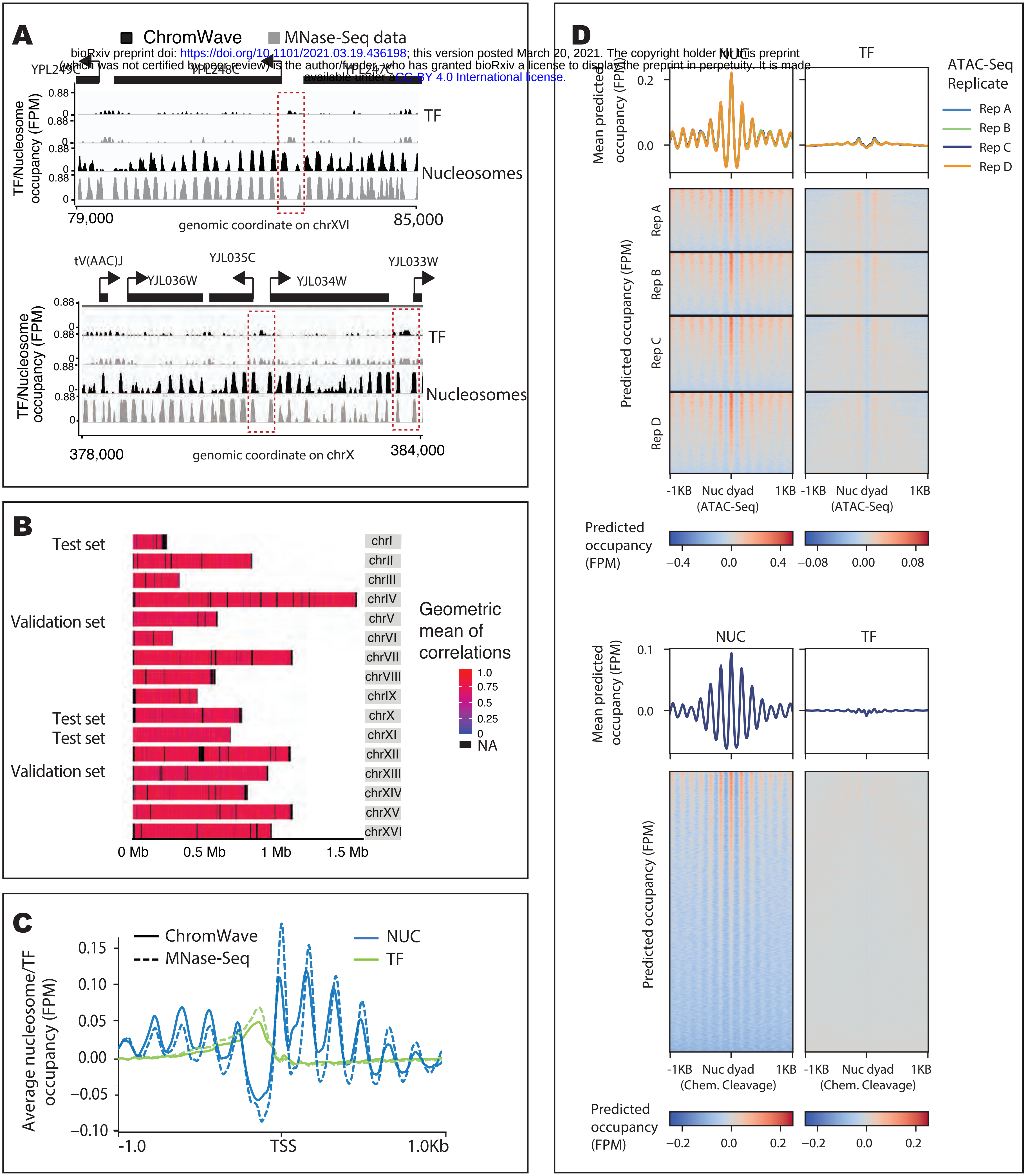

A
ChromWave
MNase-Seq data
YPL249C
YPL248C
YPL247C
TF/Nucleosome occupancy (FPM)
0.88
0
TF
Nucleosomes
79,000
genomic coordinate on chrXVI
85,000
tV(AAC)J
YJL036W
YJL035C
YJL034W
YJL033W
378,000
genomic coordinate on chrX
384,000
B
Test set
Validation set
Test set
Test set
Validation set
chrI
chrII
chrIII
chrIV
chrV
chrVI
chrVII
chrVIII
chrIX
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXV
chrXVI
0 Mb
0.5 Mb
1 Mb
1.5 Mb
Geometric mean of correlations
1.0
0.75
0.5
0.25
0
NA
C
ChromWave
MNase-Seq
NUC
TF
Average nucleosome/TF occupancy (FPM)
0.15
0.10
0.05
0.00
−0.05
−0.10
-1.0
TSS
1.0Kb
D
NUC
TF
ATAC-Seq Replicate
Rep A
Rep B
Rep C
Rep D
Mean predicted occupancy (FPM)
0.2
0.0
Predicted occupancy (FPM)
-1KB
Nuc dyad (ATAC-Seq)
1KB
Predicted occupancy (FPM)
−0.4
0.0
0.4
−0.08
0.00
0.08
NUC
TF
0.1
0.0
-1KB
Nuc dyad (Chem. Cleavage)
1KB
−0.2
0.0
0.2

Figure 2

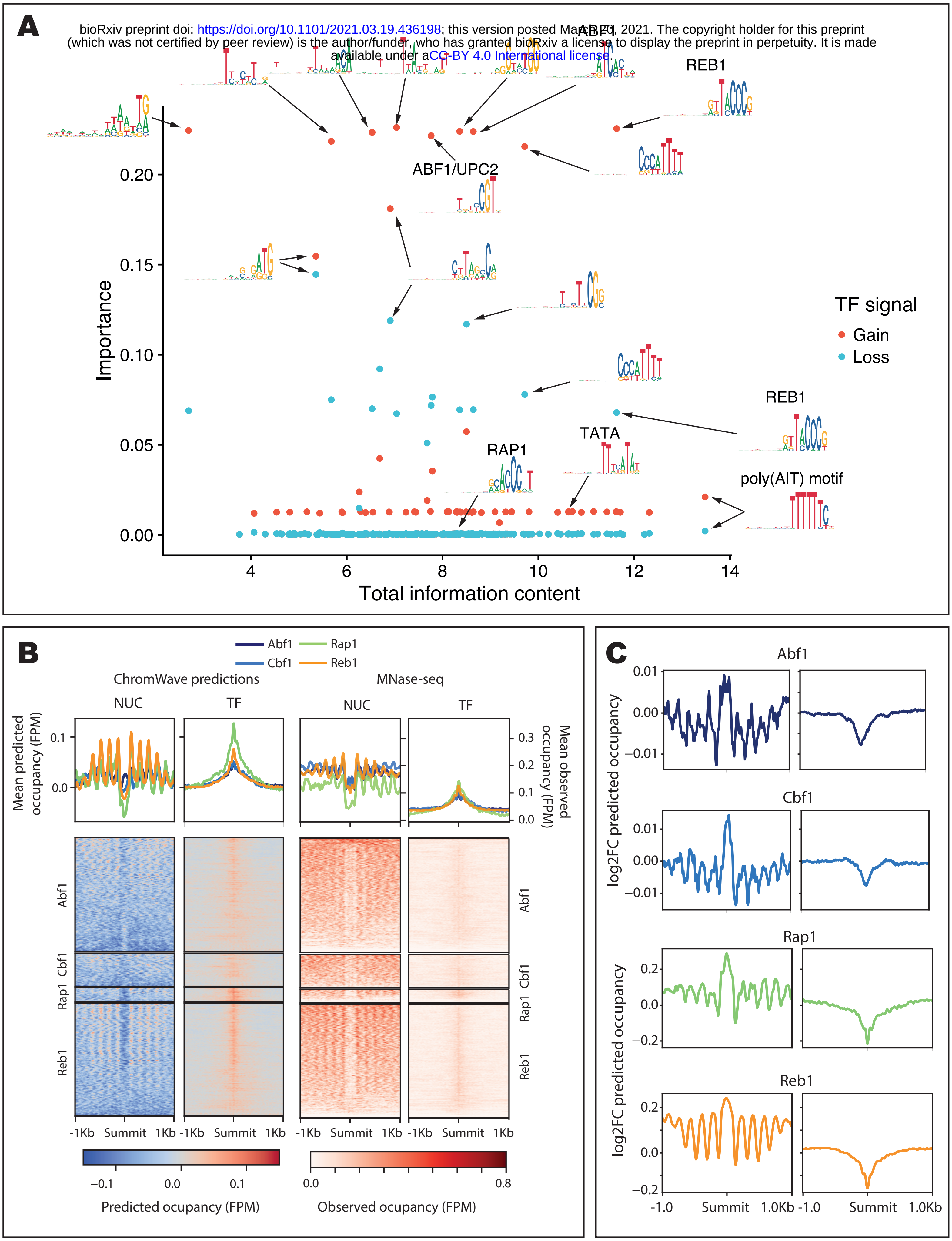

A
ABF1
REB1
ABF1/UPC2
Importance
0.20
0.15
0.10
0.05
0.00
TF signal
Gain
Loss
REB1
TATA
RAP1
poly(A|T) motif
4
6
8
10
12
14
Total information content
B
Abf1
Rap1
Cbf1
Reb1
ChromWave predictions
MNase-seq
NUC
TF
NUC
TF
Mean predicted occupancy (FPM)
0.1
0.0
Mean observed occupancy (FPM)
0.3
0.2
0.1
0.0
Abf1
Cbf1
Rap1
Reb1
-1Kb
Summit
1Kb
-0.1
0.0
0.1
Predicted ocupancy (FPM)
0.0
0.8
Observed ocupancy (FPM)
C
Abf1
Cbf1
Rap1
Reb1
log2FC predicted occupancy
0.01
0.00
-0.01
0.2
0.0
-0.2
-1.0
Summit
1.0Kb

Figure 3

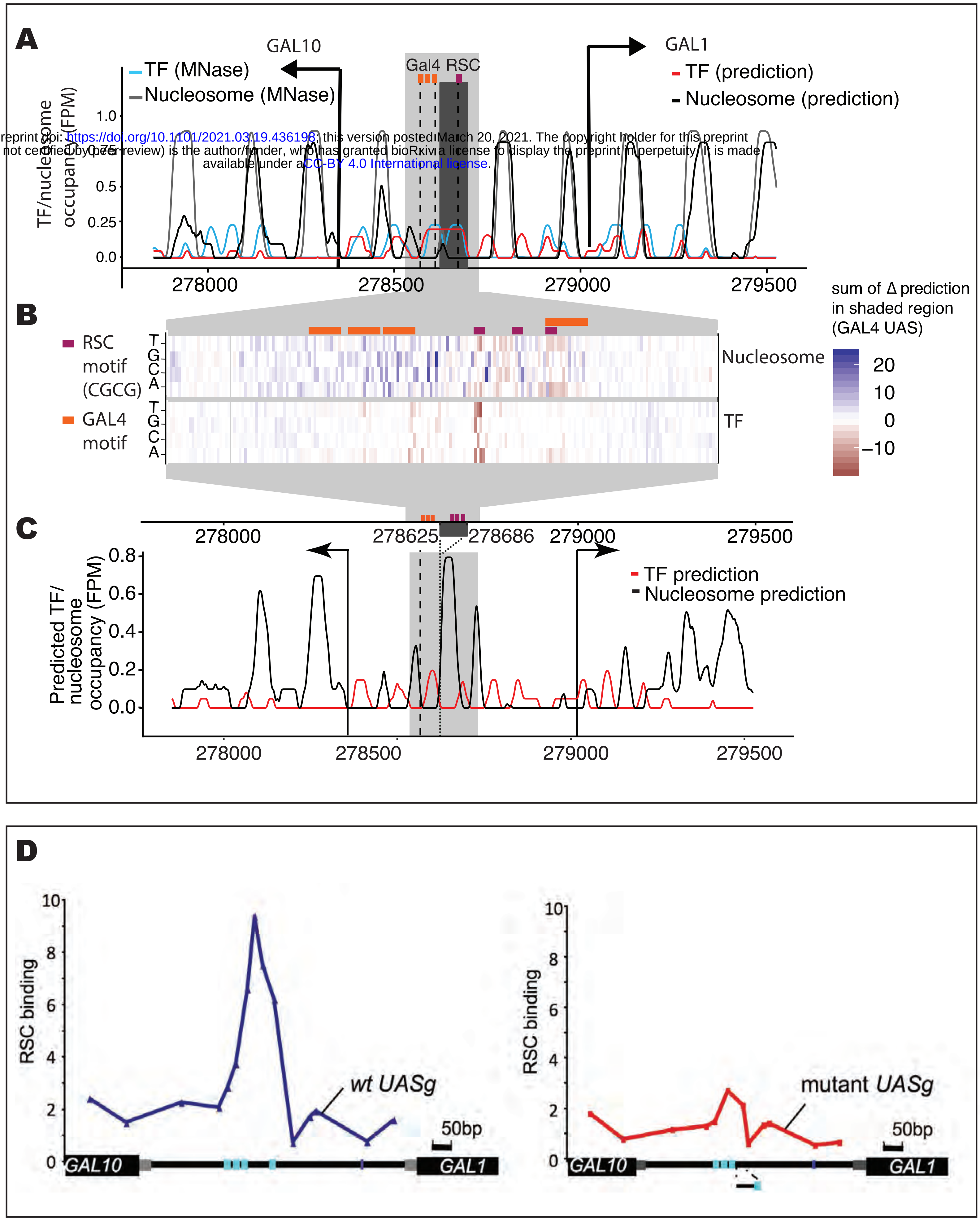

A
GAL10
Gal4 RSC
GAL1
TF (MNase)
Nucleosome (MNase)
TF (prediction)
Nucleosome (prediction)
TF/nucleosome occupancy (FPM)
1.0
0.75
0.5
0.25
0.0
278000
278500
279000
279500
B
RSC motif (CGCG)
GAL4 motif
T
G
C
A
Nucleosome
TF
sum of Δ prediction in shaded region (GAL4 UAS)
20
10
0
−10
C
278000
278625
278686
279000
279500
Predicted TF/ nucleosome occupancy (FPM)
0.8
0.6
0.4
0.2
0.0
TF prediction
Nucleosome prediction
278000
278500
279000
279500
D
RSC binding
10
8
6
4
2
0
wt UASg
50bp
GAL10
GAL1
RSC binding
10
8
6
4
2
0
mutant UASg
50bp
GAL10
GAL1

Figure 4

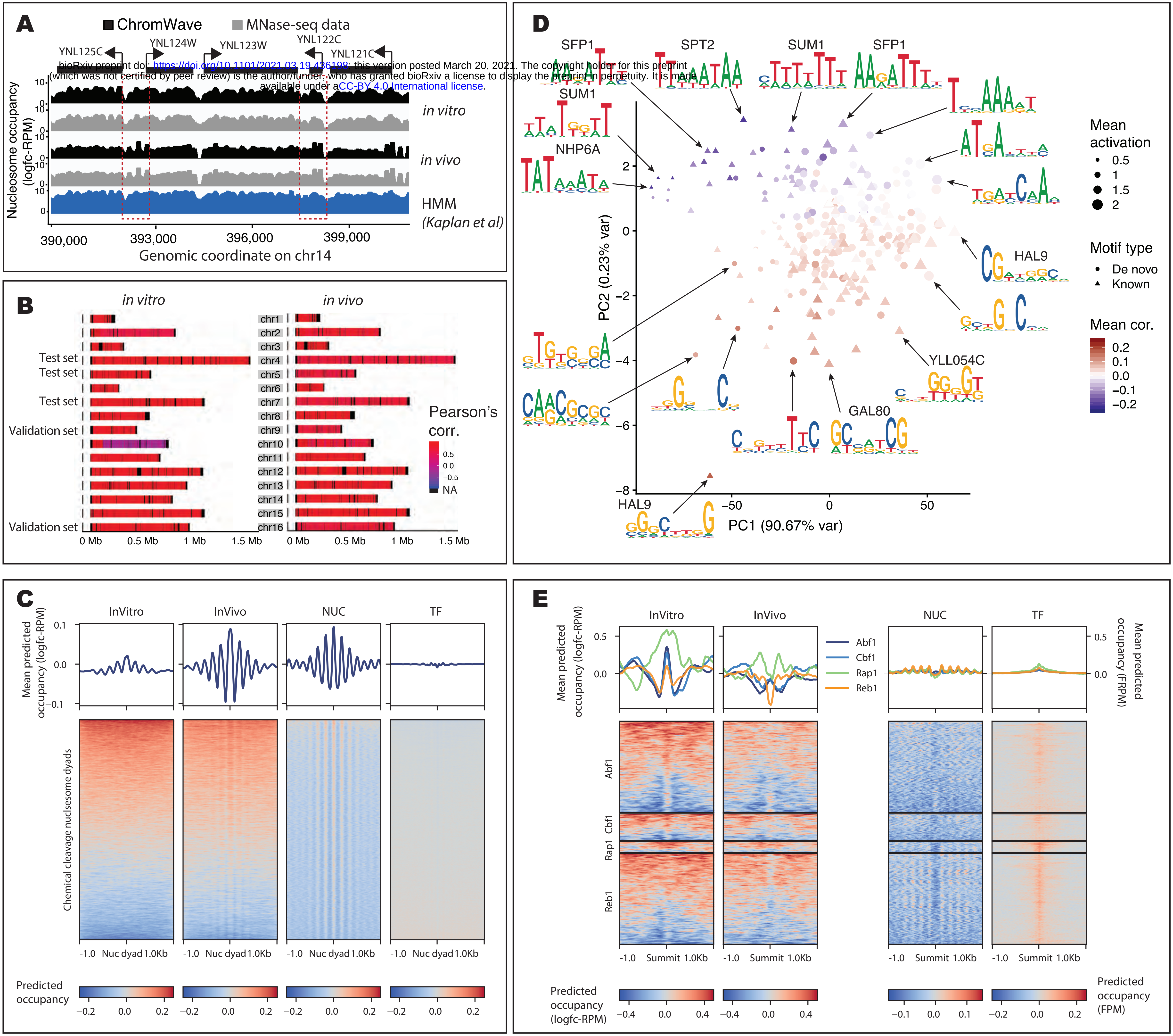

A
ChromWave
MNase-seq data
YNL125C
YNL124W
YNL123W
YNL122C
YNL121C
Nucleosome occupancy (logfc-RPM)
in vitro
in vivo
HMM (Kaplan et al)
390,000
393,000
396,000
399,000
Genomic coordinate on chr14
B
in vitro
in vivo
Test set
Test set
Test set
Validation set
Validation set
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
0 Mb
0.5 Mb
1 Mb
1.5 Mb
Pearson's corr.
0.5
0.0
−0.5
NA
C
InVitro
InVivo
NUC
TF
Mean predicted occupancy (logfc-RPM)
0.1
0.0
−0.1
Chemical cleavage nuclsesome dyads
-1.0 Nuc dyad 1.0Kb
Predicted occupancy
−0.2
0.0
0.2
D
SFP1
SPT2
SUM1
SFP1
SUM1
NHP6A
HAL9
YLL054C
GAL80
HAL9
PC2 (0.23% var)
PC1 (90.67% var)
−50
0
50
2
0
−2
−4
−6
−8
Mean activation
0.5
1
1.5
2
Motif type
De novo
Known
Mean cor.
0.2
0.1
0.0
−0.1
−0.2
E
InVitro
InVivo
NUC
TF
Abf1
Cbf1
Rap1
Reb1
Mean predicted occupancy (logfc-RPM)
Mean predicted occupancy (FRPM)
0.5
0.0
-1.0 Summit 1.0Kb
Predicted occupancy (logfc-RPM)
−0.4
0.0
0.4
−0.1
0.0
0.1
−0.2
0.0
0.2
Predicted occupancy (FPM)

Figure 5

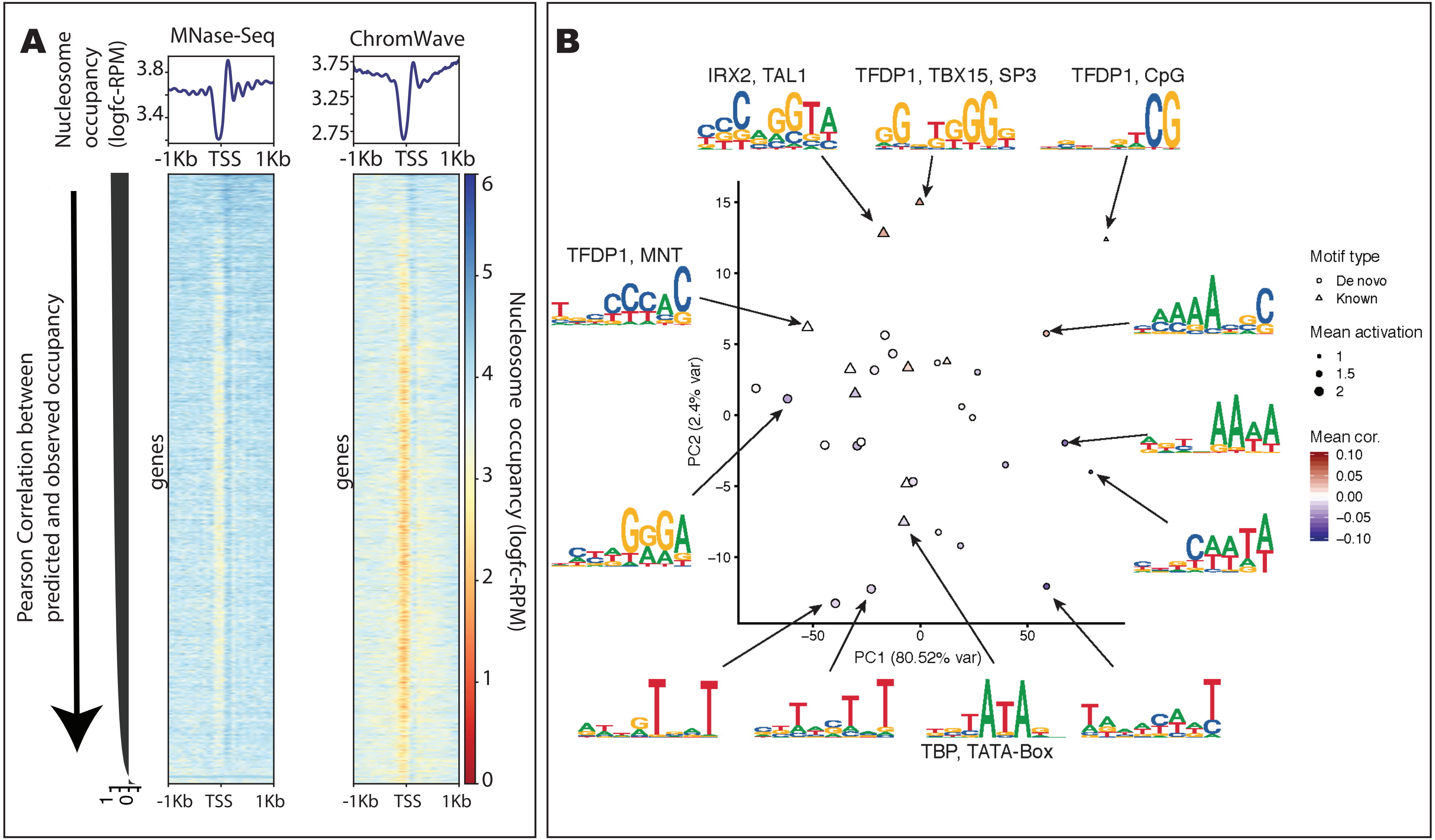
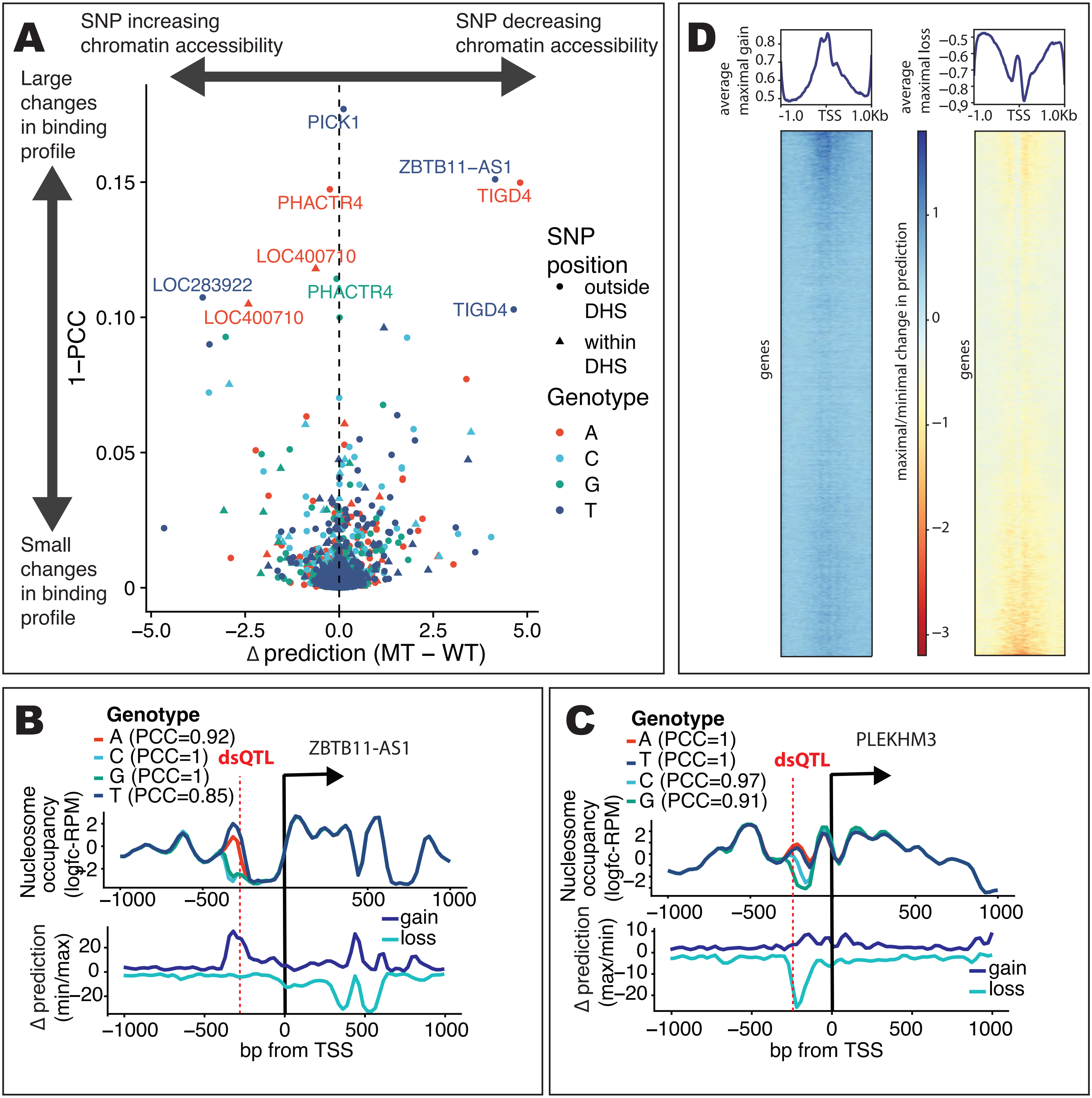

A
Nucleosome occupancy (logfc-RPM)
MNase-Seq
3.8
3.4
-1Kb
TSS
1Kb
ChromWave
3.75
3.25
2.75
-1Kb
TSS
1Kb
Pearson Correlation between predicted and observed occupancy
genes
genes
1
0
-1Kb
TSS
1Kb
-1Kb
TSS
1Kb
6
5
4
3
2
1
0
Nucleosome occupancy (logfc-RPM)
B
IRX2, TAL1
TFDP1, TBX15, SP3
TFDP1, CpG
TFDP1, MNT
TBP, TATA-Box
PC2 (2.4% var)
PC1 (80.52% var)
15
10
5
0
−5
−10
−50
0
50
Motif type
De novo
Known
Mean activation
1
1.5
2
Mean cor.
0.10
0.05
0.00
−0.05
−0.10

Figure 6



A
SNP increasing chromatin accessibility
SNP decreasing chromatin accessibility
Large changes in binding profile
Small changes in binding profile
1−PCC
0
0.05
0.10
0.15
PICK1
ZBTB11−AS1
PHACTR4
TIGD4
LOC400710
LOC283922
PHACTR4
TIGD4
LOC400710
−5.0
−2.5
0.0
2.5
5.0
Δ prediction (MT − WT)
SNP position
● outside DHS
▲ within DHS
Genotype
● A
● C
● G
● T
B
Genotype
A (PCC=0.92)
C (PCC=1)
G (PCC=1)
T (PCC=0.85)
dsQTL
ZBTB11-AS1
Nucleosome occupancy (logfc-RPM)
2
0
−2
−1000
−500
500
1000
gain
loss
Δ prediction (min/max)
20
0
−20
−1000
−500
0
500
1000
bp from TSS
C
Genotype
A (PCC=1)
T (PCC=1)
C (PCC=0.97)
G (PCC=0.91)
dsQTL
PLEKHM3
Nucleosome occupancy (logfc-RPM)
2
0
−2
−1000
−500
500
1000
Δ prediction (max/min)
10
0
−10
−20
gain
loss
−1000
−500
0
500
1000
bp from TSS
D
average maximal gain
0.8
0.7
0.6
0.5
−1.0
TSS
1.0Kb
genes
average maximal loss
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
TSS
1.0Kb
maximal/minimal change in prediction
1
0
−1
−2
−3
genes

Figure 7

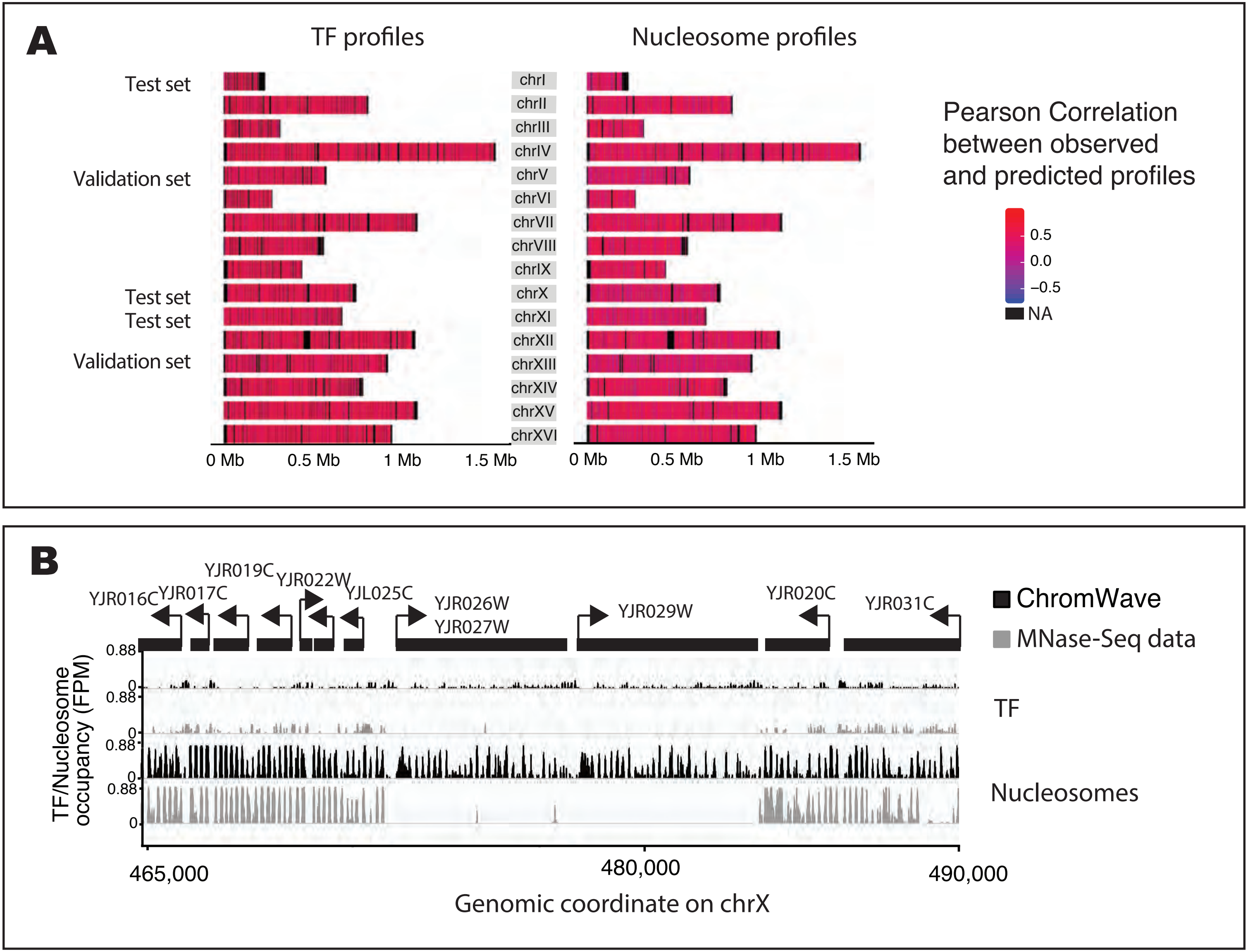

A
TF profiles
Nucleosome profiles
Test set
Validation set
Test set
Test set
Validation set
chrI
chrII
chrIII
chrIV
chrV
chrVI
chrVII
chrVIII
chrIX
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXV
chrXVI
0 Mb
0.5 Mb
1 Mb
1.5 Mb
0 Mb
0.5 Mb
1 Mb
1.5 Mb
Pearson Correlation between observed and predicted profiles
0.5
0.0
−0.5
NA
B
YJR016C
YJR017C
YJR019C
YJR022W
YJL025C
YJR026W
YJR027W
YJR029W
YJR020C
YJR031C
ChromWave
MNase-Seq data
TF
Nucleosomes
TF/Nucleosome occupancy (FPM)
0.88
0
0.88
0
0.88
0
0.88
0
465,000
480,000
490,000
Genomic coordinate on chrX

Figure S1

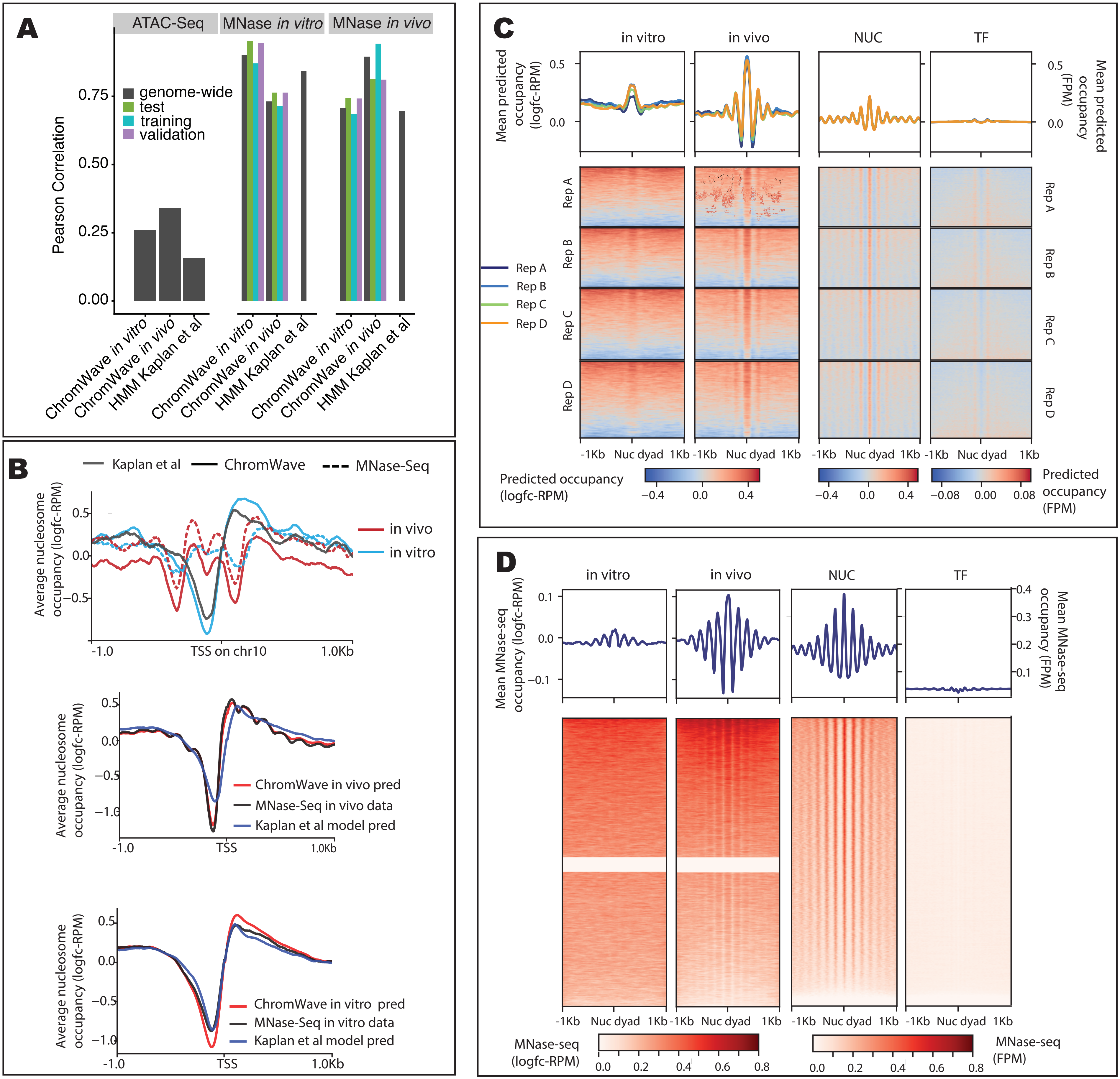

A
ATAC-Seq
MNase *in vitro*
MNase *in vivo*
Pearson Correlation
0.00
0.25
0.50
0.75
genome-wide
test
training
validation
ChromWave in vitro
ChromWave in vivo
HMM Kaplan et al
B
Kaplan et al
ChromWave
MNase-Seq
Average nucleosome occupancy (logfc-RPM)
in vivo
in vitro
0.5
0.0
−0.5
-1.0
TSS on chr10
1.0Kb
ChromWave in vivo pred
MNase-Seq in vivo data
Kaplan et al model pred
−1.0
TSS
ChromWave in vitro pred
MNase-Seq in vitro data
C
in vitro
in vivo
NUC
TF
Mean predicted occupancy (logfc-RPM)
Mean predicted occupancy (FPM)
Rep A
Rep B
Rep C
Rep D
-1Kb
Nuc dyad
1Kb
Predicted occupancy (logfc-RPM)
−0.4
0.0
0.4
−0.08
0.00
0.08
Predicted occupancy (FPM)
D
Mean MNase-seq occupancy (logfc-RPM)
Mean MNase-seq occupancy (FPM)
0.1
0.0
−0.1
0.4
0.3
0.2
MNase-seq (logfc-RPM)
MNase-seq (FPM)
0.6
0.8

Figure S2

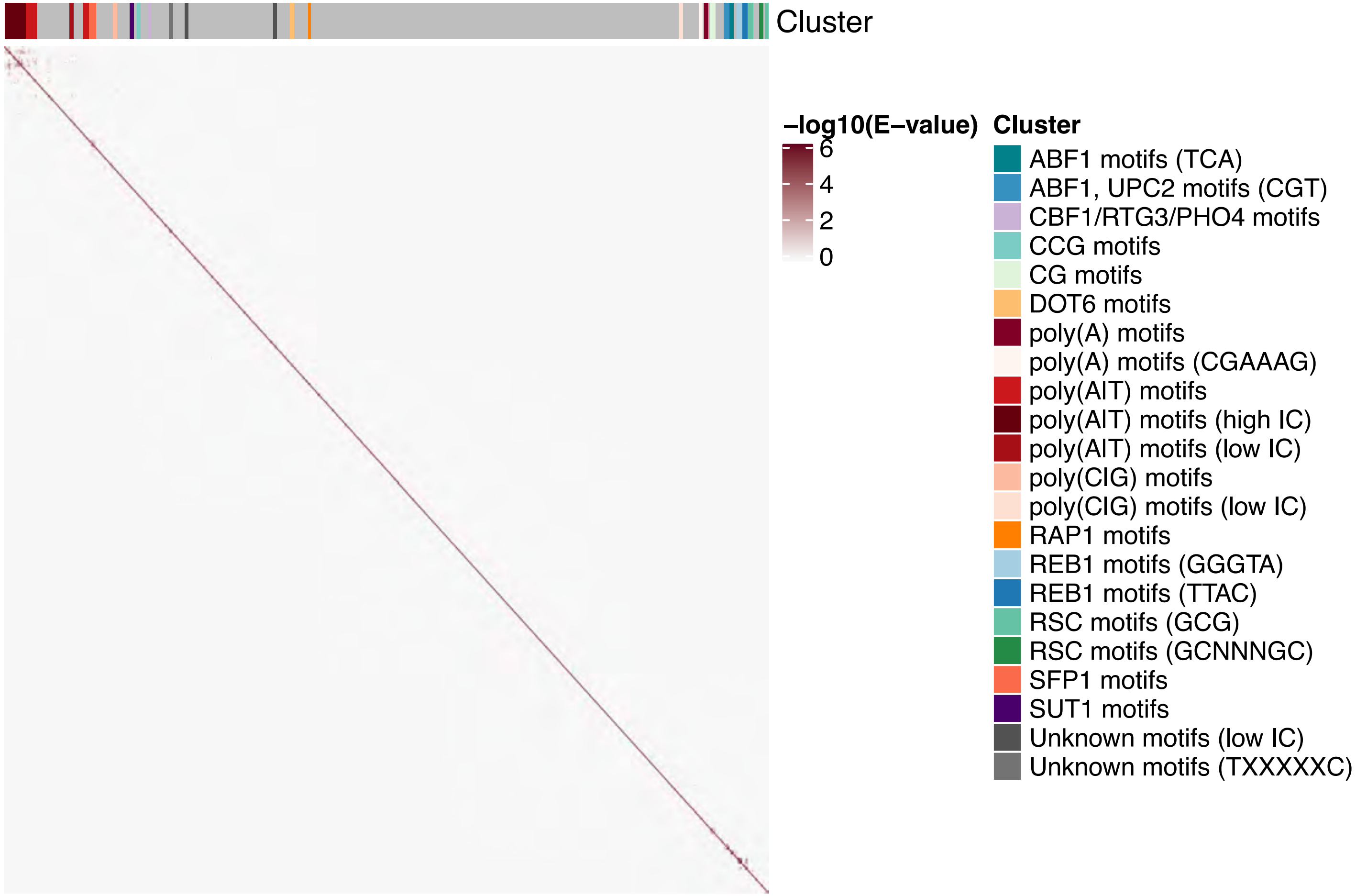

Cluster
-log10(E-value)
6
4
2
0
Cluster
ABF1 motifs (TCA)
ABF1, UPC2 motifs (CGT)
CBF1/RTG3/PHO4 motifs
CCG motifs
CG motifs
DOT6 motifs
poly(A) motifs
poly(A) motifs (CGAAAG)
poly(AIT) motifs
poly(AIT) motifs (high IC)
poly(AIT) motifs (low IC)
poly(CIG) motifs
poly(CIG) motifs (low IC)
RAP1 motifs
REB1 motifs (GGGTA)
REB1 motifs (TTAC)
RSC motifs (GCG)
RSC motifs (GCNNNGC)
SFP1 motifs
SUT1 motifs
Unknown motifs (low IC)
Unknown motifs (TXXXXXC)

Figure S3

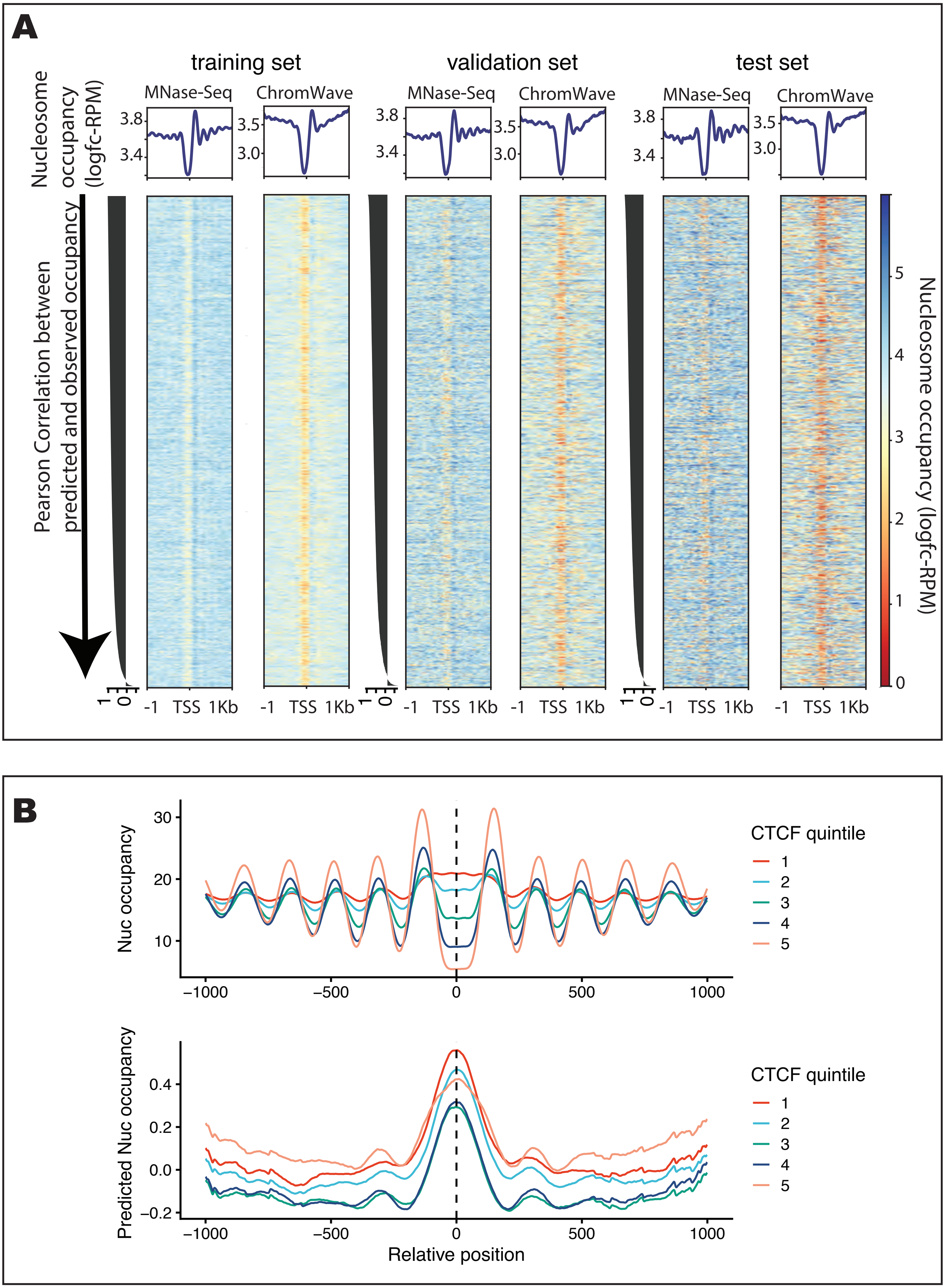

A
training set
validation set
test set
MNase-Seq
ChromWave
Nucleosome occupancy (logfc-RPM)
3.8
3.4
3.5
3.0
Pearson Correlation between predicted and observed occupancy
1
0
-1
TSS
1Kb
Nucleosome occupancy (logfc-RPM)
5
4
3
2
1
0
B
Nuc occupancy
30
20
10
−1000
−500
0
500
1000
CTCF quintile
1
2
3
4
5
Predicted Nuc occupancy
0.4
0.2
0.0
−0.2
Relative position

Figure S4