Nucleosome Free Regions in Yeast Promoters Result from Competitive Binding of Transcription Factors That Interact with Chromatin Modifiers

Evgeniy A. Ozonov, Erik van Nimwegen*

Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland

Abstract

Because DNA packaging in nucleosomes modulates its accessibility to transcription factors (TFs), unraveling the causal determinants of nucleosome positioning is of great importance to understanding gene regulation. Although there is evidence that intrinsic sequence specificity contributes to nucleosome positioning, the extent to which other factors contribute to nucleosome positioning is currently highly debated. Here we obtained both in vivo and in vitro reference maps of positions that are either consistently covered or free of nucleosomes across multiple experimental data-sets in Saccharomyces cerevisiae. We then systematically quantified the contribution of TF binding to nucleosome positiong using a rigorous statistical mechanics model in which TFs compete with nucleosomes for binding DNA. Our results reconcile previous seemingly conflicting results on the determinants of nucleosome positioning and provide a quantitative explanation for the difference between in vivo and in vitro positioning. On a genome-wide scale, nucleosome positioning is dominated by the phasing of nucleosome arrays over gene bodies, and their positioning is mainly determined by the intrinsic sequence preferences of nucleosomes. In contrast, larger nucleosome free regions in promoters, which likely have a much more significant impact on gene expression, are determined mainly by TF binding. Interestingly, of the 158 yeast TFs included in our modeling, we find that only 10–20 significantly contribute to inducing nucleosome-free regions, and these TFs are highly enriched for having direct interations with chromatin remodelers. Together our results imply that nucleosome free regions in yeast promoters results from the binding of a specific class of TFs that recruit chromatin remodelers.

Citation: Ozonov EA, 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(8): e1003181. doi:10.1371/journal.pcbi.1003181

Editor: Ilya Ioshikhes, Ottawa University, Canada

Received December 21, 2012; **Accepted** July 4, 2013; **Published** August 22, 2013

Copyright: © 2013 Ozonov, van Nimwegen. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Part of the the work presented here was supported by SystemsX.ch through the CellPlasticity project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: erik.vannimwegen@unibas.ch

Introduction

The genomes of all eukaryotic organisms are packaged into nucleosomes, which are the fundamental units of chromatin, each composed of approximately 147 base pairs (bp) of DNA wrapped around a histone octamer. Recent developments in technologies for measuring chromatin marks by chromatin immunoprecipitation (ChIP) on microarrays (ChIP-Chip) or by sequencing (ChIP-seq) have enabled the construction of genome-wide maps of nucleosome positions and modifications at high resolution across various conditions. These experimental data have revealed that nucleosomes are not uniformly distributed across the genome but rather that transcription start and termination sites are relatively depleted of nucleosomes [1,2]. Furthermore, nucleosome positioning has been shown to vary across physiological conditions [3].

It has long been accepted that nucleosomes have intrinsic sequence preferences which influence nucleosome positioning, e.g. [4–6]. At the same time, it has also long been known that barriers in the DNA can cause nucleosomes to be ‘statistically positioned’ relative to such barriers, introducing a periodic pattern of nucleosome occupancy on both sides of the barrier [7]. Given the fact that nucleosomes may cover more than 80% of the genome [1], it is therefore also conceivable that a relatively small number of barriers on the DNA, in combination with statistical positioning relative to these barriers, determines most of the observed nucleosome positioning. For example, recent work suggests that nucleosome occupancy patterns around TSSs could at least partly be explained by such statistical positioning [8].

Probably the most obvious class of candidate molecules that could introduce condition-specific barriers on the DNA are sequence-specific transcription factors (TFs). Indeed, for some specific promoters in *S. cerevisiae* it has been established that binding of TFs is a major determinant of nucleosome positioning in the promoter region, e.g. [9–11]. Moreover, the resulting nucleosome positioning has major effects on gene regulation from these promoters. In addition, for a few TFs it has been established that their binding induces local nucleosome exclusion genome-wide [1,12–14].

Although it is thus clear that both intrinsic sequence preferences of nucleosomes and competitive binding of other DNA binding factors play a role in nucleosome positioning, the relative importance of these factors have come under intense debate in recent years. For example, it has been proposed that the positioning of nucleosomes, in particular in *S. cerevisiae*, is mainly determined by intrinsic sequence preference of the nucleosomes, i.e. [15]. In this view, nucleosomes are mainly positioned by a

Author Summary

The DNA of all eukaryotic organisms is packaged into nucleosomes, which cover roughly 80% of the genome. As nucleosome positioning profoundly affects DNA accessibility to other DNA binding proteins such as transcription factors (TFs), it plays an important role in transcription regulation. However, to what extent nucleosome positioning is guided by intrinsic DNA sequence preferences of nucleosomes, and to what extent other DNA binding factors play a role, is currently highly debated. Here we use a rigorous biophysical model to systematically study the relative contributions of intrinsic sequence preferences and competitive binding of TFs to nucleosome positioning in yeast. We find that, on the one hand, the phasing of the many small spacers within dense nucleosome arrays that cover gene bodies are mainly determined by intrinsic sequence preferences. On the other hand, larger nucleosome free regions (NFRs) in promoters are explained predominantly by TF binding. Strikingly, we find that only 10–20 TFs make a significant contribution to explaining NFRs, and these TFs are highly enriched for directly interacting with chromatin modifiers. Thus, the picture that emerges is that binding by a specific class of TFs recruits chromatin modifiers which mediate local nucleosome expulsion.

'code' in the DNA sequence and the accessibility of the DNA to TFs is downstream of this sequence-guided nucleosome positioning. However, these conclusions were challenged by several studies which suggested nucleosome sequence specificity can only explain a modest fraction of nucleosome positioning, and that statistical positioning likely also plays an important role [1,2,16,17]. More recently, several groups have undertaken further experimental investigations into this question, in particular by experimentally comparing nucleosome positioning *in vivo* and *in vitro* [18,19]. Although there is general agreement that these experimental studies confirmed that both intrinsic sequence preferences and the competitive binding of TFs play a role in nucleosome positioning, different authors came to strikingly different, and often seemingly contradictory conclusions regarding which of these factors play a dominant role [20–24]. It is thus clear that, rather than lacking sufficient experimental data, the current challenge in furthering our understanding of the determinants of nucleosome positioning lies in the quantitative interpretation of this data.

Here we show that, by analyzing existing experimental data in combination with rigorous computational modeling, important novel insights can be gained that reconcile previous seemingly contradictory observations, and that suggest a new picture of the mechanisms regulating nucleosome positions. In particular, we use a biophysical model to quantitatively assess the role of TFs in determining nucleosome positioning in *S. cerevisiae*, to assess which aspects of nucleosome positioning TFs contribute to most, and to identify whether there are subsets of TFs that play a predominant roles in this process. *S. cerevisiae* is a particularly attractive system for such an analysis because extensive nucleosome positioning data are available, and because it is essentially the only organism in which sequence-specificities are available for the very large majority of TFs.

Rather than assuming that intrinsic sequence preferences determine nucleosome positioning and that TF binding occurs preferentially at those regions not covered by nucleosomes, or vice versa, assuming that TF binding sets boundaries in the DNA against which nucleosomes are statistically positioned, in our model the TF binding and nucleosome positioning patterns are determined by a dynamic competition of all TFs and nucleosomes for binding to the DNA. Our model incorporates both the sequence preferences of the nucleosomes and of all TFs in a thermodynamic setting, and rigorously calculates the resulting equilibrium occupancies genome-wide as a function of the concentrations of all TFs and the nucleosomes.

Using this model in combination with experimental data we find that TF binding makes a substantial contribution to nucleosome positioning but only at a specific subset of genomic positions. In particular, the linker regions between nucleosomes can be clearly divided into two classes based on their size: the large majority of linkers is small ($\approx 15\,\mathrm{bp}$) and occurs within large nucleosome arrays in gene bodies, whereas a minority of linkers is large ($>80\,\mathrm{bp}$) and occurs predominantly in promoters. Our results show that the phasing of the small linkers within nucleosome arrays, and thereby the majority of nucleosome positioning genome-wide, is mainly determined by sequence preferences of nucleosomes. In contrast, the larger nucleosome free regions in promoters, which are likely most relevant for effects on gene expression, are mainly determined by competitive binding of TFs. By applying our model to data on nucleosome positioning *in vitro* we also confirm that the ability of TFs to explain nucleosome positioning in promoters is restricted to *in vivo* data. Thus, our model provides a quantitative and mechanistic explanation for the observed discrepancies between *in vivo* and *in vitro* nucleosome positioning. Most strikingly, our results also show that, rather than all TFs contributing roughly equally to the competition with nucleosomes, the effect of TFs on nucleosome positioning is restricted to a relatively small set of about $10-20$ TFs. Although one might expect that these TFs are simply the highest expressed TFs with the largest number of TFBSs genome-wide in the conditions in which the experiments were performed, we find this not to be the case. Instead, we find that these TFs are highly enriched for having known protein-protein interactions with chromatin remodeling complexes, histones, and chromatin modification enzymes. Thus, the mechanistic picture suggested by our results is that there is a specific class of TFs who, upon binding to the DNA, recruit chromatin modifiers that then mediate local expulsion of nucleosomes.

Results

A biophysical model of TF and nucleosome binding to genomic DNA

To rigorously investigate the competition between TFs and nucleosomes for binding to DNA, and the role of TFs in nucleosome positioning, we take a statistical mechanics approach in which we explicitly consider all possible non-overlapping binding configurations to the genome for nucleosomes and a large set of TFs, assigning a probability to each configuration using standard Boltzmann-Gibbs statistics. The basic approach, which uses dynamic programming to efficiently sum over all possible binding configurations, has been used in computational methods for analysis of transcription regulation for over a decade, e.g. [17,25–28], and has been used more recently to specifically investigate the effect of competitive binding of nucleosomes and TFs [29,30]. Here we use this approach to comprehensively investigate the role of TFs in determining nucleosome positioning. We employ an unprecendented complete set of 158 TF binding models, we investigate the dependence on the concentrations of these TFs, and we also introduce tunable sequence-specificities for all TFs and nucleosomes.

The model is explained in detail in the Materials and Methods. Briefly, each TF t is assumed to bind DNA segments of a fixed length l_t and, for any length-l_t DNA segment s, a binding energy $E(s|t)$ is determined. The energies $E(s|t)$ are calculated from a weight matrix representation of the TF's binding sites [31] and involve a tunable scale parameter γ_t which controls the sequence-specificity of the TF. To obtain energy matrices for the large majority of sequence-specific TFs in *S. cerevisiae* we used a collection of 158 WMs that we curated previously [32] and that are based on a combination of ChIP-chip and *in vitro* binding data. Notably, while the WMs allow us to determine how the binding energy (measured in units k_BT) varies across positions in the genome for each TF, the WMs do not allow us to determine the sequence-independent contribution to binding energy, i.e. the overall 'stickines' of each TF for DNA. To compare binding energies across TFs we set the sequence-independent contribution to the binding energy such that all TFs have equal overall affinity for the DNA (see Materials and Methods).

Of the computational work done on nucleosome positioning, probably most effort has been invested in developing models for nucleosome sequence-specificity based on data from both *in vivo* and *in vitro* nucleosome binding, e.g. [15,18]. Exploiting analytical results from statistical mechanics, Locke et al. [24] rigorously inferred the energies of nucleosome binding from high-throughput data and used these to evaluate several models of different complexity for the sequence specificities of nucleosomes. The results from this study suggested that the sequence specificity of nucleosomes can be captured by fairly simple models. As we discuss below, our own analysis suggests that the performance of different models of nucleosome sequence specificity depends on the precise data-set and performance evaluation method used, but that all models make highly correlated predictions (Figure 1A). Of the models analyzed, the model of [18] gave robustly high performance across data-sets and we use this model in our study. In particular, we assume that nucleosomes bind to DNA segments of 147 nucleotides and determine an energy of binding $E(s|\text{nucl})$ for any length 147 segment s using a generalization of the model of [18], involving a scale parameter γ_{nucl} that controls the sequence specificity of the nucleosomes, analogous to the scale parameters γ_t for the TFs (see Materials and Methods). The parameter γ_{nucl} allows us to investigate the effect of enhancing or decreasing the nucleosome sequence specificity. For example, when setting $\gamma_{\text{nucl}} = 0.4$, the variation in nucleosome binding energies across different sequences is reduced to 40% of the energy variations predicted by the model of [18].

As mentioned above, the model assumes that any DNA segment can only be bound by a single TF or a nucleosome at a time. Although it is likely that there are exceptions to this simplification, it is generally accepted that TFs and nucleosomes compete for binding to DNA. In absence of specific information as to which TFs compete with nucleosomes and which can co-bind with nucleosomes, we make the simplifying assumption that all TFs compete with nucleosomes, as has been done previously by others [29,30]. Like previous approaches, e.g. [8,15,22,29], our model also assumes that the average occupancy profiles across a population of cells are well approximated by their thermodynamic equilibrium averages. Notably, given that there are many ATP-driven processes that cause nucleosome turnover and displacement by chromatin remodelers, it is not a priori clear that this equilibrium assumption holds. Ours and previous computational approaches thus essentially assume that these ATP-driven processes act mainly to affect kinetics, i.e. to allow nucleosomes

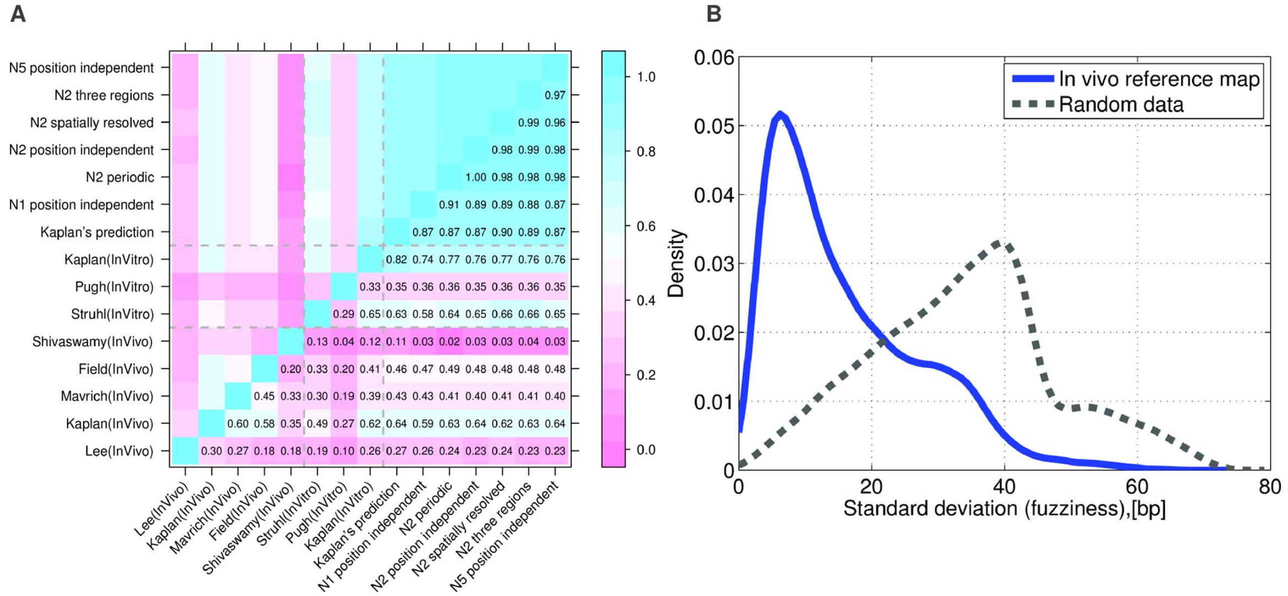

Figure 1. Reproducibility of *in vitro* and *in vivo* nucleosome data across different experiments and performance of nucleosome sequence-specificity models. A: Pearson correlation coefficients of the per-base nucleosome coverage between various experimental data-sets measuring nucleosome occupancy either *in vivo* [1,3,18,38,56] or *in vitro* [18,19,58], and predictions from a number of models of nucleosome sequence-specificity [18,24]. **B:** Reproducibility of annotated nucleosome positions across the *in vivo* data-sets. For each annotated nucleosome in the reference map of [41], we calculated the standard deviation in the annotated positions of the corresponding nucleosomes across the 6 data-sets used to construct the map. The blue curve shows the distribution of standard deviations across nucleosomes. The grey dotted curve shows the analogous distribution that is obtained using randomized data (see Materials and Methods). The high reproducibility of nucleosome positions across different data-sets justifies the use of binary data, i.e. positions of ''linkers'' and ''nucleosomes'', instead of Pearson correlation for evaluation of the performance of computational models for predicting nucleosome positions.
doi:10.1371/journal.pcbi.1003181.g001

to resample their positions, without systematically biasing their positioning. Some recent evidence appears to support this assumption [33].

The model considers all possible non-overlapping configurations C of TFs and nucleosomes bound along the genome. For each configuration C, a total energy $E(C|\mathrm{c},\gamma)$ is calculated. This energy depends on the concentrations of nucleosomes c_{nucl} and all TFs c_t, which we collectively denote as c, and also on all energy scale factors γ that determine sequence-specificity (Materials and Methods). The probability $P(C|\mathrm{c},\gamma)$ to find a cell in configuration C is then given by the standard Boltzmann-Gibbs formalism as

$$P(C|\mathrm{c},\gamma) = \frac{e^{-\beta E(C|\mathrm{c},\gamma)}}{Z}, \tag{1}$$

where $\beta = 1/(kT)$ is the inverse temperature, Z is the partition sum, and we have explicitly indicated that these probabilities depend on the concentrations c and scale factors γ. As explained in Materials and Methods, both the partition sum and the fractions of the time each TF t is bound at each genomic position can be calculated efficiently using standard dynamic programming techniques.

In summary, given a set of input concentrations c for all TFs and nucleosomes, the model efficiently calculates the equilibrium binding frequencies of all TFs and nucleosomes across the entire genome. Note that, because all TFs and nucleosomes are in competition for binding to the DNA, the occupancy of any factor to a sequence segment of the genome in principle depends, not only on the concentration of this factor and its affinity to the sequence segment, but on the concentrations of all other factors and their affinities to all other locations in the genome. Thus, the TF and nucleosome occupancy profiles across the genome can be changed by varying the concentrations c and scale factors γ. In particular, these parameters can be optimized to maximize the agreement with experimentally determined nucleosome occupancy profiles.

Comparing model predictions with experimental nucleosome position profiles

Many experimental studies have been carried out to map nucleosome positions in eukaryotic species, e.g. [34–37], and in *Saccharomyces cerevisiae* in particular, e.g. [1–3,18,19,38,39], so that several data-sets of nucleosome positions in *S. cerevisiae* are available. In order to determine how to meaningfully compare computational predictions with these experimental data, we first performed a comparative analysis of several experimental data sets. Patterns of nucleosome positioning that are typically highlighted in publications, such as the nucleosome-depleted regions upstream of the transcription start sites (TSSs) and well-positioned nucleosomes immediately downstream of TSS, involve genome-wide averages of nucleosome occupancy across a class of positions. Such average patterns are robust to fluctuations and are shared by all data-sets.

Previous works have assessed the performance of models of nucleosome sequence specificity by determining both the predicted and experimentally observed nucleosome occupancies across individual regions of the genome, and by calculating the Pearson correlation of these nucleosome occupancy profiles. To assess the validity of such an approach, we calculated Pearson correlations between observed occupancy profiles of several experimental data-sets (both *in vivo* and *in vitro*) as well as several models of nucleosome sequence specificity (Figure 1A). This shows that, unfortunately, the occupancy profiles correlate only weakly across different experimental data-sets, with Pearson correlation coefficients typically ranging from $r = 0.2$ to $r = 0.45$ for *in vivo* data-sets, and only marginally higher for *in vitro* data-sets. This large variability across data-sets may to some extent be due to biases of the technological platforms. For example, it is well known that the nucleotide composition and propensity to form secondary structures of the reads can systematically bias the read counts in ChIP-seq by more than 10-fold [20,40]. Variations in details of the ChIP protocol are likely also responsible for some of the variation across data-sets, and previous studies have indicated that MNase digestion bias may also systematically affect nucleosome positioning data [23,24]. Since all experiments were performed in YPD, true biological variation is likely only a minor source of variation in these data.

In contrast to the experimental data, the occupancy profiles predicted by the different computational models are all highly correlated. Moreover, the correlations across models for a given data-set vary much less than the correlations for a given method vary across data-sets. For example, all models consistently perform better on *in vitro* than on *in vivo* data. Among the *in vivo* data-sets, all methods perform by far best on the *in vivo* data of Kaplan et al.[18] (which is also far more correlated with *in vitro* data than any other *in vivo* data-set) and far worst on the *in vivo* data of Shivaswamy et al. [3]. Thus, comparison of different models with existing data supports the conclusions of [24] that different models of nucleosome-specificity perform similarly in explaining nucleosome positioning. Since the model of Kaplan et al. [18] exhibits highest performance for the majority of *in vivo* and *in vitro* data-sets, we chose to use this model in our analysis. However, the weak correlation of nucleosome occupancy profiles across data-sets shows that assessing the performance of computational predictions by directly comparing predicted and observed nucleosome occupancies is highly problematic. A meaningful comparison of computational models requires that one first extracts those features of the nucleosome positioning that are reproducible across experimental data-sets.

In contrast to the absolute value of the ChIP signal, we observed that the positions of local maxima and minima in nucleosome occupancy are much better reproduced across data-sets. This reproducibility of the 'peaks and troughs' in the nucleosome occupancy profile has been observed previously [41], and has been used to create a reference set of 'nucleosome' and 'linker' segments. In this procedure, local maxima and minima are used to annotate nucleosomes and linkers in each data-set. These annotations are then intersected, with reference nucleosomes placed at the consensus positions of regions annotated as nucleosomes in all data-sets, and reference linkers the regions free of nucleosomes in all annotations. That the positions of annotated nucleosomes are highly reproducible across data-sets, especially compared to raw coverage and compared to nucleosome maps based on randomized data, is illustrated in Figure 1B. The annotated positions of individual nucleosomes across different data-sets typically vary by less than 10 base pairs from the reference position (blue curve in Figure 1B) and the vast majority of annotated nucleosome positions vary by less than 20 bp from the reference position. In contrast, on randomized data positions of annotated nucleosomes typically vary by roughly 40 bp from the reference position (dotted curve in Figure 1B).

In summary, although ideally we would like to test whether computational models can predict relative nucleosome occupancies across the genome, it is not possible to meaningfully perform such an assessment given the variability observed in the experimental data. We thus evaluate the performance of different models by assessing their ability to predict nucleosome and linkers that occur consistently across different data-sets. We use the reference set annotated by [41] consisting of roughly $60'000$ annotated linker regions and $21'000$ annotated nucleosomes, that together cover about 50% of the genome, to assess the performance of the model in predicting *in vivo* nucleosome positioning. In addition, we have applied a similar annotation procedure (Materials and Methods) to produce a reference set of nucleosomes and linkers from 3 *in vitro* data-sets, which we use to assess the performance of the model in predicting nucleosome positioning *in vitro*.

To assess the model's performance we compare the predicted nucleosome coverage at annotated linker and nucleosome segments. That is, instead of comparing the predicted and observed absolute occupancies, we assess the model's ability to predict local maxima and minima in nucleosome occupancy, that occur consistently across data-sets. As described in Materials and Methods, based on the predicted nucleosome coverage, we classify each segment as either nucleosome or linker, and then calculate the *mutual information* I between the predicted and experimentally measured classification. Finally, we normalize this mutual information by the entropy H of the experimental classification to obtain the fraction $F=I/H$ of information that is captured by the model's predictions, i.e. F runs from 0 (random predictions) to 1 (perfect predictions). An F value of 0.2 means that the model captures 20% of all the information needed to specificy which of the genomic segments correspond to nucleosomes and which to linkers. We will refer F as the 'quality score'. As mutual information is the fundamental measure of dependence between two distributions [42,43], we consider the quality score F the most rigorous quantification of model performance. However, as we show below, highly similar results are obtained with other performance measures that are popular in machine learning, such as area under the ROC curve (AUC).

Optimal fits to nucleosome positioning require weak nucleosome sequence specificity

We first tested what quality score can be obtained by the intrinsic sequence specificity of the nucleosomes, i.e. leaving all TFs out of the model, and how the quality of the fit depends on the sequence specificity of the nucleosomes. Figure 2A shows the quality scores F that are obtained for different scale factors γ_{nucl} on nucleosome sequence specificity (with 0 representing no sequence preference whatsoever and 1 representing the specificity used in Kaplan et al. [18]). The optimal fit is obtained for $\gamma_{\text{nucl}} \approx 0.47$, which corresponds to significantly lower nucleosome sequence specificity than those used in Kaplan et al. [18]. That is, for the model of [18], the standard deviation of nucleosome binding energies is approximately $1.64 k_B T$ across the genome (0.97kcal/mole), whereas we observe optimal fits for roughly 2-fold lower variations in binding energies (roughly $0.77 k_B T$). Moreover, the quality score depends weakly on γ_{nucl} and becomes small only for extremely small sequence specificities.

These results may seem contradictory, given that the sequence-specificity model of Kaplan et al. was developed specifically with the aim of explaining nucleosome positioning. However, Kaplan et al. optimized the overall Pearson correlation between predicted and observed nucleosome coverage, which depends strongly on the variation in absolute nucleosome occupancies. In contrast, the quality score F depends mainly on the locations of local maxima and minima in the occupancy, and much less on the absolute amount of variation in nucleosome occupancy. To investigate this further, we compared the distribution of nucleosome occupancies for the model with different values of γ_{nucl} with the distribution of nucleosome occupancies for the model of Kaplan et al. and the experimentally observed distribution of nucleosome occupancies for the data of Lee et al. [1] (Materials and Methods, and note that very similar distributions are obtained from other experimental data-sets; Figure S1 in Text S1).

As shown in Figure 2B, the model of Kaplan et al. [18] predicts an overall nucleosome coverage that is dramatically lower than our fits, i.e. with a median nucleosome coverage of about 0.3. Such a coverage distribution is strongly at odds with the experimental data which shows that, rather than 30%, about 80% of the genome is covered by nucleosomes, e.g. [1,3,44,45]. It is likely that the unrealistically low nucleosome occupancy of Kaplan et al. [18] is

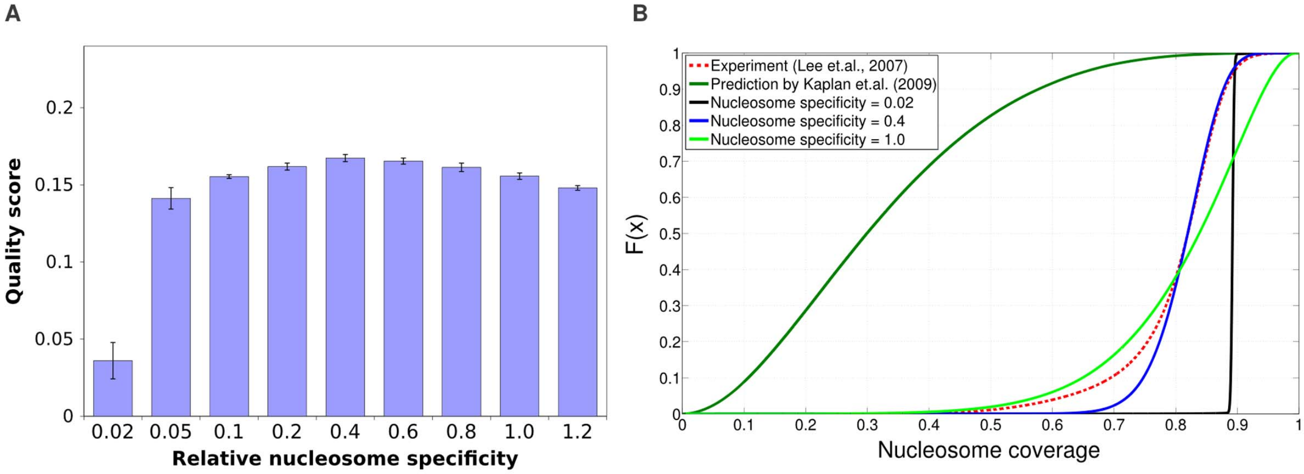

Figure 2. Performance of models that include only nucleosome sequence specificity. A: Fraction of information regarding experimentally annotated linker and nucleosome positions explained by the nucleosome-only model (quality score, vertical bars) as a function of relative nucleosome specificity. The relative nucleosome specificity is controlled by the scale factor γ_{nucl}, where $\gamma_{\text{nucl}}=1.0$ corresponds to the sequence specificity of the model of Kaplan et al. [18], for which the binding energy of the nucleosomes has a standard-deviation of $1.64 k_B T = 0.97\text{kcal/mole}$ across the genome. The error-bars indicate standard-errors across 5 separate test sets. **B:** Experimentally observed cumulative distribution of nucleosome coverages (fraction of time a given genomic position is covered by a nucleosome) from [1] (red dotted line) and cumulative distributions of predicted nucleosome coverage of the models of [18] (dark green line) and our model using nucleosome specificity scale parameters of $\gamma_{\text{nucl}}=0.02$ (black line), $\gamma_{\text{nucl}}=0.4$ (blue line), and $\gamma_{\text{nucl}}=1.0$ (light green line).
doi:10.1371/journal.pcbi.1003181.g002

an artefact of optimizing the Pearson correlation in nucleosome coverage, since this objective function favors high variance in predicted nucleosome coverage, and does not penalize the mismatch in the average nucleosome coverage.

For our model, the coverage distribution indeed strongly depends on the nucleosome specificity. Strikingly, by far the best fit between the observed and predicted coverage distribution occurs precisely at the specificity that maximizes our quality score (i.e. at $\gamma_{\mathrm{nucl}}=0.4$). This demonstrates that, in contrast to the predictions of Kaplan et al. [18], our fits produce realistic nucleosome coverage profiles, in spite of not specifically optimizing these coverage profiles. In fact, at the optimal nucleosome specificity, the predicted and experimentally observed nucleosome coverage distribution is virtually identical for the 70% of base pairs in the genome with highest nucleosome coverage (blue and red curves in Figure 2B). The main deviation between model and experimental data is that the model fails to predict regions with low nucleosome coverage that are observed experimentally. Indeed, as we will see below, whereas the model correctly predicts almost all nucleosomes, the model fails to correctly predict a substantial fraction of linker regions as nucleosome free.

In summary, optimizing the quality score F produces much more realistic fits to the nucleosome coverage distribution than previous models, and shows that the best fits are obtained with only weak nucleosome sequence-specificity.

Transcription factor binding plays a major role in explaining nucleosome free regions at promoters

We next investigated to what extent competition with TFs improves the predicted nucleosome positioning. We first considered models in which, besides the nucleosomes, there is only a single TF. For each of these models we fitted the 4 parameters (i.e. the concentrations and sequence specificity of both nucleosomes and the TF) using simulated annealing, and calculated the quality score F obtained with this model using 80/20 cross-validation (Materials and Methods). We ranked TFs by the z-statistic they obtained in cross-validation (Materials and Methods), and then investigated what quality scores F can be obtained using the top 5, 10, 20 and top 30 TFs, refitting all concentrations and sequence specificity parameters. We find that adding the TFs clearly increases the quality of the predictions on the test-sets, although the improvement is relatively small, i.e. from $F\approx 0.17$ to $F\approx 0.2$, Figure 3A. Given this modest increase in F and the large number of parameters involved when including many TFs in parallel, one may wonder whether these results are affected by overfitting. However, as shown in Figure S2 in Text S1, the observed F scores on train and test sets are essentially identical. In addition, adding the TFs to the model further improves the match between the observed and predicted nucleosome occupancy distribution (Figure S1 in Text S1).

As already observed in [41], the length distribution of linkers is bimodal. The large majority of linkers is short, around on average 15 bps in length, corresponding to short linkers within arrays of nucleosomes. There is a second class, corresponding to roughly 25% of all annotated linkers, that are much longer, i.e. each more than 80 bps long. We will refer to these longer linkers as 'nucleosome free regions' (NFRs). We next asked whether TFs contribute more to explaining the positioning of the short linkers or the longer NFRs. Moreover, as TFs are expected to bind predominantly to promoter regions, we also investigated whether the contribution of the TFs to explaining nucleosome positioning is most significant in promoters (defined as running from 500 bp upstream to 500 bp downstream of TSS). We find that, generally, inclusion of the TFs leads to a substantially larger increase in performance for promoter regions, and TFs contribute much more to explaining NFRs than explaining small linkers (Figure S3 in Text S1). In particular, considering NFRs and nucleosomes in promoter regions, inclusion of TFs almost doubles the quality score F, i.e. from 0.23 to 0.38, Figure 3A, red bars. As an aside, we note that these observations do not depend on assessing the model's performance by the quality score F. As shown in Figure S4 in Text S1, we find essentially the same results when assessing the model's performance using ROC curves, and the area under the curve (AUC) is almost perfectly correlated ($r=0.99$) with the quality score F. It is also noteworthy that, both when predicting all linkers genome-wide or NFRs in promoters, even though up to

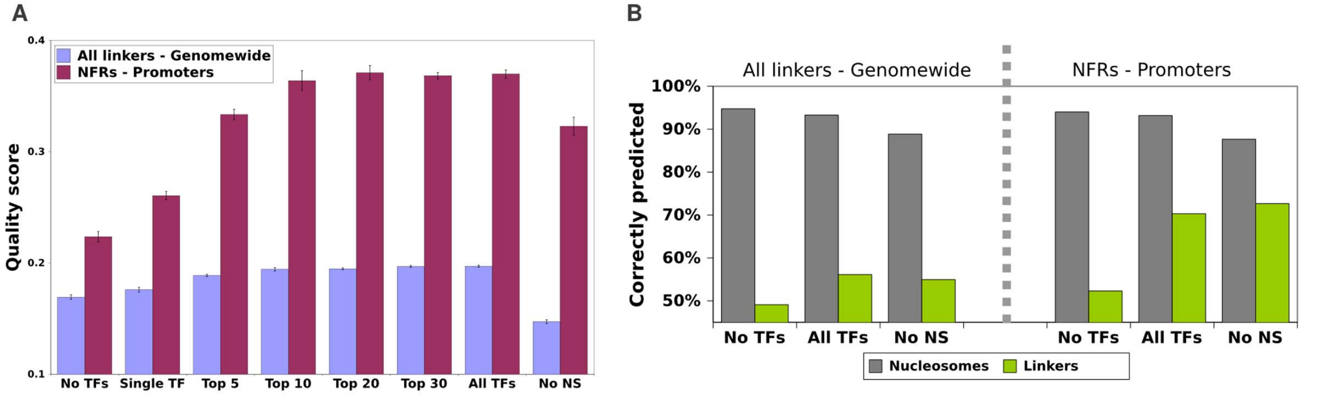

Figure 3. Incorporating competition with TFs improves predicted nucleosome positioning, particularly in promoter regions. A: Ability to predict nucleosome positioning as a function of the number of TFs used in the model. The bars show the fraction of all information regarding nucleosome positioning explained (quality score F) by each model. Results are shown for, from left to right, the model including only nucleosomes (no TFs), only the best TF, the top 5 TFs, top 10 TFs, etcetera. The rightmost pair of bars correspond to a model including all TFs but without any sequence specificity for the nucleosomes $\gamma_{\mathrm{nucl}}=0$. Blue bars correspond to quality scores for predicting all nucleosomes and linkers genome-wide and red bars correspond to quality scores for predicting nucleosomes and nucleosome free regions (long linkers) within promoters. The error bars show standard-error across 5 independent test-sets. **B**: Fractions of correctly predicted nucleosomes (grey bars) and linkers (green bars) for, from left to right, the model with nucleosome sequence specificity and no TFs, the model with all TFs, and the model with all TFs but no nucleosome sequence specificity. The left half of the figure shows results for predicting all linkers and nucleosome genome-wide, and the right half for predicting NFRs and nucleosomes in promoters.

doi:10.1371/journal.pcbi.1003181.g003

158 TFs can be incorporated, the model essentially reaches its optimal performance after adding the first $10-20$ TFs. We investigate this in more detail below.

It thus appears that TFs contribute not so much to explaining positioned nucleosomes, but rather explain the location of longer NFRs, especially in promoters. Further supporting this observation, the rightmost pair of bars in Figure 3A shows the performance of the model including all TFs but with nucleosome sequence specificity removed, i.e. $\gamma_{\mathrm{nucl}}=0$. We see that removing nucleosome sequence specificity only modestly affects the ability of the model to predict NFRs in promoters. In contrast, the performance on predicting all linkers genome-wide drops significantly when nucleosome sequence specificity is removed, even falling clearly below the performance of the model without TFs. This is further confirmed by closer examination of the errors that the fitted models make (Figure 3B).

For all models, the large majority of nucleosomes is correctly predicted and the fraction of correctly predicted nucleosomes is most strongly affected by removing the sequence specificity of the nucleosomes, i.e. from 95% correct for the model with only nucleosome sequence specificity to 88% for the model with all TFs and no nucleosome specificity. The fraction of correctly predicted linkers is much smaller, e.g slightly below 50% for the model without TFs. Adding the TFs to the model consistently increases the fraction of correctly predicted linkers, and this increase does not require nucleosome sequence specificity. When considering all linkers genome-wide, the increase in correctly predicted linkers is relatively modest, i.e. from 50% to 56%. However, for NFRs in promoters the fraction of correctly predicted NFRs increases from 50% to around 70%. In summary, correctly predicting the phasing of nucleosome arrays over gene bodies crucially depends on nucleosome sequence specificity and is only weakly affected by including TFs, whereas correctly predicting NFRs is strongly dependent on inclusion of the TFs and is almost independent of nucleosome sequence specificity.

Characterization and additional validation of the fitted model

To characterize the biophysical properties of the fitted model we first determined the overall statistics of nucleosome and TF occupancies (Figure 4A). Nucleosomes cover more than 80% of the genome, and most of the remaining regions of the genome are uncovered, with all TFs combined covering less than 1% of the genome. The top 10 TFs with the highest genomic coverage occupy between 0.15% and 0.02% of the genome, corresponding to roughly 1500 and 200 binding sites genome-wide.

For the nucleosomes and the top 10 TFs with highest genomic coverage in the fitted model we also determined the mean and standard-deviation of the binding energies at their binding sites, and the entropy of the distribution of binding probabilities per site (Materials and Methods). The latter quantity is low whenever the TF's coverage results from strong sites with high frequencies of binding, and is high when the TF's coverage comes from a large set of weak sites with lower binding frequencies. The results (Figure 4) show, first of all, that the binding sites of nucleosomes have both the lowest binding energy and the lowest variation in binding energies, i.e. they are the least sequence specific. Interestingly, the top 10 TFs clearly fall into 2 classes: a set of TFs (ABF1, REB1, ORC1, and RSC30) that are highly sequence specific and have strong binding sites, and a class of much less sequence specific TFs (PHO2, NHP6A, etcetera) that bind at a much larger number of weaker sites.

As has been observed previously, e.g. [1,2], averaged nucleosome coverage profiles show a characteristic pattern relative to the starts of genes with a nucleosome depleted region immediately upstream of TSS, followed by a well-positioned nucleosome immediately downstream of TSS and a periodic pattern of nucleosome coverage downstream into the gene body. Although the nucleosome sequence specificity by itself, i.e. without including TFs, reproduces some of this pattern at the $5'$ end of genes (Figure 5A), the observed nucleosome depleted region and the oscillatory pattern into the gene body is much weaker than observed experimentally. As an additional test of the validity of our model, we checked whether inclusion of the TFs improves this average coverage profile relative to gene starts and ends.

We find that adding TFs to the model significantly improves the match between the theoretically predicted and experimentally observed nucleosome coverage pattern at the $5'$ ends of genes (Figure 5A). It is noteworthy that the nucleosome-depleted region

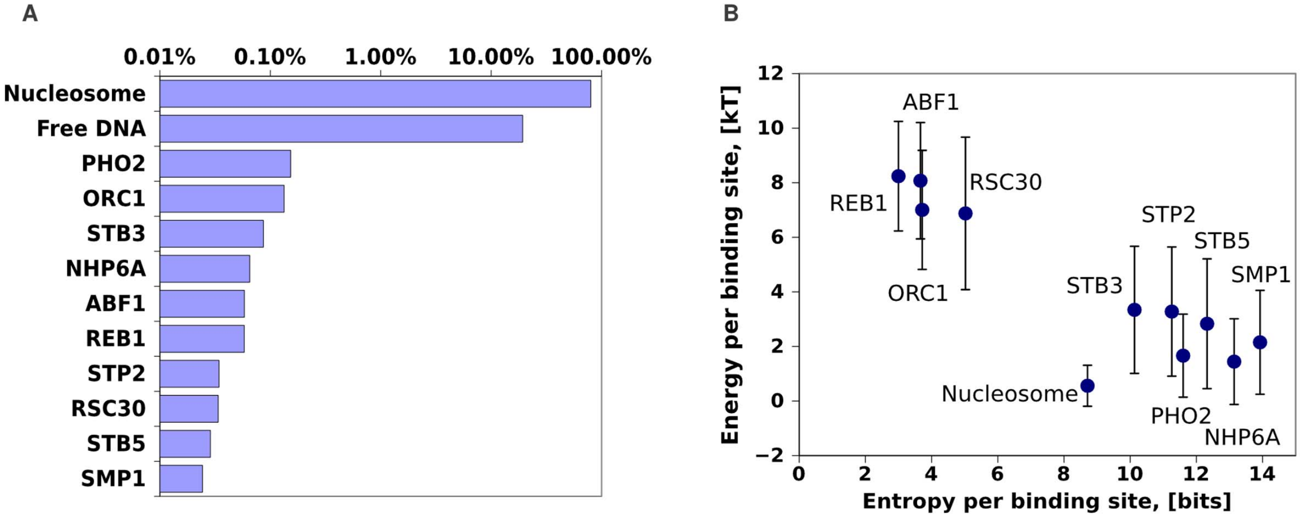

Figure 4. Biophysical properties of the fitted model. A: Average fraction of the genome covered by nucleosomes, free DNA, and the top 10 TFs with highest coverage. **B:** Average and standard-deviation of the binding energies (in units k_BT) at binding sites for nucleosomes and the top 10 TFs with highest coverage (vertical axis), against the average entropy per binding site of the distribution of binding probabilities for the corresponding TFs (horizontal axis).
doi:10.1371/journal.pcbi.1003181.g004

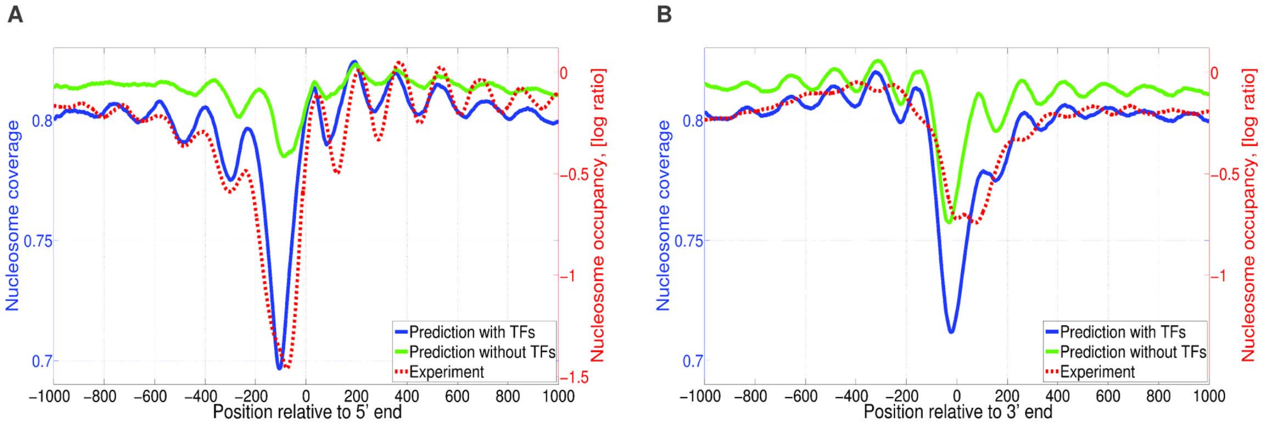

Figure 5. Predicted and observed nucleosome profiles around 5′ and 3′ ends of genes. A: Averaged nucleosome coverage near transcription starts. Each curve shows the average nucleosome coverage at different positions relative to transcription start averaged over all genes. Red dashed lines correspond to experimentally measured nucleosome coverage (data from [1], right vertical axis). The solid lines correspond to the predicted nucleosome coverage by the model including only nucleosomes (light green) and the model including all TFs (blue), left vertical axis. **B**: Averaged nucleosome coverage near transcription ends. Curves are as described for panel A.
doi:10.1371/journal.pcbi.1003181.g005

immediately upstream of TSS coincides with a peak in the overall predicted binding of TFs (Figure S5C in Text S1), further illustrating the role of TFs in establishing nucleosome depletion in these regions. A local peak in TF binding is also predicted immediately downstream of the 3′ ends of genes (Figure S5D in Text S1). Although at the 3′ ends of genes, the inclusion of the TFs also improves the match between the theoretical predictions and the experimentally observed nucleosome coverage, the experimental data and predictions clearly disagree (Figure 5B). First, the width of the experimentally observed NFR is twice as big as the width of the predicted NFR. Second, the oscillations exhibited by the experimentally-determined distribution are not as pronounced as predicted by the model. This lack of a match can likely be attributed to the role of RNA polymerase. Our model considers only 158 TFs and, in particular, does not consider the effects of binding of general transcription factors and RNA polymerase. Experimental data on the positioning of the largest subunit of Pol II - Rpo21, and the general transcription factor Sua7 shows that these factors localize at 3′ ends of genes [46], suggesting that they may contribute to the nucleosome free region observed at the 3′ ends of genes (Figure S6 in Text S1). This is further supported by the analysis in [47], which shows that rapid removal of Polymerase from 3′ end regions increases local nucleosome occupancy.

As another validation of the model, we investigated whether the predicted TF binding matches experimental observations. For example, we compared the intergenic regions predicted to be targeted by the TFs Abf1, Reb1, and Sum1, with the observed target intergenic regions according ot the ChIP-chip data of [48]. This shows that, in spite of the fact that the model was only optimized to fit nucleosome positioning, the fitted model also accurately predicts which regions are targeted by these TFs (Figure S7 in Text S1).

It is important to stress that, although we assess the model's performance by these global statistics, it predicts the precise locations of individual nucleosomes, NFRs, and TF binding sites. The full genome-wide nucleosome and TF coverage predictions obtained with the model including the TFs are made available through our SwissRegulon server www.swissregulon.unibas.ch/ozonov, allowing users to investigate in detail which NFRs at which promoters are explained by the binding of particular TFs. To illustrate the detailed comparison of the model's predictions and observed nucleosome occupancies Figure 6 shows the measured nucleosome coverage, the predictions of the model with and without TFs, and the predicted coverage of TFs, in two genomic regions. As the figure shows, whereas the locations of small peaks and troughs in occupancy across arrays of nucleosomes are reasonably well captured by nucleosome sequence specificity alone, competition with TF binding is needed to explain the occurrence of larger nucleosome free regions, which occur predominantly in promoters. Importantly, it is likely precisely this latter class of regions that are crucial for the effects of nucleosome positioning on gene expression.

However, this detailed comparison also reveals that, whereas the locations of TF binding typically matches the centers of observed NFRs, the predicted shape of these NFRs differs considerably between the model and the experimental observations. In particular, NFRs tend to be much narrower in the model's predictions than in the experimental data. This suggests that, although TF binding determines the genomic location where nucleosome depletion is observed, the observed nucleosome exclusion is more substantial than predicted from the steric hindrance between TFs and nucleosomes. This suggests that TF binding may recruit aditional factors involved in nucleosome exclusion. We return to this observation below.

Only a small subset of TFs, enriched for interacting with chromatin modifiers, crucially affects nucleosome positioning

Our model incorporates the role of TFs through a simple competition for binding DNA and one might thus naively expect that all TFs that are expressed in YPD would contribute similarly to explaining nucleosome positioning, maybe in proportion to the number of their binding sites in the genome. However, we observed above (Figure 3A) that when consecutively adding more TFs to the model, the performance already assymptotes after $10-20$ TFs. This could be due to redundancies in the contributions of the TFs, i.e. if sites for different TFs cluster in particular genomic regions, then binding by only a subset of the TFs will suffice to explain the occurrence of NFRs in these regions, and adding more TFs to the model would not further improve

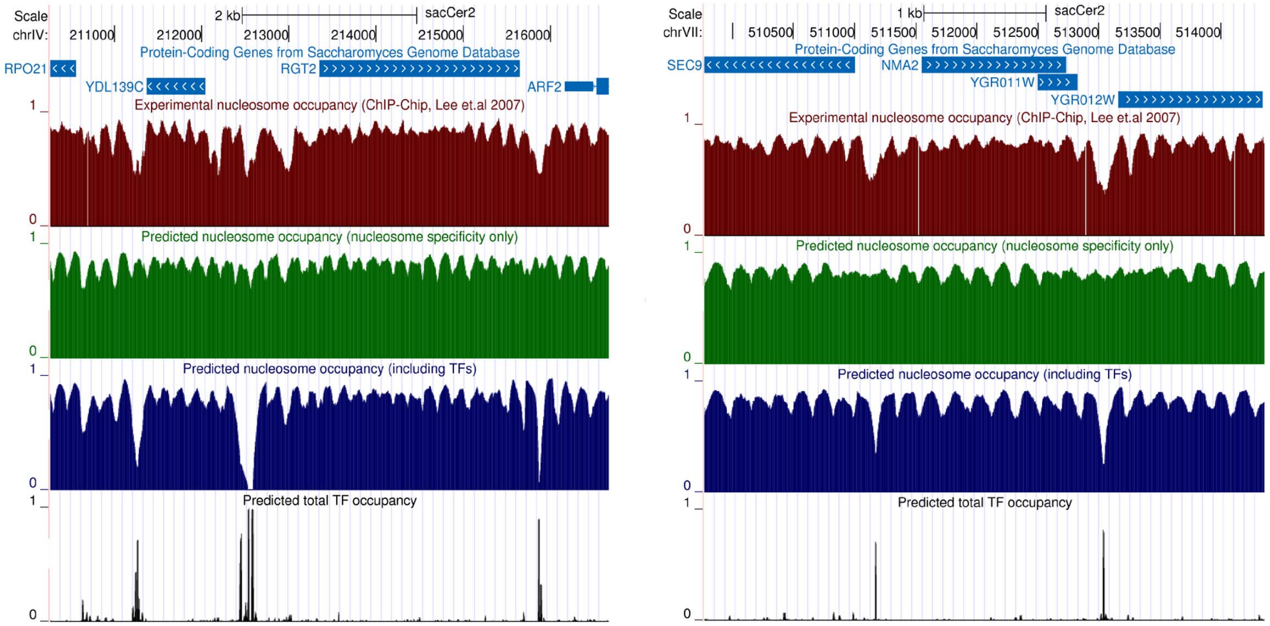

Figure 6. Illustration of the measured nucleosome occupancy and model predictions within individual genomic regions. Each panel shows a section of the yeast genome within our genome browser (swissregulon.unibas.ch/ozonov), with the tracks corresponding to, from top to bottom, chromosomal location, annotated genes, the measured nucleosome coverage based on the data from [1], the predicted nucleosome coverage using the model without TFs, the predicted nucleosome coverage using the model including TFs, and the total predicted TF coverage, i.e. summing over all TFs. Within the genome browser the coverage of individual TFs can also be displayed.
doi:10.1371/journal.pcbi.1003181.g006

performance. Alternatively, it may be that there is a specific class of TFs that contribute much more to nucleosome positioning than other TFs.

To investigate this, we used 80/20 cross-validation on 5 independent training and test sets to assess, for each of the 158 TFs, whether a model containing only nucleosomes and the single TF statistically significantly outperforms the model with only nucleosome specificity, quantifying the significance by a z-statistic (Materials and Methods). Figure 7A shows the distribution of z-statistics obtained for the 158 TFs (blue dots), together with the distribution of z-statistics expected by chance (brown dotted curve). As the figure shows, only $15-20$ of the TFs significantly improve the predictions, indicating that there is indeed a specific class of TFs that dominate in explaining NFRs. Indeed, the large majority of all other TFs obtain quality scores on the test sets that are either the same or worse than the model without any TFs (Figure S8 in Text S1).

As another validation, we checked whether the ability of this subset of TFs to explain nucleosome positioning is a specific property of the sequence specificities of yeast's TFs. That is, it is in principle conceivable that among *any* set of WMs with similar information content and sequence composition, a few will be able to help explain nucleosome positioning. To test this we constructed a set of synthetic WMs by randomly shuffling the columns of the original WMs, and fitted models with these 158 TFs in exact analogy to our fits with the original WMs. As shown in Figure 7A (green dots), none of the shuffled WMs perform better than expected by chance, confirming that the ability to explain nucleosome positioning is unique to the specific set of $15-20$ yeast WMs that we identified.

As a final test, we also evaluated whether the real WMs can explain the nucleosome positioning that is observed *in vitro* (Materials and Methods). On the one hand, since no TFs are present in the conditions at which the *in vitro* experiments are performed, the TFs should in principle not contribute to nucleosome positioning. On the other hand, as the raw *in vivo* and *in vitro* occupancies are significantly correlated (Figure 1A), one might expect that the TF WMs can still positively contribute to explaining *in vitro* nucleosome positioning. It is thus striking that none of the real yeast WMs performs better than expected by chance in explaining *in vitro* nucleosome positioning (Figure 7A, red dots), i.e. including TFs does not help explaining *in vitro* nucleosome positioning. This shows that the actions of a specific set of $15-20$ TFs are crucial for explaining the differences between *in vivo* and *in vitro* nucleosome occupancies.

Figure 7B lists the top 20 TFs and shows their quality scores on the test sets (results for all TFs are shown in Table S1). The fact that only around 20 TFs contribute significantly to nucleosome positioning raises the question of what distinguishes these TFs from the others and we investigated a number of hypotheses. One might hypothesize that the top TFs are simply those that are highest expressed in YPD, or those which occupy most sites genome-wide. However, expression data indicates that these TFs are not particularly highly expressed in YPD compared to other TFs (Figure S9 in Text S1, data from [49]). Consistent with this, the genome-wide number of binding sites, as observed in genome-wide ChIP-chip experiments (Figure S10 in Text S1), is not generally higher for these TFs. Thus, the role of these TFs in nucleosome positioning is not simply the result of increased binding or expression in YPD. Notably, for a considerable number of TFs our model predicts essentially no binding sites, and not all of these TFs are low expressed in YPD. It is conceivable that the low number of predicted sites for these TFs indicates that these TFs do not compete with nucleosomes but can bind to DNA which is wrapped around a nucleosome. We also investigated whether the top 20 TFs have particularly high or low information content and found that this is not the case (Figure S11 in Text S1).

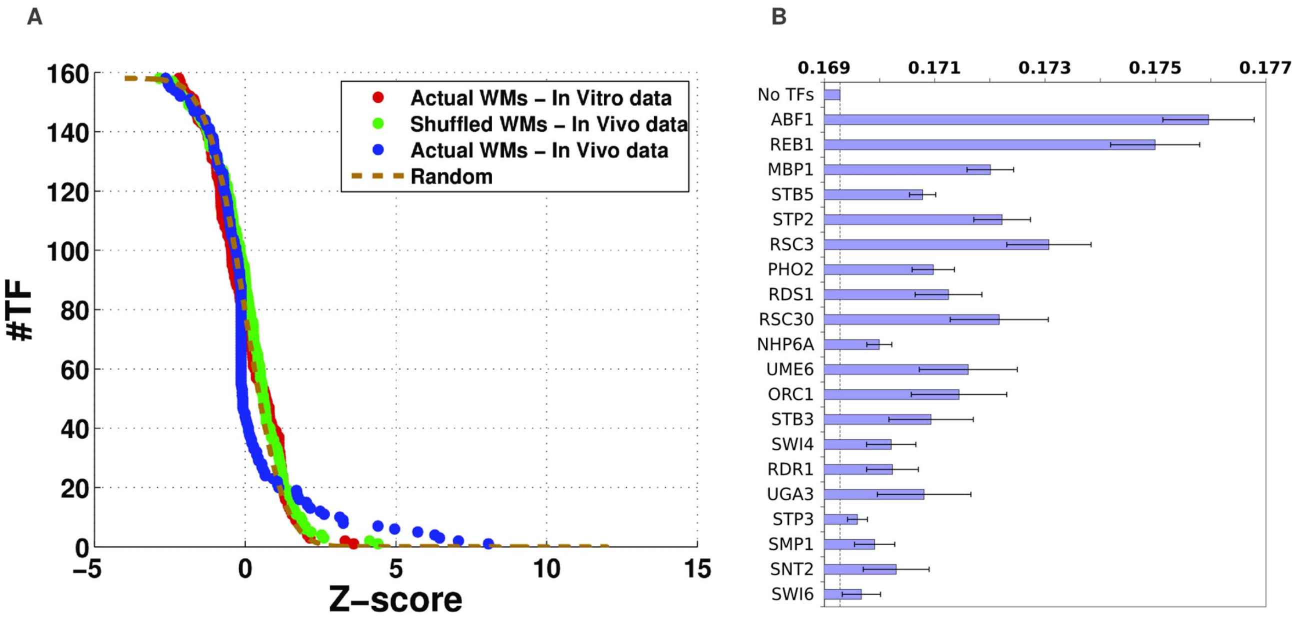

Figure 7. Only approximately 20 TFs contribute significantly to nucleosome positioning. A: For each TF an average quality score F across 5 test-sets was determined using the model containing nucleosomes and the corresponding TF. TFs were then ordered by the z-statistic $z=(F-F_{\text{noTFs}})/s_e$, with F_{noTFs} the quality score of the model without any TFs, and s_e the standard-error across the 5 test-sets (see Materials and Methods). The panel shows the reverse cumulative distribution of z-statistics observed across the 158 TFs (blue dots) together with the expected standard-normal distribution expected for random predictions (brown dotted curve). Note that about 20 TFs have z-statistics larger than expected by chance. The green dots show the reverse-cumulatives of z-statistics for the fits obtained with WMs in which the columns of each WM have been randomly shuffled. The red dots show the reverse-cumulatives of z-statistics obtained when fitting the original WMs to the *in vitro* map of nucleosome positions. Note that both the green and red dots closely follow the distribution expected by chance. **B**: The top 20 TFs that contribute most to *in vivo* nucleosome positioning sorted by their z-statistic. The bars show the average quality score F and standard-error s_e for each TF.
doi:10.1371/journal.pcbi.1003181.g007

However, when we manually inspected the functional annotation of the top 20 TFs, we noticed that roughly half of these TFs are known to be involved in chromatin remodeling (Table S1). Since, among our 158 TFs only 27 have been previously implicated in chromatin remodeling or nucleosome positioning, this amounts to a highly significant enrichment among our top 20 TFs (p-value 0.0016, see Materials and Methods). This suggested that the top 20 TFs may be characterized by interacting directly with chromatin modification machinery. To investigate this more systematically we investigated the occurrence of known direct protein-protein interactions between TFs and

1. Histones
2. Enzymes that modify histones
3. Proteins that are subunits of chromatin remodeling complexes

(see Materials and Methods). As detailed in Table 1, we find that our top 20 TFs are highly significantly enriched for direct protein-protein interactions with all 3 categories, showing the strongest enrichment for interacting directly with proteins in chromatin remodeling complexes. These results strongly suggest that our top 20 TFs are characterized by their ability to locally recruit chromatin modifiers.

The fact that only those TFs that interact directly with chromatin modifiers contribute significantly to explaining NFRs has interesting implications for the mechanisms of nucleosome positioning. It suggests that the creation of NFRs depends on the actions of chromatin modifiers whose activities lead to local expulsion of nucleosomes from the DNA. That is, the mechanistic picture that emerges is that, initially, the competition between TFs and nucleosomes for binding DNA, as implemented in our model, determines where TFs will end up binding DNA. Subsequently, in those places where TFs from the specific class that can recruit chromatin modifiers are bound, the recruitment of these modifiers will lead to local expulsion of the nucleosomes, leaving a larger region depleted of nucleosomes. This mechanistic picture also explains our previous observation that the predicted NFRs tend to be much narrower than those observed in the data.

Table 1. Statistical analysis of protein-protein interactions between TFs and chromatin remodeling complexes, histone modification enzymes, and histones.

Class	Total links	Links among top 20 TFs	p-value	Enrichment
Chromatin remodeler complexes	287	77	$9.2*10^{-11}$	3.26
Histone modification enzymes	369	74	$4.1*10^{-5}$	1.58
Histones	103	34	$7.3*10^{-8}$	2.6
All three classes	718	176	$4.1*10^{-18}$	1.94

For all yeast TFs we counted the number of 'links', i.e. known direct protein-protein interactions, with proteins from the functional categories shown in the first column. The second column shows the total number of links with all TFs, and the third column the number of links with the top 20 TFs that most significantly explain nucleosome positioning. The fourth column shows the p-value for the enrichment of links among the top 20 TFs using a hypergeometric test, and the 5 column shows the fold enrichment.
doi:10.1371/journal.pcbi.1003181.t001

Discussion

It is generally accepted that the packaging of DNA by nucleosomes in eukaryotes can modulate the accessibility of TFs to their cognate sites and thereby have major effects on gene regulation. In recent years there have been significant experimental efforts to determine nucleosome positioning patterns genome-wide, and to analyzing how these nucleosome-positioning patterns are established. As we discussed in the introduction, there has been a considerable debate as to whether nucleosome positioning in *Saccharomyces cerevisiae* is predominantly controlled by intrinsic sequence specificity of the nucleosomes, or that statistical positioning around barriers introduced by other DNA binding factors is more important for nucleosome positioning, and different researchers have presented seemingly contradictory results in this regard. We feel that these apparent contradictions may be reconciled by the results presented here.

The large majority of annotated nucleosomes and linkers genome-wide concern the phasing of short linkers within dense arrays of nucleosomes, mainly inside genes. We find that the positioning of these nucleosomes and short linkers crucially depends on the sequence specificity of the nucleosomes, and that TFs contribute relatively little to their positioning. Therefore, predicting all linkers and nucleosomes on a genome-wide scale, the sequence specificity of the nucleosomes provides the main contribution to explaining their positions. In contrast, we find that nucleosome specificity contributes little to explaining larger nucleosome free regions, especially those within promoter regions. As our modeling shows, NFRs in promoters are predominantly explained by the DNA binding of a specific class of $10-20$ transcription factors. Thus, while genome-wide locations of nucleosomes and short linkers are predominantly determined by nucleosome sequence-specificity, the large nucleosome free regions in promoters that likely contribute much more significantly to gene regulation, are determined mainly through the competitive binding of TFs. Importantly, the fact that competition with TFs can not help explain the *in vitro* nucleosome positioning shows that the contributions of the TFs is restricted to *in vivo* positioning. Thus, the competitive binding of TFs provides a quantitative and mechanistic explanation for the differences between *in vivo* and *in vitro* nucleosome occupancies.

That nucleosome free regions in promoters result from a competition between TF and nucleosome binding is supported by a number of recent studies of individual promoters, e.g. [9–11,50]. In these studies the interplay of TF and nucleosome binding determines positions of NFRs and the resulting accessibility pattern has major consequences for gene expression. Our results suggest that this mechanism is not restricted to a few promoters, but is the typical situation genome-wide. Thus, whereas nucleosome sequence specificity does have a major impact on genome-wide nucleosome positioning, precisely those aspects of nucleosome positioning that have most impact on gene regulation are rather determined by the competition between nucleosomes and TF binding.

Another major result from our study is that less than 20 of the 158 TFs that we analyzed appear to have a significant effect on nucleosome positioning. As we have shown, these TFs are not characterized by particularly high expression or large numbers of binding sites in YPD, nor do they possess particular sequence specificities or DNA binding domains. Instead, our analysis suggests that these TFs engage in specific protein-protein interactions with chromatin remodelers, thereby effecting nucleosome eviction much more dramatically than other TFs.

Although the final predictions of our statistical mechanical model are quite competent, i.e. in promoters 96% of all nucleosomes and 70% of all NFRs are correctly identified, they are still far from perfect. This raises the question as to what additional elements are missing from the model. The main error the model makes is failing to identify roughly one third of nucleosome free regions as nucleosome free. This suggests that the model misses additional factors that promote displacement of nucleosomes. As most sequence-specific TFs in yeast are already represented in the model, and our results suggest that only a small fraction of these TFs significantly affect nucleosome positioning, it seems unlikely that the missing sequence-specific TFs play a major role in the overall quality of the results. In contrast, as shown in Figure S6 in Text S1, general TFs including the RNA polymerase itself may play an important role in nucleosome positioning. In this context it has also been suggested [19] that the well-positioned nucleosome immediately downstream of TSS may result from a direct interaction between general transcription factors and the RNA polymerase with this nucleosome. Thus, including the recruitment and binding of general TFs and RNA polymerase will likely further improve the model.

In addition, TF binding can recruit chromatin modifying enzymes that displace nucleosomes and alter histone tails. The fact that experimentally observed NFRs are typically wider than the theoretically predicted ones suggest that the TF binding recruits chromatin modifiers which lead to a larger region of nucleosome exclusion than given by the TF binding itself. Thus, feed-back from TF binding to nucleosome modification and ejection as mediated by chromatin remodelers is a major feature that could improve the model's predictions. In summary, the picture that emerges from our study is that the binding of a specific class of $10-20$ TFs determines local recruitment of chromatin remodelers, which then mediate local expulsion of nucleosomes. The latter may further positively feed-back on TF binding and thereby expand and stabilize the nucleosome-free regions.

Although this work has focused on yeast, the competition between nucleosomes and TFs for binding DNA may even be more crucial for transcription regulation in higher eukaryotes. For example, in multi-cellular eukaryotes many gene regulatory elements occur in distal enhancers, i.e. local clusters of TF binding sites a few hundred base pairs in length, to which a combination of TFs binds to effect transcription at a promoter that can be hundreds of kilobases away. Recent mapping of enhancers based on chromatin marks has suggested that these enhancers are bound and activated in a highly tissue- and condition-specific manner [51,52]. An attractive simplified model for such tissue-specific binding is that nucleosomes by default cause DNA to be inaccessible and that TF binding is too weak to access individual TF binding sites. Only in areas where a cluster with many binding sites for precisely that subset of TFs that is highly expressed in the condition will these TFs jointly outcompete the nucleosomes and create a region of DNA accessibility and TF binding, i.e. similar to the qualitative model presented in [53]. We believe that the statistical mechanics model that we have used here, might also be useful to quantitatively investigate such models of enhancer function.

Materials and Methods

A statistical mechanical model of competitive binding of proteins to the DNA

Based on a combination of ChIP-chip data, *in vitro* binding data, and computational analysis [12,54,55], we previously curated [32] a collection of 158 position specific weight matrices (WMs) representing the sequence-specificities of 158 *S. cerevisiae* TFs. We let $w_t(i,\alpha)$ denote the WM probability that position i in a binding site for TF t contains nucleotide α. Consequently, the probability that a binding site for TF t has sequence s is given by

$$P(s|t) = \prod_{i=1}^{l_t} w_t(i,s_i), \tag{2}$$

where l_t is the length of the WM for TF t and s_i is the nucleotide at position i in sequence segment s. For our statistical mechanical model we wish to determine energies $E(s|t)$ for the binding of sequence segment s to TF t. We make the standard assumption that the binding energy is a sum of individual contributions from different nucleotides in the site, i.e.

$$E(s|t) = E_t^c + \sum_{i=1}^{l_t} E_t(i,s_i), \tag{3}$$

where E_t^c is a sequence-independent contribution to the binding energy. Under this assumption, the sequence-specific energy components $E_t(i,\alpha)$ can be shown [27,31] to be related to the WM components through

$$E_t(i,\alpha) = -\gamma_t \log[w_t(i,\alpha)], \tag{4}$$

where γ_t is a scale parameter, and the binding energy is expressed in units of $k_B T$.

There has been a significant amount of effort into modeling the sequence specificity of nucleosomes using data from both *in vivo* and *in vitro* experiments, e.g. [1,15,18,24]. As shown in Figure 1A, different models of nucleosome sequence-specificity give predicted occupancies that are very highly correlated, and the model of [18] exhibits the most robustly high performance. We thus took the model of [18] as the basis for calculating binding energies $E(s|\text{nucl})$ of the nucleosome to each possible 147 bp stretch s. Specifically, the raw probability $P(s|\text{nucl})$ of a 147 bp long sequence segment s under Kaplan et al's model can be obtained using the "nucleosome_prediction.pl" script, that is provided by the authors on their website, with default parameters and using the option "raw_binding". Using this we define a binding energy under the Kaplan model as

$$E_{\text{kaplan}}(s) = -\log[P(s|\text{nucl})] + c, \tag{5}$$

In order to allow us to tune the sequence specificity of the nucleosomes, we introduce a similar scale parameter γ_{nucl} to obtain

$$E(s|\text{nucl}) = \gamma_{\text{nucl}} E_{\text{kaplan}}(s). \tag{6}$$

Note that, at $\gamma_{\text{nucl}} = 1$, the sequence-specificity of this model will be equal to that of Kaplan et al's model, whereas at $\gamma_{\text{nucl}} = 0$ nucleosomes will have no sequence preferences whatsoever. For notational simplicity, in the following we will consider the nucleosome as just another member of the set T of all DNA binding factors t.

Let C denote a (non-overlapping) configuration of TFs and nucleosomes bound to the genome and let S_t denote all segments in the genome where a binding site for factor t occurs. Using the standard Gibbs-Boltzmann approach, the probability of finding the cell in configuration C is given by

$$P(C|c,\gamma) = \frac{1}{Z} \prod_t \prod_{s \in S_t} c_t e^{-\beta E(s|t)}, \tag{7}$$

where c_t is the concentration of TF t, $\beta = 1/(k_B T)$ is the inverse temperature, and Z is the partition function

$$Z = \sum_C \prod_t \prod_{s \in S_t} c_t e^{-\beta E(s|t)}. \tag{8}$$

Note that the probability depends on the scale factors γ through the dependence of the binding energies $E(s|t)$ on the scale factors.

Note that, since we will be fitting the scale factors γ_t, we can define

$$\tilde{\gamma}_t = \beta \gamma_t \tag{9}$$

and fit the $\tilde{\gamma}_t$. For notational simplicity, we will drop the tilde and refer to these rescaled gammas as simply γ_t. Note that this is equivalent to measuring the energy in units of $k_B T$.

Using only information about known binding sites, i.e. the WM entries w_α^i, we cannot determine the sequence-independent contribution E_t^c for each TF, which essentially controls how generally 'sticky' the TF is to DNA. To allow the comparison of binding energies of different TFs on a common scale we set E_t^c such that, in the limit of low TF concentrations, each TF has equal binding to the yeast genome. Specifically, we set E_t^c such that the average $\langle e^{-E(s|t)} \rangle = 1$, when averaging over all sequence segments s in the genome.

Using this reparametrization the probability of a configuration becomes simply

$$P(C|c,\gamma) = \frac{1}{Z} \prod_t \prod_{s \in S_t} c_t e^{-\gamma_t E_t^c + \gamma_t \sum_i \log[w_t(i,s_i)]}. \tag{10}$$

Figure 8 shows a cartoon illustrating various configurations C and the factors contributing to their probabilities.

The partition function can be calculated efficiently using recursion relations variously known as transfer matrices or dynamic programming, and this has been routinely used in the field to sum over non-overlapping configurations of hypothesized binding sites, e.g. [25–27,29,30]. Let Z_n denote the partition sum for all configurations up to position n in a given chromosome. We then have

$$Z_n = Z_{n-1} + \sum_t Z_{n-l_t} c_t e^{-\gamma_t E_t^c + \gamma_t \sum_{i=1}^{l_t} \log[w_t(i, s_{n-l_t+i})]}. \tag{11}$$

Similarly, we can calculate the 'backward' partition sums B_n from position n to the end of the chromosome. Finally, the probability that a binding site for factor t covers positions $(n+1)$ through

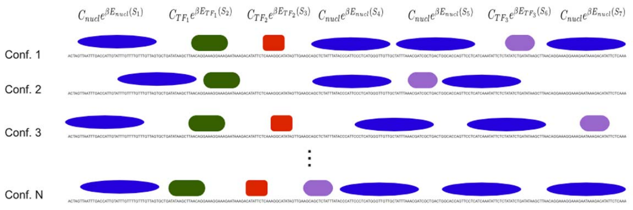
$C_{nucl}e^{\beta E_{nucl}(S_1)}$
$C_{TF_1}e^{\beta E_{TF_1}(S_2)}$
$C_{TF_2}e^{\beta E_{TF_2}(S_3)}$
$C_{nucl}e^{\beta E_{nucl}(S_4)}$
$C_{nucl}e^{\beta E_{nucl}(S_5)}$
$C_{TF_3}e^{\beta E_{TF_3}(S_6)}$
$C_{nucl}e^{\beta E_{nucl}(S_7)}$
Conf. 1
Conf. 2
Conf. 3
⋮
Conf. N

Figure 8. Illustration of example configurations of proteins bound to DNA. The top line indicates contributions from the individual binding sites to the overall probability of the configuration. Note that for illustration purposes, the sizes of TFs and nucleosomes are not shown to scale, e.g. the sizes of nucleosome footprints are much larger in reality.
doi:10.1371/journal.pcbi.1003181.g008

$(n+l_t)$ is given by

$$P(t,n|c,\gamma)=\frac{Z_n c_t e^{-\gamma_t E_t^c+\gamma_t \sum_{i=1}^{l_t}\log[w_t(i,s_{n+i})]}B_{n+l_t+1}}{Z_L}, \qquad (12)$$

where L is the chromosome length. The occupancy of factor t to position n is then given by $O(t,n|c,\gamma)=\sum_{i=n-l_t+1}^{n}P(t,n|c,\gamma)$. Thus, given a set of scale factors γ and concentrations c, we can efficiently calculate the occupancies of all 158 TFs and the nucleosomes across the entire yeast genome.

Experimentally determined positions of nucleosomes and linkers

To compare the 'raw' occupancies as predicted by various models of nucleosome specificity and measured across several *in vivo* and *in vitro* experiments, we first downloaded the per base occupancy predictions provided by [18] and [24] and used these predicted occupancies directly. We also obtained raw data from the experiments [1,3,18,38,56]. To obtain per-base nucleosome occupancies we calculated, for the ChIP-seq data, the number of reads overlapping each position and log-transformed these read counts. For the ChIP-chip data we log-transformed the chip signal. We observed that there is a very small number of positions for which sometimes aberrantly high or low signals are reported. To avoid having these outliers skew the observed correlations we removed the 0.5% of genomic positions with highest signal and 0.5% with lowest signal. We then directly calculated Pearson correlation coefficients between all data-sets and all predictions.

For the *in vivo* data, we make use of the reference map of nucleosomes and linkers for *S. cerevisiae* growing in YPD that was constructed by combining 6 different experimental data-sets in [41]. We only retained nucleosomes that were observed in all 6 datasets and have occupancy bigger then 80% (according to the authors' annotation). This set contained 21′252 nucleosomes covering 26% of the *S. cerevisiae* genome, and covers approximately 90% of all annotated nucleosomes in [41]. Linkers were defined as regions lying in between segments that were annotated as nucleosomes in any of the 6 data-sets. This set contained 60′448 linkers covering 26% of the *S. cerevisiae* genome. As observed in [41] the distribution of linker lengths is bimodal and we separately considered 'short linkers' (less than 80 bps long) and 'nucleosome free regions' (longer than 80 bps) in our analysis. There were 45′981 short linkers and 14′467 nucleosome free regions, covering 9% and 17% of the genome, respectively. We also separately considered the quality of the predicted nucleosome positions in promoter regions, defined as running from 500 bps upstream to 500 bps downstream of the TSS for each gene. The TSS definitions, as well as the definitions of the 3′ ends of genes, were taken from [57].

To assess the reproducibility of annotated nucleosome positions across the 6 experimental data-sets we calculated, for every nucleosome in the reference annotation, the standard-deviation in the positions of the associated annotated nucleosomes in each of the 6 data-sets. To compare the reproducibility of the annotated nucleosomes with what may be expected by chance, given the annotation procedure, we created randomized data-sets in which each sequencing read is mapped to a randomly chosen location in the genome. We then applied the same annotation procedure to this randomized data and calculated standard-deviations of the positions of annotated nucleosomes in the same way.

We constructed a reference map of *in vitro* nucleosome positioning using 3 independent data-sets from [18,19,58] using a procedure analogous to the one used in [41]. To annotate nucleosomes for every data-set we first run the GeneTrack software [59] using parameters $e=294$ (width of the exclusion zone corresponding to configurations with non-overlapping nucleosomes), $s=20$ (width of the smoothing gaussian kernel), $u=d=73$ (half-width of the peak) and $F=1$ (cut-off for peak height). The values of parameters e and u and d are dictated by the 147 bp width of the nucleosome footprint. Since the width $s=20$ of the smoothing kernel is much smaller than the nucleosome width, the final nucleosome annotation is insensitive to the precise width of this kernel. Similarly, raising the cut-off F by 2-fold or 4-fold would only slightly reduce the number of called nucleosomes (i.e. 1% and 5% respectively) and not substantially affect the results presented in the paper. We use the annotated nucleosomes as input to GeneTrack (with the same settings), i.e. as if each annotated nucleosome were a read, to produce annotated reference nucleosomes. We retained the roughly 75% of annotated reference nucleosomes that occur in all 3 data-sets, leaving 18′867 reference nucleosome genome-wide. Reference linkers were defined as regions not covered by nucleosomes in any of the annotations. There were 30′824 such linkers genome-wide.

Assessing the match between predicted nucleosome coverage and experimental nucleosome positioning

To compare the experimentally annotated linker and nucleosome regions with the predicted nucleosome coverage we proceeded as follows. For a given set of parameters, i.e. concentrations c and scale parameters γ, we first calculate the median of the predicted nucleosome occupancy across each annotated linker and nucleosome region. Given a critical median occupancy level O_{crit}, we then classified each region as either 'nucleosome' n when its median occupancy was larger than O_{crit} and 'linker' l when its median occupancy was less than or equal to O_{crit}. We then determined the fraction of regions both predicted and annotated as nucleosome $P_{nn}(O_{\text{crit}})$, the fraction of regions predicted as nucleosome and annotated as linker $P_{nl}(O_{\text{crit}})$, the fraction of regions predicted as linker and annotated as nucleosome $P_{ln}(O_{\text{crit}})$, and the fraction both predicted and annotated as linkers $P_{ll}(O_{\text{crit}})$. Using these we determined the *mutual information* between the predictions and the annotations based on the experimental data:

$$I(O_{\text{crit}},c,\gamma)=\sum_{i,j\in\{n,l\}}P_{ij}(O_{\text{crit}})\log\left[\frac{P_{ij}(O_{\text{crit}})}{P_i(O_{\text{crit}})P_j^e}\right], \qquad (13)$$

where $P_i(O_{\text{crit}})$ is the fraction of all regions predicted as i, P_j^e is the fraction of regions annotated as j, and we have explicitly indicated that this mutual information depends on the concentrations c and scale factors γ used in the predictions. We then define the mutual information $I(c,\gamma)$ as the maximal mutual information that can be obtained varying the critical occupancy O_{crit}, i.e.

$$I(c,\gamma)=\max_{O_{\text{crit}}}[I(O_{\text{crit}},c,\gamma)]. \qquad (14)$$

Finally, to normalize the mutual information on a more intuitive scale, we divide by the maximal possible mutual information, i.e. the entropy of the experimentally observed distribution:

$$H=-P_n^e\log[P_n^e]-P_l^e\log[P_l^e], \qquad (15)$$

to obtain

$$F(c,\gamma) = \frac{I(c,\gamma)}{H}. \tag{16}$$

Thus, $F(c,\gamma)$ is the fraction of the information regarding nucleosome and linker positioning that is captured by the predictions, which we refer to as the *quality score*. We calculate the mutual informations I and quality score F in an entirely analogous manner when considering a particular subset of experimentally annotated nucleosomes and linkers, i.e. excluding short linkers and/or focusing only on promoter regions.

To obtain predicted nucleosome coverage distributions we simply calculate the predicted occupancy at each position in the genome as described above. To obtain nucleosome coverage distributions from different experimental data-sets we proceeded as follows. As has been observed previously [20], especially for ChIP-seq data-sets, the variance in read coverage along the genome is too large to be consistent with the known overall nucleosome coverage of roughly 80%. Consequently, a naive normalization in which one assumes read-coverage to be directly proportional to nucleosome occupancy would lead to unrealistically low overall nucleosome coverage. To address this, we normalize the data by rescaling log read-coverage, similar to the normalization procedure we developed previously for next-generation sequencing data [60].

Specifically, for ChIP-chip data (from a tiling array with 4 bp resolution) we obtain a signal x_i corresponding to the log-ratio of signal from the nucleosome and background sample for each probe i along the genome. Similarly, for ChIP-seq data we extend each read to length 147 bp and defined the 'signal' x_i at each genomic position i as the logarithm of the number of reads overlapping position i. We assume that the signal x_i is *proportional* to the logarithm of the probability P_i that a nucleosome is bound to the corresponding segment in the genome, i.e

$$x_i = \lambda \log(P_i) + c, \tag{17}$$

where λ and c are unknown constants. We determine c and λ by demanding that the *average* coverage probability matches the experimentally observed average nucleosome coverage of 0.8, and that all coverage probabilities P_i must lie in the interval [0,1]. Finally, there is a small number of probes (0.1 percent of all probes) with an abnormally high signal x_i and we removed these outliers before fitting c and λ. As shown in Figure S1 in Text S1, this procedure leads to highly similar coverage distributions for different data-sets.

Predicted average nucleosome coverage profiles around transcription starts and ends were obtained by simply averaging the predicted nucleosome coverage at different positions relative to TSS and transcription end over all genes. We similarly averaged the experimental coverage profiles relative to transcription starts and ends.

Model fitting

To optimize the concentration and specificity scaling parameters (c,γ) we used the Melder-Mead algorithm in combination with a simulated annealing algorithm that is implemented in the GNU Scientific Library (GSL). To avoid over-fitting when fitting different models with varying numbers of parameters we used a 80/20 cross-validation scheme for each model and data-set. That is, for each data-set and model, we randomly divide the data-set of annotated nucleosomes and linkers into 5 equally sized sub-sets. We then perform the parameter fitting 5 independent times, each time optimizing the parameters on 80% of the data and then evaluating the final quality score of the model on the 'test-set' containing the remaining 20% of the data. Whereever quality scores are shown we show the average quality score and its standard-error across the 5 test-sets.

For the *in vivo* reference set of nucleosomes and linkers, we first performed optimizations of the nucleosome-only model with different (fixed) values of the specificity scaling parameter γ_{nucl}, i.e. optimizing only the concentration c_{nucl}. For both the *in vivo* and *in vitro* reference sets we optimized the two-parameter nucleosome-only model (obtaining an optimal $\gamma_{\text{nucl}} = 0.47$ for the *in vivo* data, and $\gamma_{\text{nucl}} = 0.41$ for the *in vitro* data). After this we fixed the nucleosome specificity and concentration to their optimal values and, for the *in vivo* data, fitted the model with all TFs, fitting the concentrations and scale parameters for all TFs.

For the biophysical characterization of the fitted model, we first averaged the fitted concentrations c and scale parameters γ over the 5 training sets. We then calculated the predicted posterior binding probabilities $P(t,n|c,\gamma)$ for every factor t (i.e. the nucleosomes and all TFs) at every position n in the yeast genome. For each factor t, we then calculated the fraction of the genome f_t covered by this protein: $f_t = l_t \sum_n P(t,n|c,\gamma)/L_{genome}$, where l_t is the length of the footprint of protein t and L_{genome} is the length of the yeast genome. We also calculated the average binding energy $\langle E_t \rangle$ of the binding sites of each protein t, i.e. $\langle E_t \rangle = \sum_n E_{t,n} P(t,n|c,\gamma)/[\sum_n P(t,n|c,\gamma)]$, and its standard deviation $\sigma(E_t) = \sqrt{\langle E_t^2 \rangle - \langle E_t \rangle^2}$. Here $E_{t,n}$ is the binding energy of protein t at position n, measured in units $k_B T$. Finally, we calculated the average entropy H_t per binding site:

$$H_t = \frac{-\sum_n P(t,n|c,\gamma)\log_2[P(t,n|c,\gamma)] + (1-P(t,n|c,\gamma))\log_2[1-P(t,n|c,\gamma)]}{\sum_n P(t,n|c,\gamma)}. \tag{18}$$

To calculate the information content for a TF t, as shown in Figure S11 of Text S1, we used the standard formula

$$IC(t,\gamma_t) = \sum_{i=1}^{l_t} \sum_{\alpha \in \{A,C,G,T\}} \omega_t(i,\alpha) \log_2[\frac{\omega_t(i,\alpha)}{p_\alpha}], \tag{19}$$

where the $p_\alpha = 0.25$ are background probabilities (which we chose uniform) and the $\omega_t(i,\alpha)$ are the weight matrix entries. Note that, to incorporate the scaling parameter γ_t, the weight matrix entries are rescaled according to:

$$\omega_{\text{scaled}}(i,\alpha) = \frac{[\omega_{\text{unscaled}}(i,\alpha)]^{\gamma_t}}{\sum_{\alpha'} [\omega_{\text{unscaled}}(i,\alpha')]^{\gamma_t}}. \tag{20}$$

To assess the contribution of different TFs we fitted, for each TF, the model with nucleosomes and this single TF. For each TF we calculated, on each of the 5 test-sets, the difference dF between the quality score using only the nucleosome, and the quality score with the TF added, and determined the mean $\langle dF \rangle$ and standard error $SE = \sigma(dF)/\sqrt{5}$ over the 5 test-sets. We then ranked the TFs by the z-statistic $z = \langle dF \rangle / SE$. These fits and statistics were obtained separately for both the *in vivo* and the *in vitro* data. Finally, we also created a set of 158 randomized WMs by, for each WM, randomly shuffling the columns of the WM. Note that this randomization conserves both the sequence composition and the information scores of the WMs. We then performed the fitting with these 158 randomized WMs and obtained *z*-statistics in the precise same way.

For the *in vivo* data we then also fitted models including the top 5, 10, 20, and 30 TFs from the list ranked by their *z*-statistic, re-optimizing all parameters. Finally, to assess the contribution of the nucleosome specificity when TFs are added for the *in vivo* data, we fitted the model including all TFs, but without nucleosome sequence specificity, i.e. setting $\gamma_{\text{nucl}} = 0$.

Annotating chromatin related TFs

To annotate TFs with known roles in chromatin dynamics we used the Gene Ontology (GO) annotations available from the Saccharomyces cerevisiae genome database. We considered a TF 'chromatin related' when its GO annotation included any of the following categories:

- GO:0016568 chromatin modification.
- GO:0006338 chromatin remodeling.
- GO:0008301 DNA bending activity.
- GO:0031491 nucleosome binding.
- GO:0003682 chromatin binding.
- GO:0033698 Rpd3C(L) A histone deacetylase complex which deacetylates histones across gene coding regions.

Finally, we also added the TFs identified in [12] to this list. To calculate the over-representation of 'chromatin related' TFs among the top 20 TFs effecting nucleosome positioning, we performed a simple hypergeometric test.

Protein-protein interactions between TFs, histones, and chromatin remodelers

We first annotated yeast proteins that are either (1) part of chromatin remodeling complexes, (2) histone modification enzymes, or (3) histones themselves. Subunits of chromatin remodeler complexes were taken from [61,62]. As subunits of histone modification enzymes we took genes that have GO annotation "covalent chromatin modification" and all children GO categories, i.e. histone methylation, acetylation etcera (108 genes in total). Information about protein-protein interactions were downloaded from the STRING database (http://www.string-db.org, file 'protein.links.detailed.v9.0.txt.gz'), using only experimental evidence with a cutoff of 400. After determining all known protein-protein interactions between the 158 TFs and the three classes of proteins (histones, histone modification enzymes, and subunits of chromatin remodeling complexes) we calculated enrichment of interactions between each class and the top 20 TFs that significantly explain nucleosome positioning. To assess the significance of the enrichment we used a simple hypergeometric test. The results are listed in Table 1.

Supporting Information

Table S1 Information for every transcription factor about Z-scores, fitted parameters and protein-protein interactions with chromatin remodeling complexes, histone modification enzymes and histones.
(XLS)

Text S1 Supplementary file. This pdf file contains the supplementary Figures S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11. The supplementary Table S1 is provided in separate excel file.
(PDF)

Acknowledgments

We thank the members of the van Nimwegen lab for useful discussions and feed-back.

Author Contributions

Conceived and designed the experiments: EvN. Performed the experiments: EAO. Analyzed the data: EAO. Contributed reagents/materials/analysis tools: EAO EvN. Wrote the paper: EAO EvN.

References

1. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39: 1235–1244.
2. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res 18: 1073–1083.
3. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, et al. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol 6: e65.
4. Simpson RT, Stafford DW (1983) Structural features of a phased nucleosome core particle. Proc Natl Acad Sci USA 80: 51–55.
5. Struhl K (1985) Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc Natl Acad Sci USA 82: 8419–8423.
6. Lowary PT, Widom J (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J Mol Biol 276: 19–42.
7. Kornberg RD, Stryer L (1988) Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucleic Acids Res 16: 6677–6690.
8. Mobius W, Gerland U (2010) Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. PLoS Comput Biol 6: e1000891.
9. Floer M, Wang X, Prabhu V, Berrozpe G, Narayan S, et al. (2010) A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. Cell 141: 407–418.
10. Bai L, Ondracka A, Cross FR (2011) Multiple sequence-specific factors generate the nucleosomedepleted region on CLN2 promoter. Mol Cell 42: 465–476.
11. Wang X, Bai L, Bryant GO, Ptashne M (2011) Nucleosomes and the accessibility problem. Trends Genet 27: 487–492.
12. Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, 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.
13. Koerber RT, Rhee HS, Jiang C, Pugh BF (2009) Interaction of transcriptional regulators with specific nucleosomes across the saccharomyces genome. Mol Cell 35: 889–902.
14. Ganapathi M, Palumbo MJ, Ansari SA, He Q, Tsui K, et al. (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.
15. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature 442: 772–778.
16. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, et al. (2007) Nucleosome positioning signals in genomic DNA. Genome Res 17: 1170–1177.
17. Chevereau G, Palmeira L, Thermes C, Arneodo A, Vaillant C (2009) Thermodynamics of intragenic nucleosome ordering. Phys Rev Lett 103: 188103.
18. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458: 362–366.
19. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, et al. (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol 16: 847–852.
20. Stein A, Takasuka TE, Collings CK (2010) Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? Nucleic Acids Res 38: 709–719.
21. Kaplan N, Hughes TR, Lieb JD, Widom J, Segal E (2010) Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. Genome Biol 11: 140.
22. Segal E, Widom J (2009) What controls nucleosome positions? Trends Genet 25: 335–343.
23. Chung HR, Dunkel I, Heise F, Linke C, Krobitsch S, et al. (2010) The effect of micrococcal nuclease digestion on nucleosome positioning data. PLoS ONE 5: e15754.

24. Locke G, Tolkunov D, Moqtaderi Z, Struhl K, Morozov AV (2010) High-throughput sequencing reveals a simple model of nucleosome energetics. Proc Natl Acad Sci U S A 107: 20998–21003.
25. Bussemaker HJ, Li H, Siggia ED (2000) Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. Proc Natl Acad Sci USA 97: 10096–100.
26. Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002) Computational detection of genomic cisregulatory modules, applied to body patterning in the early drosophila embryo. BMC Bioinformatics 3: 30.
27. van Nimwegen E (2007) Finding regulatory elements and regulatory motifs: a general probabilistic framework. BMC Bioinformatics 8 Suppl 6: S4.
28. Schwab DJ, Bruinsma RF, Rudnick J, Widom J (2008) Nucleosome switches. Phys Rev Lett 100: 228105.
29. Wasson T, Hartemink AJ (2009) An ensemble model of competitive multi-factor binding of the genome. Genome Res 19: 2101–2112.
30. Raveh-Sadka T, Levo M, Segal E (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. Genome Res 19: 1480–1496.
31. Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins: Statisticalmechanical theory and application to operators and promoters. J Mol Biol 193: 723–750.
32. Chen K, van Nimwegen E, Rajewsky N, Siegal ML (2010) Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. Genome Biol Evol 2: 697–707.
33. Tolkunov D, Zawadzki KA, Singer C, Elfving N, Morozov AV, et al. (2011) Chromatin remodelers clear nucleosomes from intrinsically unfavorable sites to establish nucleosome-depleted regions at promoters. Mol Biol Cell 22: 2106–2118.
34. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ (2006) Flexibility and constraint in the nucleosome core landscape of caenorhabditis elegans chromatin. Genome Res 16: 1505–1516.
35. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132: 887–898.
36. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the drosophila genome. Nature 453: 358–362.
37. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, et al. (2008) A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18: 1051–1063.
38. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, et al. (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol 4: e1000216.
39. Whitehouse I, Rando OJ, Delrow J, Tsukiyama T (2007) Chromatin remodelling at promoters suppresses antisense transcription. Nature 450: 1031–1035.
40. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10: R32.
41. Jiang C, Pugh BF (2009) A compiled and systematic reference map of nucleosome positions across the saccharomyces cerevisiae genome. Genome Biol 10: R109.
42. Shannon CE (1948) A mathematical theory of communication. Bell Sys Tech Journal 27.
43. Jaynes ET (2003) Probability Theory: The Logic of Science. Cambridge University Press.
44. Kornberg RD (1974) Chromatin structure: a repeating unit of histones and dna. Science 184: 868–871.
45. Jansen A, Verstrepen KJ (2011) Nucleosome positioning in Saccharomyces cerevisiae. Microbiol Mol Biol Rev 75: 301–320.
46. Venters BJ, Pugh BF (2009) A canonical promoter organization of the transcription machinery and its regulators in the saccharomyces genome. Genome Res 19: 360–371.
47. Fan X, Moqtaderi Z, Jin Y, Zhang Y, Liu XS, et al. (2010) Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3′-end formation. Proc Natl Acad Sci USA 107: 17945–17950.
48. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.
49. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, et al. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. Nat Biotechnol 27: 652–658.
50. Lam FH, Steger DJ, O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. Nature 453: 246–250.
51. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reect global cell-type-specific gene expression. Nature 459: 108–112.
52. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, et al. (2009) Chip-seq accurately predicts tissuespecific activity of enhancers. Nature 457: 854–858.
53. Mirny LA (2010) Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci U S A 107: 22534–22539.
54. Siddharthan R, Siggia ED, van Nimwegen E (2005) Phylogibbs: A Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1: e67.
55. Gordan R, Hartemink AJ, Bulyk ML (2009) Distinguishing direct versus indirect transcription factor-dna interactions. Genome Res 19: 2090–2100.
56. Mavrich T, Ioshikhes I, Venters B, Jiang C, Tomsho L, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res 18: 1073–1083.
57. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by rna sequencing. Science 320: 1344–1349.
58. Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, et al. (2011) A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science 332: 977–980.
59. Albert I, Wachi S, Jiang C, Pugh BF (2008) GeneTrack–a genomic data processing and visualization framework. Bioinformatics 24: 1305–1306.
60. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, et al. (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. Genome Biology 10: R79.
61. Smith CL, Horowitz-Scherer R, Flanagan JF, Woodcock CL, Peterson CL (2003) Structural analysis of the yeast SWI/SNF chromatin remodeling complex. Nat Struct Biol 10: 141–145.
62. Bao Y, Shen X (2007) SnapShot: chromatin remodeling complexes. Cell 129: 632.