CHAPTER 11

Quantitative Models of the Mechanisms that Control Genome-Wide Patterns of Animal Transcription Factor Binding

Tommy Kaplan[*,†] **and Mark D. Biggin**[‡]

[*]Department of Molecular and Cell Biology, California Institute of Quantitative Biosciences, University of California, Berkeley, California, USA

[†]School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel

[‡]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA

Abstract
I. Introduction
II. Overview of Model/Algorithm
  A. Alternate Modeling Strategies
  B. Generalized Hidden Markov Models
  C. Experimental Datasets
  D. Model Overview
III. Biological Insights
  A. A Simple Model is Mildly Successful
  B. Allowing Transcription Factor Competition Does Not Improve the Predictions' Accuracy
  C. Expanding the Model to Three Dimensions With Single Nucleus Resolution Has Only Slight Effects
  D. Predicting Nucleosome Location Does Not Improve the Model's Predictive Power
  E. DNA Accessibility Data Greatly Improve DNA binding Predictions
  F. Modeling Direct Cooperative DNA Binding Does Not Affect Model Performance
  G. Implications for Determining Transcription Factor DNA Occupancy *in vivo*
IV. Open Challenges
V. Computational Methods
  A. Generalized Hidden Markov Models
  B. Model-based Simulation of Chromatin

METHODS IN CELL BIOLOGY, VOL 110

0091-679X/10 $35.00
DOI 10.1016/B978-0-12-388403-9.00011-4

C. Thermodynamic Modeling of Protein–DNA Interactions Using Boltzmann Ensembles
D. Optimization of Model Parameters
VI. Glossary
VII. Further Reading
References

Abstract

Animal transcription factors drive complex spatial and temporal patterns of gene expression during development by binding to a wide array of genomic regions. While the *in vivo* DNA binding landscape and *in vitro* DNA binding affinities of many such proteins have been characterized, our understanding of the forces that determine where, when, and the extent to which these transcription factors bind DNA in cells remains primitive.

In this chapter, we describe computational thermodynamic models that predict the genome-wide DNA binding landscape of transcription factors *in vivo* and evaluate the contribution of biophysical determinants, such as protein–protein interactions and chromatin accessibility, on DNA occupancy. We show that predictions based only on DNA sequence and *in vitro* DNA affinity data achieve a mild correlation ($r = 0.4$) with experimental measurements of *in vivo* DNA binding. However, by incorporating direct measurements of DNA accessibility in chromatin, it is possible to obtain much higher accuracy ($r = 0.6$–0.9) for various transcription factors across known target genes. Thus, a combination of experimental DNA accessibility data and computational modeling of transcription factor DNA binding may be sufficient to predict the binding landscape of any animal transcription factor with reasonable accuracy.

I. Introduction

Animal transcription factors each bind to many thousands of DNA regions throughout the genome in cells (Boyer *et al.*, 2005; Georlette *et al.*, 2007; MacArthur *et al.*, 2009; Robertson *et al.*, 2007; Zeitlinger *et al.*, 2007; reviewed by Biggin, 2011). While many of the most highly occupied regions are functional *cis*-regulatory regions and are evolutionarily conserved, many thousands of other genomic regions that are bound at lower levels *in vivo* do not appear to be functional targets (Carr and Biggin, 1999; MacArthur *et al.*, 2009). It is, therefore, a critical challenge to quantitatively predict the DNA binding levels of regulatory transcription factors in cells and to determine the biochemical mechanisms that direct these complex patterns of factor occupancy.

Animal transcription factors recognize short (5–12 bp) sequences of DNA that occur with high frequency throughout the genome (Wunderlich and Mirny, 2009), yet most occurrences of these recognition sites are not detectably bound *in vivo* (Carr and

Biggin, 1999; Li *et al.*, 2008; Liu *et al.*, 2006). There are several mechanisms that could account for this discrepancy between predicted and observed transcription factor DNA binding in cells. Competitive inhibition of binding at those DNA recognition sites that overlap sequences occupied either by other sequence-specific factors (Stanojevic *et al.*, 1991) or nucleosomes (Agalioti *et al.*, 2000; Cosma *et al.*, 1999; Narlikar *et al.*, 2002) could selectively inhibit DNA occupancy at these sites. In addition, direct or indirect cooperative interactions between transcription factors bound at close by recognition sites could increase their occupancy at other genomic locations (Buck and Lieb, 2006; Mann *et al.*, 2009; Miller and Widom, 2003; Zeitlinger *et al.*, 2003). Here we describe a computational modeling strategy that can analyze the relative influence of each of these biochemical mechanisms on the overall pattern of transcription factor DNA binding *in vivo* (Kaplan *et al.*, 2011). A glossary is provided to explain key technical terms used in describing the computational modeling (Section VII).

II. Overview of Model/Algorithm

A. Alternate Modeling Strategies

Many computational algorithms have been developed for predicting *in vivo* DNA binding. Crudely, these studies fall into two categories:

Qualitative models aim at identifying statistically significant occurrences of DNA binding sites or *cis*-regulatory regions (Agius *et al.*, 2010; Ernst *et al.*, 2010; Frith *et al.*, 2001; Granek and Clarke, 2005; Narlikar *et al.*, 2007; Narlikar and Ovcharenko, 2009; Rajewsky *et al.*, 2002; Ramsey *et al.*, 2010; Schroeder *et al.*, 2004; Sinha, 2006; Sinha *et al.*, 2003; Ward and Bussemaker, 2008; Whitington *et al.*, 2009; Won *et al.*, 2010). These computational methods usually rely on modeling the background distribution of transcription factor DNA recognition sites and focus on identifying significant *p*-values, that is, sites where the background hypothesis is rejected. These algorithms can identify a subset of *cis*-regulatory binding sites and provide a putative transcriptional regulatory architecture for an organism by connecting regulators to a set of putative target genes. They are less adequate, however, for predicting the levels of transcription factor DNA occupancy, which has been shown to be critical for relating DNA binding patterns to biological function (Carr and Biggin, 1999; MacArthur *et al.*, 2009).

Quantitative models, on the other hand, estimate the occupancy of a factor along the genome. Statistically, they aim to calculate the binding probability (hence, the percent of time or cells) at which the protein binds a specific DNA locus. These methods are, thus, more suitable for modeling the continuous quantitative landscape of transcription factor DNA occupancy as measured by genome-wide chromatin immunoprecipitation (ChIP) studies. An additional advantage of the quantitative approach is its natural generative probabilistic settings, which allow for easy integration of external data, such as chromatin state, the concentration of transcription factors in cells, or interactions between neighboring proteins (He *et al.*, 2009; He *et al.*, 2010; Roider *et al.*, 2007; Wasson and Hartemink, 2009).

In addition to direct quantitative models of transcription factor DNA binding, a related set of models have focused on predicting the gene transcription patterns driven by predefined DNA *cis*-regulatory regions. These studies generally use thermodynamic models to predict transcription factor DNA binding within known *cis*-regulatory regions as well as the resulting expression patterns driven by these target regions in animal embryos (He *et al.*, 2010; Kazemian *et al.*, 2010; Raveh-Sadka *et al.*, 2009; Segal *et al.*, 2008; Zinzen *et al.*, 2006). Three-dimensional changes in the concentration of regulatory transcription factors result in differential occupancy at the same DNA locus over different nuclei, which in turn results in different expression outputs in each cell. Unfortunately, these models do not explicitly train their models (or test them) on experimental *in vivo* DNA binding data, and limit their scope to predict the expression levels driven by specific target genes or *cis*-regulatory regions across the embryo. Therefore, their success in predicting *in vivo* DNA occupancy cannot be directly assessed.

B. Generalized Hidden Markov Models

Most direct quantitative algorithms for predicting transcription factor DNA binding rely on a probabilistic framework based on generalized hidden Markov models (gHMMs). These models use inference algorithms to estimate the occupancies (or DNA binding probability) of one or more transcription factors across any DNA sequence given their concentrations and protein–DNA binding specificities (Frith *et al.*, 2001; Granek and Clarke, 2005; Kulp *et al.*, 1996; Raveh-Sadka *et al.*, 2009; Segal *et al.*, 2008; Sinha *et al.*, 2003; Wasson and Hartemink, 2009).

We have adopted a form of gHMM for modeling transcription factor-DNA binding *in vivo,* as this class of model offers several advantages. These models have very few parameters and are therefore straightforward to optimize. Unlike most probabilistic graphical models, they offer exact inference of posterior probabilities in linear time, using a forward-backward dynamic programming algorithm (Durbin *et al.*, 1998; Rabiner, 1989). Finally, gHMMs are related to thermodynamic equilibrium models: they view the ensemble of all possible configurations of bound factors along the DNA as a Boltzmann distribution in which each configuration is assigned a weight (or probability) depending on its energetic state; the probability that a factor is bound at a specific location is calculated by summing the probabilities of all configurations in which it is bound (Ackers *et al.*, 1982; Buchler *et al.*, 2003; Granek and Clarke, 2005; Rajewsky *et al.*, 2002; Schroeder *et al.*, 2004; Segal *et al.*, 2008; Sinha, 2006; Wasson and Hartemink, 2009).

On the other hand, gHMMs are limited in their modeling power due to their Markovian property: these models lack any memory for past states, and so when estimating the probability of binding at a certain position the model is agnostic of other (nonoverlapping) DNA binding sites. This prevents this class of models from considering the full context in which DNA binding occurs. We will address this limitation below and offer an approximation to allow a full thermodynamic model using sampling procedures.

C. Experimental Datasets

We demonstrated our approach by modeling the genomic binding of five regulators of early embryonic anterior-posterior (A-P) patterning in *Drosophila melanogaster*: Bicoid (BCD), Caudal (CAD), Hunchback (HB), Giant (GT), and Kruppel (KR).

ChIP-seq data for the five factors in stage 5 blastoderm embryos were used to provide the measure of *in vivo* DNA occupancy (Bradley *et al.*, 2010). 20-bp-long sequence reads were mapped to the genome (Apr. 2006 assembly, BDGP Release 5). To minimize mapping noise, we only considered reads uniquely mapped to the genome with up to one mismatch. The mapped reads were then extended according to their orientation to a length of 150 bp, and binned (down-sampled) to a 10 bp resolution. Finally, the genomic binding landscape of each factor was smoothed using a running window of 10 bins (or 100 bp), to account for sampling noise.

DNA binding affinities of the five factors (expressed as position weight matrices – PWMs) were derived from previous *in vitro* measurements that used SELEX-Seq (Berkeley *Drosophila* Transcription Network Project, unpublished data; (MacArthur *et al.*, 2009) (Fig. 1). The PWM counts were normalized to probabilities,

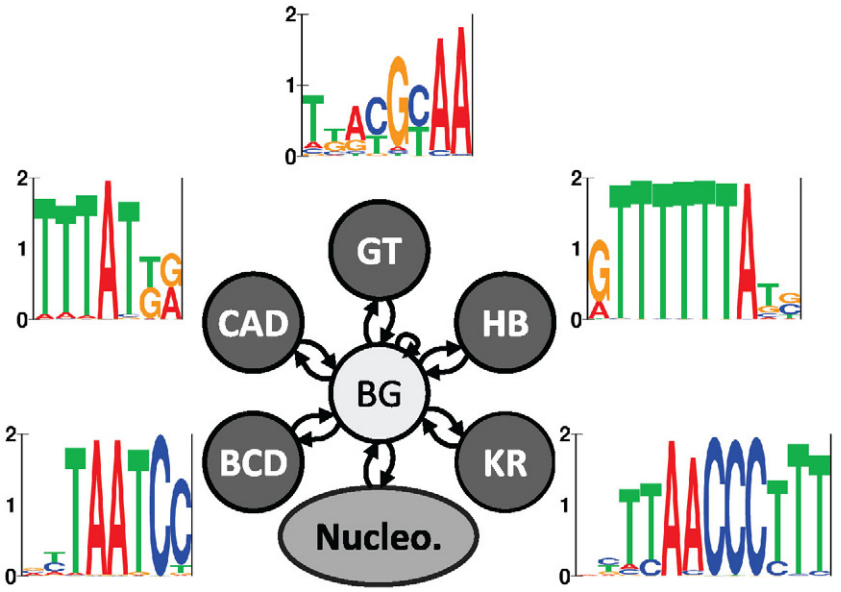

Fig. 1 The generalized hidden Markov model. Diagram of the model states, including the unbound background (BG) state, five states corresponding to the five transcription factors in the model (BCD, CAD, GT, HB, and KR), and a 141-bp-long nucleosomal binding state (Nucleo.). The emission probabilities of each transcription factor state are visualized using sequence logos that are based on position weight matrices (PWMs). (See color plate.)

after adding a pseudo-count of 0.01 to avoid zero probabilities (available at http://bdtnp.lbl.gov/gHMM). Additional sources of PWMs (Noyes *et al.*, 2008; Segal *et al.*, 2008) were also tested in the model, yielding similar results (Kaplan *et al.*, 2011). In all cases, the DNA binding specificities defined by these various experiments (i.e., the PWMs) were maintained and were not optimized as parameters in the model. We found that the tradeoff between having potentially more accurate PWMs (and a better fit to experimental *in vivo* DNA binding data) versus the cost of additional parameters to optimize was not beneficial for most factors. Moreover, using fixed PWMs from external studies prevents overfitting or drift toward additional motifs that are often present near developmental regions, such as the CAGGTAG sequence known to be bound by the transcription factor Zelda (Bradley *et al.*, 2010).

The accessibility of DNA in chromatin was obtained from DNase-seq data resulting from the DNase I digestion of isolated stage 5 blastoderm embryo nuclei (Li *et al.*, 2011; Thomas *et al.*, 2011). 34-bp-long reads were mapped to the genome by requiring unique matches with no more than two mismatches, then these were extended to a length of 150 bp, binned (down-sampled) to a 10 bp resolution, and smoothed using a running window of 10 bins.

Estimates of transcription factor protein concentrations in each nucleus were derived from three-dimensional fluorescence microscopy of *D. melanogaster* embryos at early stage 5 (Fowlkes *et al.*, 2008).

All of the above experimental datasets are available from the supplemental website for Kaplan *et al.* (2011) at http://bdtnp.lbl.gov/gHMM.

D. Model Overview

Hidden Markov models are probabilistic frameworks where the observed data (such as, in our case the DNA sequence) are modeled as a series of outputs (or emissions) generated by one of several (hidden) internal states. The model then uses inference algorithms to estimate the probability of each state along every position along the observed data. In our case, the model is composed of the various states that the DNA could be in: unbound (the background state), bound by transcription factor t_1, bound by transcription factor t_2, etc., or wrapped around a nucleosome (Fig. 1). Each state holds some probability distribution of the DNA sequences it favors (and emits according to the HMM). In our case, the background state is derived using the simple mononucleotide (single base) probability (frequency) in the genome to model the A/T distribution along the noncoding parts of the genome. The "bound" states hold a probabilistic DNA model that represents the sequences that each protein prefers to bind (its recognition sites). Additional parameters of the gHMM include the prior probabilities of entering each state, which are modeled using the transition probabilities between states. For example, a highly expressed protein that is more likely to be in the bound state along the DNA will have a higher transition probability than a protein present at lower concentrations in cells. Once the parameters of the gHMM are optimized (using a held-out set of training sequences) and given a new DNA sequence, it is straightforward to infer the probability of each state (unbound,

bound by factor t_1, bound by factor t_2, etc.) at each position along the sequence. See Section V for further details of these models.

All our computational models estimate the DNA binding probability of each transcription factor at a single-nucleotide resolution. A model-based algorithm is then used to transform these predictions into smoothed ChIP-like landscapes so they can be compared to the *in vivo* ChIP-seq measurements of protein–DNA binding (Fig. 2). For this, the length distribution of DNA fragments recovered by the ChIP process is used to simulate the overall shape of one peak, corresponding to a single DNA binding event measured by ChIP-seq. For a length distribution *c(l)*, the estimated shape F of a peak is described as:

$$F(\Delta_x) \propto \sum_{l=\Delta_x}^{\infty} c(l)$$

where Δx denotes the relative distance from the binding locus or peak center. In other words, the probability of obtaining a read Δx bp away from the binding event is proportional to the total number of reads at least Δx bp long (Capaldi *et al.*, 2008; Kaplan *et al.*, 2011).

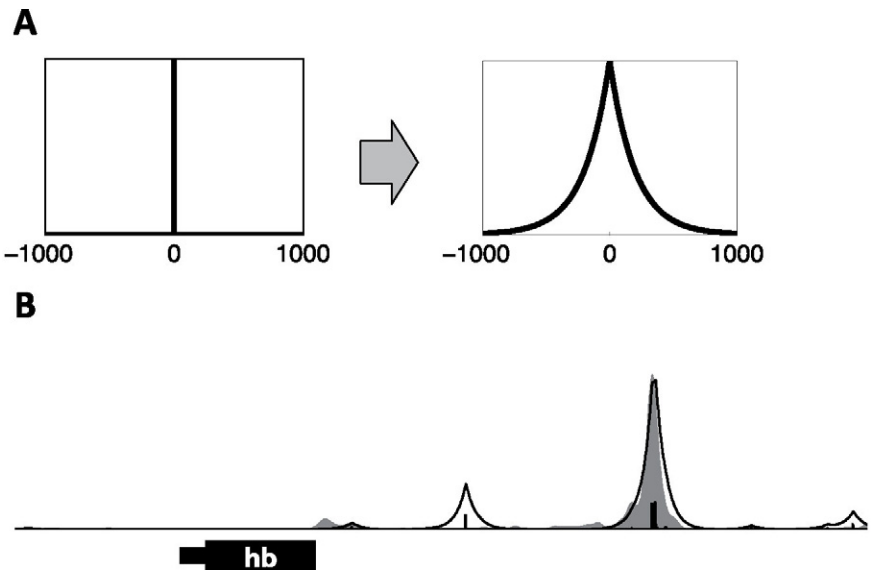

Fig. 2 From DNA binding probabilities to ChIP landscape. **(A)** Each DNA binding event (left) was transformed to a model-based estimation of expected ChIP peak shape based on the average length of the DNA fragments immunoprecipitated in the ChIP experiment (right) (Kaplan *et al.*, 2011). **(B)** This model was then used to convolve the model's binding predictions (vertical black bars) to the expected landscape of ChIP sequencing assay (thin black line), which was then compared to the measured *in vivo* DNA binding landscape (gray shaded landscape).

III. Biological Insights

A. A Simple Model is Mildly Successful

We began with the simplest model – a single transcription factor binding to DNA. This required optimizing only a single parameter, *P(t)*, for each transcription factor that corresponds to its effective concentration in nuclei and assuming, for this first simple case at least, that the protein is expressed at the same concentration in all embryo cells. We used standard optimization techniques (based on a combination of genetic algorithms and gradient ascent-based algorithm) to optimize these parameters (see section V). For each tested value of *P(t)*, we used the generalize hidden Markov model to estimate the binding probability per position, and then convoluted these predictions into the predicted DNA binding landscape.

To analyze our predictions, we compiled a list of 21 known target loci of the A-P patterning system. Each target gene was expanded by ~10 Kb upstream and downstream of the transcription unit to capture its *cis*-regulatory regions. In each analyses presented in this chapter, we trained the model parameters to optimize the fit between the predicted and the observed ChIP-seq landscapes at a set of six loci, which spanned ~87 Kb, and evaluated the trained model on the remaining set of 15 loci, which spanned ~280 Kb. To account for long genomic regions where no DNA binding is observed *in vivo* by ChIP, the training and test sets were enhanced by addition of three or five control regions, spanning a total of 100 and 221 Kb, respectively (Kaplan *et al.*, 2011).

After parameter optimization using the training set, the model was applied to the test set. The predicted DNA binding landscape around one gene in the test set is shown in Fig. 3A. The total correlation between the model predictions and measured data was quite weak when averaged over all ~500 kb of the test set ($r = 0.36$), with specific factors varying from $r = 0.15$ (GT) to $r = 0.66$ (BCD) (Kaplan *et al.*, 2011).

In addition to estimating accuracy using the correlation between the model's predictions and experimentally measured *in vivo* DNA binding, we also tried two alternatives. In one, we used distance-based measures such as the root mean square deviation (RMSD) between the predicted and measured genomic landscapes. In the second, we tried a peak-centric comparison method, where a peak calling algorithm was used to identify "bound regions" in both the predicted and the measured data and then the overlap between called peaks was compared. These alternate scoring methods resulted in qualitatively similar results to the correlation coefficients given in the rest of the text and in Fig. 4.

B. Allowing Transcription Factor Competition Does Not Improve the Predictions' Accuracy

Encouraged by the results with each transcription factor considered singly, we examined the effect of DNA binding site competition between the five factors on our ability to predict *in vivo* DNA occupancy. Overlapping DNA recognition sites can

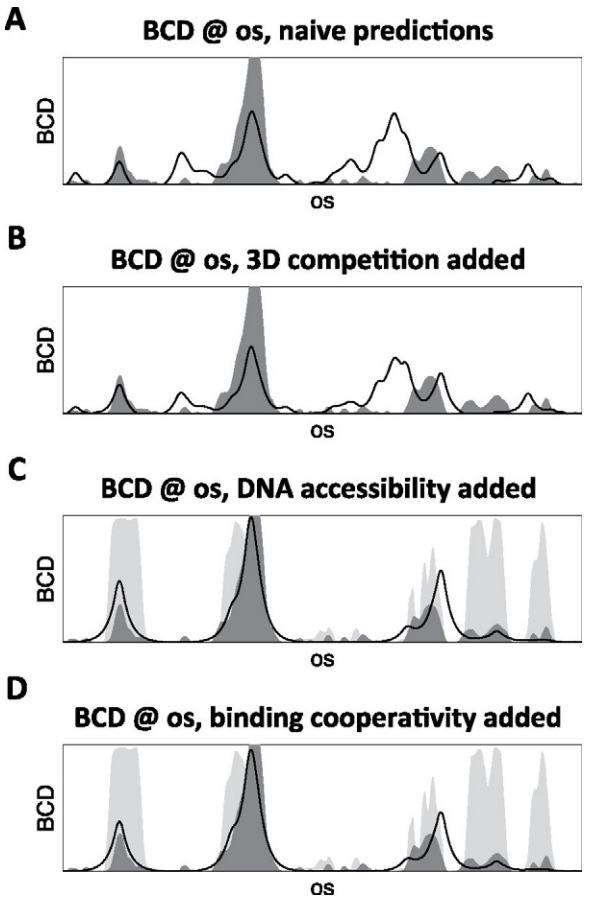

Fig. 3 High-resolution predictions of protein–DNA binding landscape. **(A)** The model's DNA binding predictions (thin black line) for BCD are compared to the measured *in vivo* DNA binding landscape (dark shaded landscape) across the 15 Kb around the *os* locus. In this example, the BCD binding landscape was predicted without considering the other transcription factors. **(B)** Same as (A), except that direct DNA binding competition between the five factors and with nucleosomes was allowed, and BCD binding was modeled independently in each of 6,078 nuclei of the stage 5 blastoderm embryo. **(C)** Same as (B), but also incorporating a nonuniform DNase I hypersensitivity-based prior on transcription factor binding to account for variations in DNA accessibility (shown as light shaded landscape). **(D)** Same as (C), but further adding cooperative interactions between adjacently bound transcription factor molecules in a thermodynamic setting.

allow direct competition between transcription factors (Stanojevic *et al.*, 1991). Moreover, overlapping sites are often conserved at long evolutionary distances, suggesting an important role for inter-factor competition (Hare *et al.*, 2008). Therefore, we expanded the gHMM in our model to consider all five transcription

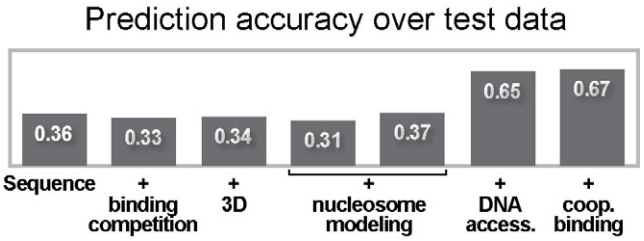

Fig. 4 Prediction accuracy at increasing degrees of model complexity. Accuracy of DNA binding predictions for the test set of 15 known A-P targets and five control loci. Shown are the correlation coefficients between model prediction and measured *in vivo* DNA binding landscape for increasing degrees of model complexity. These are, from left to right: independent predictions per transcription factor using our simplest model; allowing DNA binding site competition between transcription factors; making predictions at a single-nucleus resolution; including nucleosomes using a sequence-specific or a sequence-independent model of nucleosome binding; adding a nonuniform prior on transcription factor binding using DNA accessibility measurements; and adding cooperative DNA binding interactions in a thermodynamic setting.

factors simultaneously in a probabilistic framework (Fig. 1), where the concentrations of each factor *t* is modeled by an additional probabilistic term *P(t)*. In the single transcription factor model, binding of one protein to a recognition site did not affect the DNA occupancy of a different transcription factor at an overlapping site. In this new model, however, because the total occupancy at a site cannot exceed 1, transcription factors effectively compete for DNA binding to overlapping recognition sites. Surprisingly, this competitive model gave slightly less accurate predictions than its single factor counterpart. On the test data, the model's predictions decreased from a total correlation of 0.36 to 0.33 (see Fig. 4).

C. Expanding the Model to Three Dimensions With Single Nucleus Resolution Has Only Slight Effects

One reason why the model did not improve when competition was allowed could have been that, because we treated the embryo as a homogenous entity, the model allowed competition between transcription factors that are not expressed together at high levels in the same cells. We therefore expanded our algorithm to model the DNA binding of all transcription factors in each of the ~6000 nuclei of the embryo separately. To scale the optimized concentration parameters of the five transcription factors for each nuclei, we further scaled the prior probability *P(t)* of every transcription factor *t* proportionally to its protein expression level, as measured at a single-cell resolution (Fowlkes *et al.*, 2008). We then averaged the predicted DNA binding landscape of all nuclei to obtain whole-embryo genomic predictions, which were then compared to the (whole-embryo average) *in vivo* DNA binding measurements from ChIP-seq (Kaplan *et al.*, 2011). This slightly improved the predictions relatively to the whole-embryo predictions (Figs. 3B and 4). However, combining

DNA binding site competition and 3D expression data yields a model that is only about as effective as the simplest model. Thus, while competition between transcription factors is likely important at a subset of recognition sites, it does not appear to be a principal determinant of the overall distribution of transcription factor DNA occupancy *in vivo*.

D. Predicting Nucleosome Location Does Not Improve the Model's Predictive Power

To test if chromatin state influences the accuracy of our model, we first attempted to predict the locations of nucleosomes to enable modeling of the competition between transcription factors and nucleosomes in binding to DNA (Narlikar *et al.*, 2007; Raveh-Sadka *et al.*, 2009; Wasson and Hartemink, 2009). As there are no direct measurements of nucleosome positions from early *Drosophila* embryos, we modeled these computationally. We extended our Markov model to represent the sequence bound by a single nucleosome. This was done by including an additional state in the gHMM that comprised a sequence-independent model of nucleosome DNA binding in which nucleosomes are viewed as long "space-fillers" that, when present, prevent regulators from binding to DNA. We used a 141-bp long model of nucleosome binding, based on a fixed distribution of nucleotides as in the background state P_B of the Markov model (0.32 for A/T, 0.18 for G/C). Similarly to the transcription factor states, the nucleosomal state was assigned a prior probability term *P(t)* to reflect a fixed nucleosomal concentration along the embryo. *P(t)* was optimized together with other concentration-related parameters *P(t)* for all transcription factors. Alternatively, due to uncertainty in the literature about the contribution of DNA sequence specificity to *in vivo* nucleosome positioning, we also tested a sequence-specific model of nucleosome binding (Segal *et al.*, 2006). Neither of these nucleosomal models dramatically improved the DNA binding predictions for the five transcription factors (Fig. 4).

E. DNA Accessibility Data Greatly Improve DNA binding Predictions

A weakness of the above strategies to predict nucleosome location is that only one constitutive model is derived for all cells of the organism for all stages of development. Yet it is known that chromatin accessibility varies dramatically over time and between cells (Kharchenko *et al.*, 2011; Thomas *et al.*, 2011). Therefore, we sought to exploit direct genome-wide measurements of DNA accessibility for the same developmental stage from which the ChIP-seq data were derived (Li *et al.*, 2011; Thomas *et al.*, 2011). Interestingly, when we compared these DNA accessibility data to the predictions of the original, simple version of our gHMM, we found that the model correctly predicts DNA binding on the most highly accessible genomic regions but tended to predict stronger DNA binding than was actually measured on less accessibility regions (Kaplan *et al.*, 2011). We therefore leveraged the statistical framework of generalized hidden Markov models and incorporated

DNA accessibility data into the model as a nonuniform prior probability of regulatory binding along the genome – with regions of low accessibility being given a greatly reduced probability of binding.

The incorporation of differential DNA accessibility in this way dramatically boosted the model's accuracy by almost twofold to a correlation of $r = 0.67$ with the measured *in vivo* occupancy data when averaged over all the ~500 kb test set, with the factor-specific correlation varying from 0.58 (HB) to 0.79 (BCD) (Figs. 3C and 4).

In addition to the sigmoidal prior described in Section V, we investigated additional methods to transform the DNA accessibility data DD_x into probabilities PD_x. First, we tried to linearly scale the accessibility data DD_x and limit the maximal PD_x values at one. This resulted with slightly less accurate predictions ($r = 0.66$ on test data). Also, we tried an even simpler model using a step function, namely modeling PD_x as one value below some minimal value of DD_x, and another value above it. Even this naive model achieved comparable accuracy, at $r = 0.64$. This slightly reduced correlation suggests that the effect of DNA accessibility on transcription factor binding may be almost binary – low accessibility regions show almost no regulatory binding, while binding at accessible regions is modeled quite accurately by DNA sequence alone (Kaplan *et al.*, 2011).

F. Modeling Direct Cooperative DNA Binding Does Not Affect Model Performance

Although our predictions that included DNA accessibility data were reasonable, we sought to further refine our model by considering factor-factor interactions other than the simple direct competition (via overlapping recognition sites) described earlier. For example, direct physical interactions between transcription factors bound at neighboring recognition sites have often been found to increase the occupancy of one or both proteins on DNA, for both homomeric and heteromeric cooperative interactions, and to sharpen the regulatory response to changes in transcription factor concentration (Arnosti *et al.*, 1996; Small *et al.*, 1992).

Generalized hidden Markov models, however, have limited ability to model the broader context of DNA binding events, including cooperative interactions between neighboring sites. We therefore added a second, sampling-based phase to our computational model. In this phase, a large ensemble of DNA binding configurations is sampled, each with a different set of protein–DNA interactions. The probability of each configuration is then estimated based on all pairs of nearby occupied sites (up to 95 bp apart) and the parameterized energetic gain of each pair. Finally, the overall DNA binding probability at each position is quantified as a weighted sum of all sampled configurations.

By adopting a statistical mechanics perspective, the exponential space of protein–DNA binding configurations can be viewed as a canonical ensemble in a thermodynamic equilibrium. Here, the probability of each configuration is directly linked to its energetic state, including direct protein–DNA interactions, steric hindrance constraints, and cooperative interactions with neighboring factors (Ackers *et al.*, 1982; Segal *et al.*, 2008). We extended our model to capture cooperative interactions between transcription factors using a novel set of 15 parameters (one for each

nonredundant pair of the five factors), modeling the energy gain for the nearby binding of every possible pair of the five transcriptional regulators in our model.

The optimized set of cooperative DNA binding parameters includes predictions of interactions between many homomeric and heteromeric pairs (Kaplan *et al.*, 2011). These cooperativity parameters improved the predictive power of the model to a correlation of $r = 0.67$ on the test data, ranging from $r = 0.58$ (HB) to $r = 0.79$ (BCD), a marginal improvement over the Markovian approach (Figs. 3D and 4). Thus, our model suggests that cooperative interactions between transcription factor molecules have a rather limited contribution in shaping the genomic landscape of *in vivo* DNA binding (Kaplan *et al.*, 2011).

G. Implications for Determining Transcription Factor DNA Occupancy *in vivo*

The increasing availability of genome-wide *in vivo* measurements of DNA accessibility (via DNase I, FAIRE) for a variety of cell types, developmental stages, and environmental conditions, together with the laborious nature of direct ChIP measurements, suggests a mixed computational-experimental streamlined strategy for estimating the genome-wide binding landscape of proteins. While we often fail to predict transcription factor DNA binding levels from DNA sequence and *in vitro* DNA affinity measurements alone, by incorporating DNA accessibility data into a thermodynamic model, a reasonable job of quantitatively predicting the occupancy of transcription factors can be made. While such an approach should not be viewed as a substitute for systematic experimental measurement of transcription factor DNA binding *in vivo*, we believe our predictions are good enough to be useful when such experimental data are unavailable or impractical to obtain.

IV. Open Challenges

Quantitative computational models of sequence-specific protein–DNA interactions offer a fast approximation of the genomic landscape of protein–DNA binding. Nonetheless, these predictions are still far from being reliable enough to fully replace experimental *in vivo* measurements.

One of the greatest challenges for improving future models is in modeling locus-specific DNA accessibility using genomic DNase I hypersensitivity data. Our current models rely on a probabilistic platform, in which we tested various ways to transform read coverage into *a priori* DNA binding probabilities, with a sigmoid function being the most useful. While this approach worked well on relatively accessible regions (*cis*-regulatory regions and the regions flanking actively expressed genes), it was not as accurate on a full genomic scale, giving a correlation coefficient of only 0.33 for an entire chromosome arm (Kaplan *et al.*, 2011). Most false predictions arose from *bona fide* sites predicted to be strongly bound, but which show limited or no binding *in vivo* due to limited accessibility. In addition, we observed some highly accessible regions bound by several transcription factors, even in the absence of cognate sequence recognition sites. We believe that

optimizing the transformation from DNase I read densities into binding probabilities at very low and very high DNase-seq read densities could strengthen the model.

A second challenge is to improve the modeling of cooperative DNA binding (both direct and indirect). In our work to date, we applied a somewhat simple approach, where two nearby transcription factor molecules contribute some constant energetic value only if they bind in close proximity (<95 bp). It seems probable that more sophisticated methods, with a greater number of parameters, could model the biological/physical effect with greater accuracy.

Wasson and Hartemink (2009) recently used hidden Markov models to analyze transcription factor DNA binding in yeast and showed that their predictions improve as more sequence-specific transcription factors are added to the model. While we did not observe this trend with our data, possibly because we only analyzed five transcription factors, revisiting this approach with a greater number of transcription factors could be revealing.

Finally, while the direct goal of the work described in this chapter was to predict *in vivo* DNA binding from DNA sequence, *in vitro* affinity, and chromatin accessibility data, a more challenging question is to understand and predict *de novo* how dynamic patterns of DNA accessibility are themselves generated in cells. This may require correctly modeling the activities of hundreds of sequence-specific transcription factors, the chromatin remodeling proteins that they recruit, nucleosomes, and other chromatin proteins. We doubt that sufficient data or knowledge is available to yet take up this task.

V. Computational Methods

A. Generalized Hidden Markov Models

Generalized hidden Markov models were used to predict transcription factor DNA binding based on the factor concentration and the DNA sequence. We followed a thermodynamic rationale, and considered the space of all valid DNA binding configurations as a Boltzmann distribution. Under this statistical framework, the probability of each configuration, P_i, is proportional to its energetic state E_i

$$P_i \propto e^{-\beta E_i}$$

where β equals $1/k_B T$, with k_B being the Boltzmann constant and T the temperature (25 °C).

The energetic state of each configuration could therefore be calculated from its binding probability. Under this model, bound nucleotides are generated according to the protein–DNA binding preference, or PWM, of the transcription factor. The probability of a subsequence S_i to be bound by transcription factor t equals

$$P_t(S_i) = P(t) \prod_{j=0}^{l_t-1} P_j(S_{i+j}|\theta_t)$$

with $P(t)$ being the *a priori* binding probability of transcription factor t, l_t the length of the binding site for factor t, and $P_j(S_{i+j}|\theta_t)$ corresponds to the probability of the nucleotide S_{i+j}, at the j position of a binding site for factor t, as modeled by its recognition parameters θ_t. Unbound nucleotides are generated from a mononucleotide background distribution P_B (0.32 for A/T, 0.18 for G/C).

It is useful to visualize this family of models as a series of probabilistic transitions between the internal states of the model (Fig. 1). The different types of DNA sequence (unbound DNA; DNA bound by factor t, etc.) are the nodes, and the allowed transitions between states are shown as arrows in the figure. The parameters of the model correspond to the probabilities of transition between states. Each state is associated with one transcription factor; the probability of the corresponding DNA subsequence is calculated using its binding site model $P_j(S_i)$. Each configuration is viewed as one path along the internal states of the model, starting in one state at the beginning of the DNA sequence, and transitioning among the states until the end of the sequence. The full binding configuration of DNA sequence S, with multiple factors $t_1,\ldots,t_k$ bound at positions $x_1\ldots x_k$, respectively, is viewed as a path that loops into the unbound state along most of the DNA sequence except for positions $x_1\ldots x_k$ where it enters the states corresponding to the transcription factors $t_1,\ldots,t_k$. We can then write the probability of this path as:

$$P(S) = P_B(S)\prod_{i=1}^{k} P(t)\frac{P_{t_i}(S_{x_i})}{P_B(S_{x_i})}$$

Note that no overlapping binding sites are allowed in each configuration. To further account for steric hindrance, each PWM was extended by two flanking regions of 3 bp. These were modeled by a nonspecific background distribution P_B (0.32 for A/T, 0.18 for G/C). The minimal distance between two occupied sites in one binding configuration is therefore 7 bp (two 3 bp flanks plus a 1 bp transition through the unbound state).

To infer the overall binding probability of each transcription factor at each DNA position, one must account for the exponentially large number of possible configurations, while weighting each configuration based on its probability. While this task seems difficult at first, it can be solved in a linear time using the dynamic programming inference algorithm (Durbin *et al.*, 1998; Rabiner, 1989). Specifically, we use the forward-backward algorithm. First, we calculate the local probabilities of each transcription factor t to bind DNA at each position i, $U_{t,i} = P(t) * P_t(S_i)$. We then calculate the Forward Potentials $F_{t,i}$ and the Backward Potentials $B_{t,i}$ by summing the probabilities of all configurations (paths) that end (for Forward Potentials) or begin (for Backward Potentials) at position i with a binding site of t. Finally, we calculate the exact *a posteriori* probability of transcription factor t bound at position i by multiplying the forward and backward potentials. This calculates the binding probability of factor t at position i, given all possible combinations of other transcription factors along the entire sequence S.

B. Model-based Simulation of Chromatin

We used a sigmoid transformation to convert the genomic landscape of DNase I hypersensitivity data DD_x (density of sequenced reads along the genome) into the *a priori* probability PD_x of entering a bound state at position x:

$$PD_x = \frac{1}{1 + e^{-\beta\, DD_x + a}}$$

The parameters of this equation, $\alpha = 6.008$ and $\beta = 0.207$, were optimized over the training data, separately from the concentration parameters in an iterative manner (piecewise optimization). Those probabilities PD_x are then multiplied by the prior probability of binding *P(t)* for each transcription factor t in order to calculate the actual transition probability into the bound state of transcription factor t at position x along the genome.

C. Thermodynamic Modeling of Protein–DNA Interactions Using Boltzmann Ensembles

To predict transcription factor DNA binding in a full thermodynamic setting we first used the generalized hidden Markov model to analyze the underlying sequence and predict proteins' DNA binding according to the different protein concentrations within each nucleus in the *Drosophila* blastoderm stage embryo. This was used to calculate an approximate map of DNA binding. To allow for cooperative interactions between the transcription factor molecules, we then used the DNA binding probabilities described above to sample 10,000 binding configurations per sequence/run and reweighted them to account for the energetic gain due to cooperative DNA binding interactions. This was done in a thermodynamic setting, where every configuration i was reweighted by W_i

$$W_i = \exp\left(- \sum_{|x_j - x_k| < 95} C_{j,k} \right)$$

where x_j and x_k are the binding locations of factors j and k, while $C_{j,k}$ corresponds to their optimized cooperativity parameter. The reweighted samples are then averaged, and the binding probability of every factor at every position is calculated. This combination of direct gHMM calculations followed by importance-weighted sampling allows us to approximate the full thermodynamic landscape of binding using a fast framework with few parameters.

D. Optimization of Model Parameters

We optimized all the parameters in our models by focusing on the train set loci and maximizing the correlation among the model predictions and the *in vivo* measurements of transcription factor DNA binding. The prior probabilities *P(t)* of entering into the bound state for each transcription factor, which reflect the nuclear protein

concentration of each factor, were first optimized by a genetic optimization algorithm (Goldberg and Holland, 1988) with 25 generations and a population size of 15. We then further optimized the *P(t)* variables using a gradient-based trust-region algorithm (Steihaug, 1983).

VI. Glossary

***Qualitative Models of DNA Binding Sites*:** A family of computational models aimed at identifying transcription factor binding sites along a given DNA sequence.

***Quantitative Models of DNA Binding*:** A family of computational models aimed at estimating the occupancy of DNA-binding proteins along the positions of a given DNA sequence. For example, given an input sequence, a qualitative model may identify two putative recognition sites, while a quantitative model may predict that one of these sites is occupied twice as often (i.e., for longer periods of time) as the other.

***Position Weight Matrix (PWM)*:** A statistical representation of a DNA motif. Commonly used to model the DNA recognition element of a transcription factor, a PWM is a table of 4-by-*N* that records the probability of observing each of the four nucleotides at every position of the motif. These models assume independence between the *N* positions of the motif such that each nucleotide position is represented as a single column with the estimated probabilities for each of the four nucleotides. To calculate the probability of transcription factor binding at a DNA word of size N given the PWM, the probabilities given in the cells of the table that correspond to the nucleotide at each of the N positions of the word are multiplied.

***Background Model of DNA*:** A statistical representation of DNA sequences, typically used as a negative control when scanning DNA for sequence motifs. These models typically model only the general nucleotide (A-T content) of the DNA and as a result are too weak to model the entire length of a sequence-specific binding site for transcription factors.

***Thermodynamic Model*:** *According* to statistical thermodynamics, the relative amount of time a complex system with multiple states would spend in each state is related to its energetic states. Using a Boltzmann distribution, the energetic state of each configuration is used to estimate the probability of the system being in each state. For example, every position along a DNA sequence could be bound by many transcription factors, but it is more likely the system is usually in a "stable" state – such as no binding at all or binding of one or more proteins at their higher-affinity recognition sites.

***Hidden Markov Model (HMM)*:** A probabilistic framework for modeling a series of observations (in our case a DNA sequence) using a series of unobserved transitions between the internal states of the model. The parameters of the model include the probabilities of transition between the various states (the transition probabilities), and the probabilities for each of the possible outputs of each state (the emission probabilities). Using inference algorithms, HMMs are used to efficiently find the most probable explanation (path over the states of the model) of the data, or to infer the posterior probability of a given state at a given position.

***Generalized Hidden Markov Model (gHMM)*:** An extended class of HMMs that allow states with longer outputs as well as mute states with no output at all. We employ a gHMM with "bound" states that use PWM to model sequence-specific binding sites, and a "not-bound" state that uses a background model of DNA nucleotide distribution. Given a DNA sequence, we use the gHMM to infer which positions along the sequence are likely to correspond to the "bound" states and to what extent.

***Prior and Posterior Probabilities*:** In Bayesian statistics, the prior and posterior probabilities estimate the likelihood of an event before or after we take evidence into account, respectively. For example, the prior probability of a given state in model corresponds to how often we believe a given transcription factor binds DNA in general, while the posterior probability of the protein's binding depends of the actual sequence of the DNA.

***Dynamic Programming*:** A class of algorithms in computer science that solve certain problem by breaking them down into simpler overlapping subproblems.

***Forward-Backward Algorithm*:** A dynamic programming inference algorithm for calculating the posterior probability of all states at all the positions of an input series of observations. Here, we use the algorithm to estimate the posterior binding probability of each transcription factor along a sequence of DNA. First, the algorithm calculates the probabilities of each state at any position given the DNA sequence from the start until that point (forward probabilities). It then calculates the probabilities of all states given the remaining part of the DNA sequence (background probabilities). Finally, these are combined to produce the posterior probability given the full sequence.

Further Reading

More details on our model and the implications of our analysis for transcription factor DNA binding can be found in Kaplan *et al.* (2011). A companion paper providing additional biochemical arguments suggesting that chromatin accessibility plays a more important role than direct heteromeric cooperative association between transcription factors in directing factor binding in cells can be found in Li *et al.* (2011). Finally, Biggin (2011) comprehensively reviews the relationship between the continuum of transcription factor DNA occupancy levels seen in animal cells and biological function.

References

Ackers, G. K., Johnson, A. D., and Shea, M. A. (1982). Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl. Acad. Sci. USA.* **79**, 1129–1133.

Agalioti, T., Lomvardas, S., Parekh, B., Yie, J., Maniatis, T., and Thanos, D. (2000). Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* **103**, 667–678.

Agius, P., Arvey, A., Chang, W., Noble, W. S., and Leslie, C. (2010). High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Comput. Biol.* 6.

Arnosti, D. N., Barolo, S., Levine, M., and Small, S. (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214.

Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* **21**, 611–626.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956.

Bradley, R. K., Li, X. -Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., Tonkin, L. A., Biggin, M. D., and Eisen, M. B. (2010). Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *Plos Biol.* **8**, e1000343.

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. USA.* **100**, 5136–5141.

Buck, M. J., and Lieb, J. D. (2006). A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nat. Genet.* **38**, 1446–1451.

Capaldi, A., Kaplan, T., Liu, Y., Habib, N., Regev, A., Friedman, N., and O'Shea, E. (2008). Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat. Genet.* **40**, 1300–1306.

Carr, A., and Biggin, M. D. (1999). A comparison of in vivo and in vitro DNA-binding specificities suggests a new model for homeoprotein DNA binding in Drosophila embryos. *EMBO J.* **18**, 1598–1608.

Cosma, M. P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell* **97**, 299–311.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK New York.

Ernst, J., Plasterer, H. L., Simon, I., and Bar-Joseph, Z. (2010). Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* **20**, 526–536.

Fowlkes, C. C., Hendriks, C. L. L., Keränen, S. V. E., Weber, G. H., Rübel, O., Huang, M. -Y., Chatoor, S., Depace, A. H., Simirenko, L., Henriquez, C., Beaton, A., Weiszmann, R., Celniker, S., Hamann, B., Knowles, D. W., Biggin, M. D., Eisen, M. B., and Malik, J. (2008). A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell.* **133**, 364–374.

Frith, M. C., Hansen, U., and Weng, Z. (2001). Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* **17**, 878–889.

Georlette, D., Ahn, S., MacAlpine, D. M., Cheung, E., Lewis, P. W., Beall, E. L., Bell, S. P., Speed, T., Manak, J. R., and Botchan, M. R. (2007). Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb MuvB/dREAM complex in proliferating cells. *Genes Dev.* **21**, 2880–2896.

Goldberg, D., and Holland, J. (1988). Genetic algorithms and machine learning. *Machine Learning* **3**, 95–99.

Granek, J. A., and Clarke, N. D. (2005). Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.* **6**, R87.

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., and Eisen, M. B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet.* **4**, e1000106.

He, X., Chen, C. C., Hong, F., Fang, F., Sinha, S., Ng, H. H., and Zhong, S. (2009). A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One* **4**, e8155.

He, X., Samee, M. A., Blatti, C., and Sinha, S. (2010). Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput. Biol.* 6.

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.

Kazemian, M., Blatti, C., Richards, A., McCutchan, M., Wakabayashi-Ito, N., Hammonds, A. S., Celniker, S. E., Kumar, S., Wolfe, S. A., Brodsky, M. H., and Sinha, S. (2010). Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials. *PLoS Biol.* 8.

Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. K., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C., Kuroda, M. I., Pirrotta, V., Karpen, G. H., and Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485.

Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**, 134–142.

Li, X. -Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C. L. L., Chu, H. C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S. E., Knowles, D. W., Gingeras, T., Speed, T. P., Eisen, M. B., and Biggin, M. D. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol.* **6**, e27.

Li, X. Y., Thomas, S., Sabo, P. J., Eisen, M. B., Stamatoyannopoulos, J. A., and Biggin, M. D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol.* **12**, R34.

Liu, X., Lee, C. K., Granek, J. A., Clarke, N. D., and Lieb, J. D. (2006). Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528.

MacArthur, S., Li, X. -Y., Li, J., Brown, J. B., Chu, H. C., Zeng, L., Grondona, B. P., Hechmer, A., Simirenko, L., Keränen, S. V. E., Knowles, D. W., Stapleton, M., Bickel, P., Biggin, M. D., and Eisen, M. B. (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **10**, R80.

Mann, R. S., Lelli, K. M., and Joshi, R. (2009). Hox specificity unique roles for cofactors and collaborators. *Curr. Top. Dev. Biol.* **88**, 63–101.

Miller, J. A., and Widom, J. (2003). Collaborative competition mechanism for gene activation in vivo. *Mol. Cell Biol.* **23**, 1623–1632.

Narlikar, G. J., Fan, H. -Y., and Kingston, R. E. (2002). Cooperation between complexes that regulate chromatin structure and transcription. *Cell.* **108**, 475–487.

Narlikar, L., Gordan, R., and Hartemink, A. J. (2007). A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.* **3**, e215.

Narlikar, L., and Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. *Brief Funct. Genomic. Proteomic.* **8**, 215–230.

Noyes, M. B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M. H., and Wolfe, S. A. (2008). A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.* **36**, 2547–2560.

Rabiner, L. (1989). A Tutorial on hidden Markov models and selected applications in speech recognition. *P IEEE.* **77**, 257–286.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformat.* **3**, 30.

Ramsey, S. A., Knijnenburg, T. A., Kennedy, K. A., Zak, D. E., Gilchrist, M., Gold, E. S., Johnson, C. D., Lampano, A. E., Litvak, V., Navarro, G., Stolyar, T., Aderem, A., and Shmulevich, I. (2010). Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics* **26**, 2071–2075.

Raveh-Sadka, T., Levo, M., and Segal, E. (2009). Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res.* **19**, 1480–1496.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Meth.*. **4**, 651–657.

Roider, H. G., Kanhere, A., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* **23**, 134–141.

Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. *Plos Biol.* **2**, E271.

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.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**, 535–540.

Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics* **22**, e454–e463.

Sinha, S., van Nimwegen, E., and Siggia, E. D. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics* **19**(Suppl 1), i292–i301.

Small, S., Blair, A., and Levine, M. (1992). Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J.* **11**, 4047–4057.

Stanojevic, D., Small, S., and Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science (New York, NY)* **254**, 1385–1387.

Steihaug, T. (1983). The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Numerical Analysis* **20**, 626–637.

Thomas, S., Li, X. Y., Sabo, P. J., Sandstrom, R. B., Thurman, R. E., Canfield, T. D., Giste, E., Fisher, W., Hammonds, A., Celniker, S. E., Biggin, M. D., and Stamatoyannopoulos, J. A. (2011). Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.* **12**, R43.

Ward, L. D., and Bussemaker, H. J. (2008). Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* **24**, i165–i171.

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

Whitington, T., Perkins, A. C., and Bailey, T. L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.* **37**, 14–25.

Won, K. -J., Ren, B., and Wang, W. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.* **11**, R7.

Wunderlich, Z., and Mirny, L. A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* **25**, 434–440.

Zeitlinger, J., Simon, I., Harbison, C. T., Hannett, N. M., Volkert, T. L., Fink, G. R., and Young, R. A. (2003). Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395–404.

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev.* **21**, 385–390.

Zinzen, R. P., Senger, K., Levine, M., and Papatsenko, D. (2006). Computational models for neurogenic gene expression in the Drosophila embryo. *Curr. Biol.* **16**, 1358–1365.