

Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions

Marco Tompitak,^{1,*} Cédric Vaillant,² and Helmut Schiessel¹

¹Leiden University, Lorentz Institute, Leiden, the Netherlands; and ²Laboratoire de Physique, Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Lyon, France

ABSTRACT Sequences that influence nucleosome positioning in promoter regions, and their relation to gene regulation, have been the topic of much research over the last decade. In yeast, significant nucleosome-depleted regions are found, which facilitate transcription. With the arrival of nucleosome positioning maps for the human genome, it was discovered that in our genome, unlike in that of yeast, promoters encode for high nucleosome occupancy. In this work, we look at the genomes of a range of different organisms, to provide a catalog of nucleosome positioning signals in promoters across the tree of life. We utilize a computational model of the nucleosome, based on crystallographic analyses of the structure and elasticity of the nucleosome, to predict the nucleosome positioning signals in promoter regions. To be able to apply our model to large genomic datasets, we introduce an approximative scheme that makes use of the limited range of correlations in nucleosomal sequence preferences to create a computationally efficient approximation of the full biophysical model. Our predictions show that a clear distinction between unicellular and multicellular life is visible in the intrinsically encoded nucleosome affinity. Furthermore, the strength of the nucleosome positioning signals correlates with the complexity of the organism. We conclude that encoding for high nucleosome occupancy, as in the human genome, is in fact a universal feature of multicellular life.

INTRODUCTION

Nucleosomes are the fundamental packaging units of DNA that eukaryotic organisms employ to render their genomes compact enough to fit inside a cell, consisting of ~147 basepairs worth of DNA wrapped around a histone core. This packaging also restricts access to the genome: DNA bound to histones is unavailable for coupling to many other DNA-binding complexes, such as the transcriptional machinery. Therefore, the positioning of nucleosomes along the genome interacts with gene expression, as was already realized some three decades ago (1).

This interplay suggests that nucleosomes may play a role in gene regulation, and nucleosomes are in fact actively displaced to regulate gene expression (2,3). Genomic sequences may also have evolved to position nucleosomes in specific, beneficial locations. This possibility is suggested both by the fact that the degeneracy of the genetic code, in principle, allows for multiplexing

of such positioning signals with genetic information (4), and by the observation that the mutation patterns of DNA bound to histones differ from those of linker DNA (5).

Research into such nucleosome positioning signals, hardcoded into eukaryotic genomes, has veritably exploded over the last decade, primarily due to the development of experimental methods that allow for efficient genomewide nucleosome mapping (6). This research has provided insight into the importance of nucleosomal sequence preferences for chromatin organization (7), and has allowed for the creation, refinement, and testing of many models for predicting nucleosome positioning along genomes (8,9). The intrinsic nucleosome-DNA affinity of genomic sequences appears to play a significant role in vivo in positioning nucleosomes in certain regions of the genome, such as transcription start sites (TSSs) and origins of replication (7), alongside other effects like the presence of proteins that compete for the same DNA stretch or the action of chromatin remodelers (10,11).

Around the TSS of *Saccharomyces cerevisiae* (baker's yeast), nucleosomes have been found to be

Submitted July 11, 2016, and accepted for publication December 29, 2016.

*Correspondence: tompitak@lorentz.leidenuniv.nl

Editor: Tamar Schlick.

<http://dx.doi.org/10.1016/j.bpj.2016.12.041>

© 2017 Biophysical Society.



depleted on average, both in vitro and in vivo (12–18). The persistence of this depletion in vitro, in the absence of active remodeling, identifies the sequence preferences of nucleosomes as the dominant cause. Those preferences have been measured and utilized in various models to explain the observed nucleosome depletion (15,16,18–20). These nucleosome-depleted regions (NDRs) in gene promoters are thought to be encoded into the genomic sequence to allow RNA polymerases more ready access to the TSS, thereby facilitating transcription (13).

Since the earliest studies on baker's yeast, inquiries into nucleosome positioning have been extended to the genomes of many other organisms, such as *Schizosaccharomyces pombe* (21) and various other species of yeast (22), *Caenorhabditis elegans* (23,24), *Plasmodium falciparum* (25), flies (26), zebrafish (27), *Arabidopsis thaliana* (28), mice (29,30), and humans (30–35). Most of these studies were conducted in vivo, and therefore do not allow for isolation of effects encoded into the genomic sequences. This body of research shows, however, that sequence effects alone are not generally sufficient to explain in vivo observations (11). An important role is also played by the active regulation of transcription. In yeast, the promoters of actively transcribed genes show much more pronounced nucleosome depletion than those of inactive genes (21).

In human cells, as in yeast, NDRs were found in vivo only for actively expressed genes (31). However, in vitro nucleosome mapping reveals that the human genome does not share yeast's strategy of depletion-by-default. Instead, it was found that promoter regions in the human genome showed enhanced nucleosome occupancy. One interpretation is that this reflects the differentiated nature of human cells: it may be more beneficial to keep genes relatively inaccessible by default, and to actively open the promoter region only when needed (33,34). This idea seems to be countered by newer results, however, which find stronger intrinsic nucleosome-attracting regions (NARs) for housekeeping genes than for tissue-specific genes, directly opposite of what one would expect (36). Those results indicate that the function of the NARs in the human genome may be to retain nucleosomes in sperm cells (in which most nucleosomes are removed from the chromatin) and so pass on epigenetic information to the next generation.

Whichever is the case, these ideas raise the question whether the presence of an NDR in yeast versus that of an NAR in humans might be a general distinguishing feature between unicellular and multicellular life. To answer this question, we utilize a purely mechanics-based model for the sequence-dependent DNA-nucleosome affinity to predict in vitro nucleosome positioning signals, and compare the signals encoded into the promoter regions of a wide range of genomes.

MATERIALS AND METHODS

Data acquisition

All genomic sequences and gene (cDNA) data were downloaded from ensemblgenomes.org, release 31 (37). The in vitro nucleosome map produced by Kaplan et al. (18) was retrieved from GEO accession number GEO: GSE13622. The map from Valouev et al. (34) was downloaded from ccg.vital-it.ch/mga/hg18/valouev11/valouev11.html. The map from Locke et al. (38) was downloaded from <http://nucleosome.rutgers.edu/nucenergen/celegansnuc/>. The data from Ercan et al. (24) was taken directly from Fig. 1 C in that reference. TSS locations in *S. cerevisiae* were derived from David et al. (39) in the manner described in Vaillant et al. (40).

Model

We employed a statistical model inspired by that of Segal et al. (13), Field et al. (16), and Kaplan et al. (18). However, whereas their models are trained on experimental data, we employed this type of model to create a computationally inexpensive approximation to the theoretical nucleosome model recently published in Eslami-Mossallam et al. (4). The predictiveness of the Eslami-Mossallam nucleosome model has been examined in Eslami-Mossallam et al. (4), where it was found to outperform the experimentally informed models mentioned above, and in de Bruin et al. (41), where it is shown to be applicable not only to predictions for nucleosome positioning along a genome, but also the sequence-dependent response of nucleosomes to external forces.

We employed an extended version of the model presented in Segal et al. (13), which is informed by trinucleotide distributions, rather than dinucleotide distributions, because we found that this trinucleotide model leads to a more accurate approximation (see the [Supporting Material](#) and [Fig. S1](#) for more information).

The model of Segal et al. (13) requires as input position-dependent (di) nucleotide probabilities for the nucleosome. These can be derived from suitable sequence ensembles, as done in their article and its followup work. Such ensembles can also be generated in silico using the mutation Monte Carlo method of Eslami-Mossallam et al. (4) We applied the mutation Monte Carlo method to generate an ensemble of 10^7 high-affinity nucleosome sequences, from which we calculated the necessary di- and trinucleotide probability distributions. We found that the bioinformatical model approximated the full biophysical model with a root mean square deviation of 0.85 kT.

For this work, the parameterization of the nucleosome model was changed from the hybrid parameterization described in Eslami-Mossallam et al. (4), to a parameterization informed solely by crystallography data. We found that this improves its applicability to long-range effects. See the [Supporting Material](#) for more information.

Sequence analysis

For every genome analyzed, we calculated the averaged signal as follows. For every annotated gene, we looked up the location of the TSS, and extracted the 1146 bp before and after. For each of the resulting sequences, we calculated a probability landscape for nucleosome positioning using the trinucleotide model mentioned above. We would like to calculate occupancies from these landscapes and average over all genes. Unfortunately, because the probabilities vary over several orders of magnitude, the number of genes is generally not large enough to provide a meaningful average; it tends to be dominated by the highest probabilities. Therefore, we instead consider the average energy landscape for a given organism.

From the predicted probabilities, an energy landscape can be calculated up to a constant shift, because such a probability is the normalized Boltzmann weight of a state. We took the average of the energy landscapes of all the sequences as a representative energy landscape for a given organism. For each basepair (−1000 to +1000), we then calculated the nucleosome occupancy by summing the Boltzmann probabilities of all 147 nucleosome positions

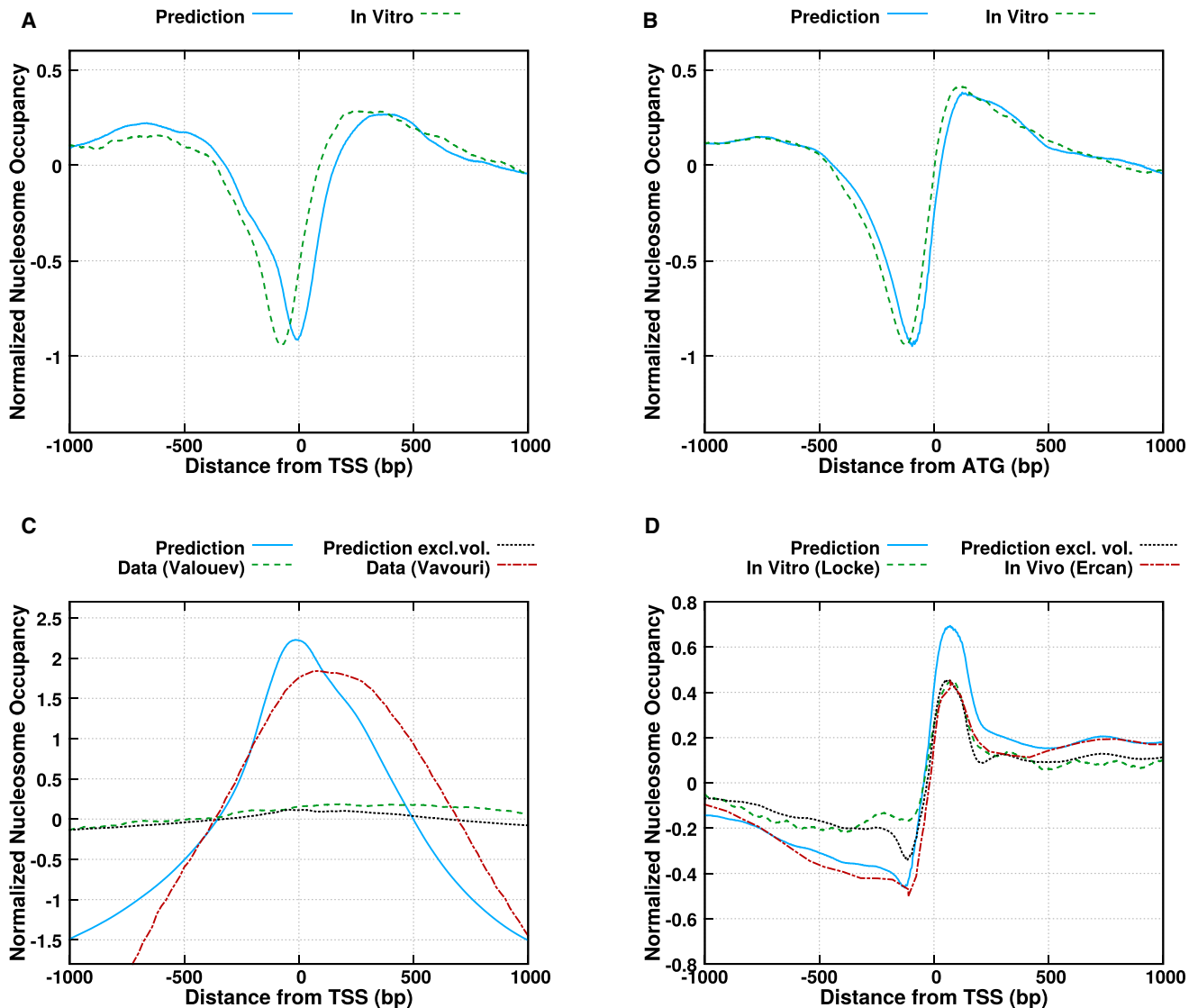


FIGURE 1 Comparison of predicted and measured intrinsic nucleosome positioning signals in promoter regions. The quantities plotted are the natural logarithms of the occupancies and the signals have been normalized such that they average to zero. (Solid blue curves) Our predictions in the limit of low nucleosome density, which give an account of the strength of the signals intrinsically encoded; (dashed green curves) in vitro measurements; (dotted black curves) predictions taking into account the steric interactions. Using the same treatment as in Chevreaux et al. (44), these curves have a free parameter $\bar{\mu} = \mu - \langle E \rangle$, i.e., the difference between the chemical potential and the average energy of the landscape, which we determined to be -8.5 kT for yeast (curves not shown due to similarity with the low-density limit), -5.7 kT for *C. elegans*, and -1.38 kT for humans. (A and B) *S. cerevisiae*, average nucleosome occupancy centered on the TSS and start codons, respectively. Data from Kaplan et al. (18). (C) Like (A), for *Homo sapiens*. The in vitro data is from Valouev et al. (34). Additionally shown is the nucleosome retention signal from Vavouri and Lehner (36). (D) Like (A), for *C. elegans*. The in vivo data is from Ercan et al. (24); the in vitro data is from Locke et al. (38). To see this figure in color, go online.

that lead to that basepair being covered by the nucleosome. This gives us a prediction of the intrinsic nucleosome affinity encoded in the genomic sequences.

RESULTS AND DISCUSSION

Opposing nucleosome occupancy signals in yeast and human genomes

The high-coverage *S. cerevisiae* nucleosome maps provide the standard testing ground for any model designed

to predict nucleosome occupancy. Applying our nucleosome affinity model (see Materials and Methods), we find we can correctly predict NDRs in the promoter regions of *S. cerevisiae*. The comparisons, for regions centered on the TSSs and on the start codons, are shown in Fig. 1, A and B, respectively.

For the human genome, a map of in vitro nucleosome occupancy has been published by Valouev et al. (34), and, as predicted by Tillo et al. (33), it reveals occupancy signals opposite to that of yeast: human promoters seem to encode

for high, rather than low, nucleosome occupancy. Vavouri and Lehner (36) similarly find an increased retention of nucleosomes when nucleosomes are depleted in human sperm cells. Correspondingly, when applying our model to the promoter regions of the human genome, we find a very strong NAR around the TSS, as can be seen in Fig. 1 C.

Initially surprisingly, the signal found by Valouev et al. (34) is an order-of-magnitude smaller than that predicted by our model and that found by Vavouri and Lehner (36). This discrepancy can be explained when we consider that the nucleosome density cannot exceed 1 per 147 bp due to excluded volume. The experiment attempts to measure enrichment of nucleosomes in the promoter regions relative to the average density of nucleosomes. Unlike in experiments that look at nucleosome depletion or retention, the excluded volume between nucleosomes puts a limit on how strong the enrichment can be in practice.

This is the reason for the discrepancy between the in vitro results of Valouev et al. (34) and ours and those of Vavouri and Lehner (36). To approximate the effects of steric interactions, we applied Percus' equation (42) to our average energy landscapes, and solved it as described in Vanderlick et al. (43). The solution depends on the chemical potential of the nucleosomes binding to the DNA (see also Chevereau et al. (44)), which we adjust to get a good fit with the in vitro data. We see that steric interactions can indeed explain the very weak signal for humans (*dotted black curve* in Fig. 1 C) as well as the apparent overshoot of our prediction for *C. elegans* (same in Fig. 1 D).

This means that at physiological conditions, the nucleosome density will be saturated at much smaller values due to steric interactions. However, we stress that independent of this saturation effect, a nucleosome at the peak of the nucleosome occupancy signal will be strongly energetically bound, and so hinder transcription if it is not actively removed, as well as be more stable under a nucleosome-depleting force.

The results of Vavouri and Lehner (36) when examining where nucleosomes are retained when they are depleted from chromatin in human sperm are more in line with our predictions, as can also be seen in Fig. 1 C. When depleting nucleosomes, excluded-volume interactions are not a constraint and our predictions can be probed. Although these authors studied a special in vivo situation, the nucleosome retention signals were found to correlate strongly with DNA sequence. Because the depletion of nucleosomes in sperm is an out-of-equilibrium process, and our model therefore does not make direct numerical predictions for this situation, we note the similarity between our predictions and the in vivo nucleosome retention signal.

We thus have interesting observations and predictions on two ends of a spectrum. A very simple, unicellular eukaryote shows nucleosome depletion as its most prominent, intrinsically encoded nucleosome positioning feature. A complex

multicellular one shows high nucleosome occupancy instead. What happens in between these two extremes?

In Fig. 1 D we present a comparison between our predicted signal for *C. elegans* and the signals found in vitro by Locke et al. (38) and in vivo by Ercan et al. (24). We find remarkable agreement in the shape of the signal, indicating that the data is indeed indicative of intrinsically encoded nucleosome positioning. Somewhat surprisingly, the in vitro and in vivo signals are similar to each other, which is not as strongly the case for yeast, and even less so for humans (see e.g., Fig. 3 in Vavouri and Lehner (36)). It has been noted that an in vivo nucleosome occupancy map of the nematode *C. elegans* lacks many of the features that distinguish in vivo maps from in vitro maps of yeast, such as strongly phased nucleosomes. Valouev et al. (23) find much flexibility in nucleosome positions in *C. elegans*. Such variability may average out some of the effects of active remodeling, rendering the two maps similar.

C. elegans seems to show a nucleosome positioning signal that is a hybrid of the signals found in the yeast and human genomes. It has an NDR upstream of the TSS, like yeast, but it also shows a significant NAR just after the TSS.

Intrinsic nucleosome positioning signals are indicative of multicellularity

The hybrid behavior in *C. elegans* may be hypothetically explained. As suggested by Tillo et al. (33), organisms may wish to tune their genomic sequences to intrinsically deactivate genes that are active only in some cell types, while intrinsically activating those that are common to all of its cells. In unicellular life, most genes will not be permanently silenced, leading to an overall average depletion signal. In complex multicellular life, the signal may be dominated by the many genes that are intrinsically deactivated, leading to an overall attractive signal. *C. elegans* may then represent a range of organisms where the two contributions are more equal, leading to both a depleted region just before the start codon (where it is also observed in yeast) and an attractive region just after (the peak in occupancy in the human genome is also skewed toward the right).

The results of Vavouri and Lehner (36), however, suggest that, at least in the human genome, the hypothesis of Tillo et al. (33) does not hold, and the function of the NARs is to retain nucleosomes in sperm cells. The hybrid signal we find in *C. elegans* may in this case similarly play a dual role of facilitating initiation of transcription, but at the same time assist in nucleosome retention.

We can extend our observation of these signals to other genomes using our model. We mapped the nucleosome positioning signals for promoters in genomes across the tree of life and discovered organisms that have intrinsically encoded NDRs and NARs, as well as many that fall into the hybrid category.

Most archaea (14 genomes analyzed) show a signal similar to that of yeast, in that a nucleosome-depleted region is the most prominent feature (Fig. S2). Archaea are unicellular organisms that do not have histone octamers, but employ only tetramers of (archaeal) histones to compactify their DNA. We expect these tetramers to obey positioning rules similar enough to nucleosomes that our model is predictive of their occupancy. We therefore analyzed the octamer affinity landscapes, for the sake of comparison to eukaryotes, even though archaea do not possess them. The signals show that these simple unicellular organisms almost all fall into the depletion-by-default category.

Fungi (seven genomes analyzed) show somewhat more diverse signals, Fig. S3. While *S. cerevisiae* has a prominent NDR, many of the other fungi analyzed lack both a localized depleted region and a localized attractive region, but retain a step-function signal centered on the TSS. Fungal cells are not highly differentiated, but some fungi are dimorphic (they switch between unicellular and filamentous states), possibly causing these more hybridlike signals.

Plants (four genomes analyzed) come in many forms, from unicellular algae to complex multicellular life. As expected, we see various signals (Fig. S4). The genome of *Chlamydomonas reinhardtii*, a unicellular alga, shows an NDR. Among the multicellular plants, we see two signals with a strong NAR, and one with hybrid behavior.

Among animals (24 genomes analyzed) we also find various signals. In worms, like *C. elegans*, we find both hybrid signals and more NAR-like signals (Fig. S5). *Drosophila melanogaster* and other members of its genus show strong hybrid signals, with a swift rise in nucleosome occupancy at the TSS (Fig. S6). Finally, the zebrafish genome and all mammalian genomes analyzed (human, chimpanzee, and mouse) have strong NARs (Fig. S7).

We see a clear separation between unicellular and multicellular organisms. Although some signals from unicellular lifeforms show some hybrid characteristics, the dominant feature is generally an NDR. All multicellular genomes, on the other hand, either encode for high nucleosome occupancy in the promoter region, or show hybrid signals. This distinction persists across the eukaryotic phylogenetic tree and is clearly visible in Fig. 2, where we have plotted a representative set of signals, divided into unicellular and multicellular classes. We finally note that these signals qualitatively correlate well with GC content (Fig. S8), suggesting that GC content is a prominent factor in shaping mechanical signals in promoter regions.

Intrinsic nucleosome positioning signals correlate with complexity

One proposed measure for organism complexity is the number of different cell types an organism possesses (45), and the ideas presented here clearly have a link to this measure. Unfortunately, numerical data describing the numbers of

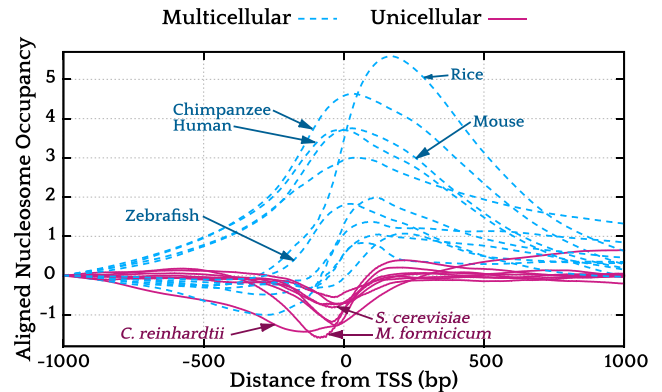


FIGURE 2 A representative selection of nucleosome positioning signals from various genomes. As a visual aid, the signals have been shifted vertically such that the logarithmic nucleosome occupancy at position -1000 is 0. The signals clearly fall into two distinct classes, based on whether the organism is unicellular or multicellular. To see this figure in color, go online.

cell types does not appear to be readily available in the literature, so we were unable to define a numerical measure of complexity. Therefore, we have restricted ourselves to ordering the organisms, by making assumptions about the cell type numbers. From simple to complex, we list: archaea, unicellular eukaryotes, filamentous and dimorphic fungi, multicellular plants, nematodes, *Drosophila* flies, zebrafish, and mammals.

We then considered the strength and direction of the NDR/NAR signals. To quantify this, we calculated the maximum and minimum of the signal and took the difference with the signal value at position -1000 relative to the start codon. We then took the largest of these two values (in the absolute sense) and designated this value as the signal's strength (not in the absolute sense; a dominant NDR gives a negative signal strength).

The signal strength as thus defined clearly distinguishes unicellular and multicellular lifeforms (Welch's t (39.051) = 10.5512, p -value 5.4×10^{-13}) and the signals for multicellular organisms show correlation with our complexity ordering (Spearman $r_s = 0.52$, p -value 82.3×10^{-3}), as shown in Fig. 3. The ordering of the organisms is almost certainly imperfect, for example because all multicellular plants have been lumped together; without more accurate knowledge of the cell type numbers, there is no way to place them more realistically. However, the NDR/NAR strengths show a tentative trend. All unicellular eukaryotes have a negative signal strength, indicating an NDR, as noted in the previous section. All multicellular eukaryotes (with one exception, *D. melanogaster*) have a stronger NAR than NDR, and the strength of this NAR roughly increases with complexity. This observation concurs with the hypothesis of Tillo et al. (33). Our expectation based on that hypothesis would be that a more differentiated organism will have more genes that are nucleosome-occupied by default, leading to a higher NAR signal. It is not clear what purpose

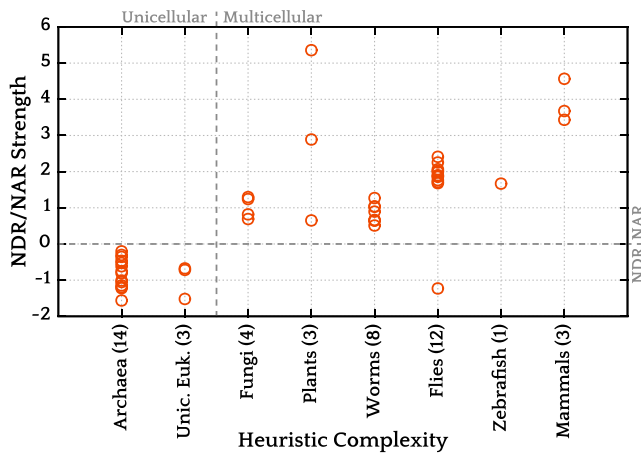


FIGURE 3 Promoter nucleosome positioning signal strength grouped by a heuristic measure of complexity of the organisms. The numbers in parentheses indicate how many genomes fall in each category. To see this figure in color, go online.

this correlation might serve in the context of nucleosome retention in the germline.

CONCLUSIONS

We found that the recently discovered fact that the human genome, unlike the yeast genome, encodes (on average) for an NAR rather than an NDR in the promoter region, is in fact a universal feature of multicellular life. The hypothesis put forth by Tillo et al. (33) is that this NAR suppresses gene transcription and that this suppression helps an organism with differentiated cell types manage its gene expression. Genes that are not needed in every cell type are suppressed by default, and only activated in those cells where they are necessary. In unicellular lifeforms, however, most genes will be in constant use, and keeping those genes easily accessible is more favorable.

On the other hand, Vavouri and Lehner (36) have found that the NARs found in humans in fact serve a different purpose, namely the retention of certain nucleosomes in sperm cells, and their study of the signals found for housekeeping genes versus tissue-specific genes directly contradicts the hypothesis of Tillo et al. (33). The NARs we find in multicellular life may therefore instead be indicative of the need to retain nucleosomes in the germ cells of multicellular organisms.

NARs are common to complex multicellular lifeforms, while almost all unicellular lifeforms we analyzed have NDRs. In-between there is a range of organisms with hybrid positioning signals. In almost all of these signals, however, the NAR is a more prominent feature than the NDR. This leads to a clear distinction between uni- and multicellular life based on the type of nucleosome positioning signals found in the promoter regions.

Furthermore, the strength of the NAR appears to increase with organism complexity. This fits the hypothesis of Tillo

et al. (33), because organisms with more cell differentiation will have more genes suppressed by an NAR (and possibly by stronger ones). If the purpose of the NARs is solely to retain nucleosomes in the germline, it seems that more complex life cares more strongly about retaining its nucleosomes and passing on epigenetic information. More research will be needed to explore this idea.

Given the presence of hybrid signals, we speculate that the encoding of NARs versus NDRs in promoter regions is not an all-or-nothing choice for organisms. Whether the NARs serve to close off genes by default, or to retain nucleosomes in the germline, they compete with an apparent need to create an NDR to facilitate the initiation of transcription. The organisms showing hybrid signals seem to strike a balance between the two.

Outlook

We hope that our results will motivate the experimental community to expand the available catalog of in vitro nucleosome maps to a greater number and variation of organisms. This will help not only verify our findings but also be of great service to any followup inquiries into the deeper nature and meaning of the signals we have found. We also suggest that nucleosome maps be generated at lower nucleosome densities, because steric hindrance will hide strong enrichment signals.

We also hope to encourage further examination of housekeeping versus tissue-specific genes in other organisms to further test the hypothesis of Tillo et al. (33), and an expansion of the results of Vavouri and Lehner (36) to other organisms, to test whether nucleosome retention in the germline is a goal served by the mechanical signals we find in the genomes of other complex organisms. If so, our results raise an intriguing question: why do more complex organisms tend to favor stronger nucleosome retention?

SUPPORTING MATERIAL

Supporting Materials and Methods, Supporting Results, and eight figures are available at [http://www.biophysj.org/biophysj/supplemental/S0006-3495\(17\)30035-8](http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)30035-8).

AUTHOR CONTRIBUTIONS

H.S. and C.V. designed the study; M.T. devised and built the model; M.T. and C.V. performed the analyses; and M.T., C.V., and H.S. contributed to the article.

ACKNOWLEDGMENTS

We thank Alain Arneodo, Benjamin Audit, Remus Dame, and Bram Henneman for discussions.

This work was supported by the Netherlands Organisation for Scientific Research (NWO/OCW), as part of the Frontiers of Nanoscience program.

SUPPORTING CITATIONS

References (46–48) appear in the Supporting Material.

REFERENCES

- Han, M., and M. Grunstein. 1988. Nucleosome loss activates yeast downstream promoters in vivo. *Cell*. 55:1137–1145.
- Becker, P. B., and J. L. Workman. 2013. Nucleosome remodeling and epigenetics. *Cold Spring Harb. Perspect. Biol.* 5:a017905.
- Lorch, Y., and R. D. Kornberg. 2015. Chromatin-remodeling and the initiation of transcription. *Q. Rev. Biophys.* 48:465–470.
- Eslami-Mossallam, B., R. D. Schram, ..., H. Schiessel. 2016. Multiplexing genetic and nucleosome positioning codes: a computational approach. *PLoS One*. 11:e0156905.
- Makova, K. D., and R. C. Hardison. 2015. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* 16:213–223.
- Tolkunov, D., and A. V. Morozov. 2010. Genomic studies and computational predictions of nucleosome positions and formation energies. *Adv. Protein Chem. Struct. Biol.* 79:1–57.
- Iyer, V. R. 2012. Nucleosome positioning: bringing order to the eukaryotic genome. *Trends Cell Biol.* 22:250–256.
- Teif, V. B. 2015. Nucleosome positioning: resources and tools online. *Brief. Bioinform.* 17:745–757.
- Liu, H., R. Zhang, ..., S. Zhou. 2014. A comparative evaluation on prediction methods of nucleosome positioning. *Brief. Bioinform.* 15:1014–1027.
- Struhl, K., and E. Segal. 2013. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* 20:267–273.
- Zhang, Z., C. J. Wippo, ..., B. F. Pugh. 2011. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*. 332:977–980.
- Yuan, G.-C., Y.-J. Liu, ..., O. J. Rando. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*. 309:626–630.
- Segal, E., Y. Fondufe-Mittendorf, ..., J. Widom. 2006. A genomic code for nucleosome positioning. *Nature*. 442:772–778.
- Albert, I., T. N. Mavrich, ..., B. F. Pugh. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*. 446:572–576.
- Lee, W., D. Tillo, ..., C. Nislow. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39:1235–1244.
- Field, Y., N. Kaplan, ..., E. Segal. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* 4:e1000216.
- Shivaswamy, S., A. Bhinge, ..., V. R. Iyer. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* 6:e65.
- Kaplan, N., I. K. Moore, ..., E. Segal. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*. 458:362–366.
- Ioshikhes, I. P., I. Albert, ..., B. F. Pugh. 2006. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* 38:1210–1215.
- Yuan, G. C., and J. S. Liu. 2008. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* 4:e13.
- Lantermann, A. B., T. Straub, ..., P. Korber. 2010. *Schizosaccharomyces pombe* genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae*. *Nat. Struct. Mol. Biol.* 17:251–257.
- Tsankov, A. M., D. A. Thompson, ..., O. J. Rando. 2010. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.* 8:e1000414.
- Valouev, A., J. Ichikawa, ..., S. M. Johnson. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18:1051–1063.
- Ercan, S., Y. Lubling, ..., J. D. Lieb. 2011. High nucleosome occupancy is encoded at X-linked gene promoters in *C. elegans*. *Genome Res.* 21:237–244.
- Bunnik, E. M., A. Polishko, ..., K. G. Le Roch. 2014. DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite *Plasmodium falciparum*. *BMC Genomics*. 15:347.
- Mavrich, T. N., C. Jiang, ..., B. F. Pugh. 2008. Nucleosome organization in the *Drosophila* genome. *Nature*. 453:358–362.
- Zhang, Y., N. L. Vastenhouw, ..., X. S. Liu. 2014. Canonical nucleosome organization at promoters forms during genome activation. *Genome Res.* 24:260–266.
- Liu, M., A. E. Seddon, ..., S. Shiu. 2015. Determinants of nucleosome positioning and their influence on plant gene expression. *Genome Res.* 25:1182–1195.
- Teif, V. B., Y. Vainshtein, ..., K. Rippe. 2012. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.* 19:1185–1192.
- Fenouil, R., P. Cauchy, ..., J. C. Andrau. 2012. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.* 22:2399–2408.
- Ozsolak, F., J. S. Song, ..., D. E. Fisher. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* 25:244–248.
- Schones, D. E., K. Cui, ..., K. Zhao. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell*. 132:887–898.
- Tillo, D., N. Kaplan, ..., T. R. Hughes. 2010. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS One*. 5:e9129.
- Valouev, A., S. M. Johnson, ..., A. Sidow. 2011. Determinants of nucleosome organization in primary human cells. *Nature*. 474:516–520.
- Gaffney, D. J., G. McVicker, ..., J. K. Pritchard. 2012. Controls of nucleosome positioning in the human genome. *PLoS Genet.* 8:e1003036.
- Vavouri, T., and B. Lehner. 2011. Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome. *PLoS Genet.* 7:e1002036.
- Kersey, P. J., J. E. Allen, ..., D. M. Staines. 2016. EnsemblGenomes 2016: more genomes, more complexity. *Nucleic Acids Res.* 44(D1):D574–D580.
- Locke, G., D. Haberman, ..., A. V. Morozov. 2013. Global remodeling of nucleosome positions in *C. elegans*. *BMC Genomics*. 14:284.
- David, L., W. Huber, ..., L. M. Steinmetz. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA*. 103:5320–5325.
- Vaillant, C., L. Palmeira, ..., A. Arneodo. 2010. A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res.* 20:59–67.
- de Bruin, L., M. Tompitak, ..., H. Schiessel. 2016. Why do nucleosomes unwrap asymmetrically? *J. Phys. Chem. B*. 120:5855–5863.
- Percus, J. K. 1976. Equilibrium state of a classical fluid of hard rods in an external field. *J. Stat. Phys.* 15:505–511.
- Vanderlick, T. K., L. E. Scriven, and H. T. Davis. 1986. Solution of Percus's equation for the density of hard rods in an external field. *Phys. Rev. A Gen. Phys.* 34:5130–5131.
- Chevereau, G., L. Palmeira, ..., C. Vaillant. 2009. Thermodynamics of intragenic nucleosome ordering. *Phys. Rev. Lett.* 103:188103.
- Valentine, J. W., A. G. Collins, and C. P. Meyer. 1994. Morphological complexity increase in metazoans. *Paleobiology*. 20:131–142.
- Olson, W. K., A. A. Gorin, ..., V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*. 95:11163–11168.
- Calladine, C. R., and H. R. Drew. 1984. A base-centred explanation of the B-to-A transition in DNA. *J. Mol. Biol.* 178:773–782.
- Becker, N. B., L. Wolff, and R. Everaers. 2006. Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Res.* 34:5638–5649.

Biophysical Journal, Volume 112

Supplemental Information

Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions

Marco Tompitak, Cédric Vaillant, and Helmut Schiessel

Supplementary Methods

Model: The model used in this work to predict nucleosome affinity is based on that of Segal *et al.* (1), which is a model for the thermodynamic probabilities for 147-base-pair sequences to reside in a nucleosome. That is, it provides a method to calculate the probability $P(S)$ of a sequence S related to the energy cost E of using a DNA molecule with this sequence to form a nucleosome:

$$(1) \quad P(S) \propto e^{-E/kT}$$

This probability depends on every one of the nucleotides that make up the sequence S . If we define S as a set of S_i with i an index running from 1 to 147, we can write

$$(2) \quad P(S) = P\left(\bigcap_i S_i\right)$$

Using the chain rule of probabilities, this can be rewritten as

$$(3) \quad P(S) = \prod_{n=1}^{147} P(S_n | \bigcap_{i=1}^{n-1} S_i)$$

This equation expresses the probability of the whole sequence as simply the product of all the separate base pairs in the sequence. The catch is that the probabilities of the base pairs are all interdependent; the probability for S_n depends on the values of S_1 through S_{n-1} .

The way the model of Segal *et al.* is obtained is by assuming that long-range correlations between base pairs can be neglected in the expression above. Specifically, they assume that the probability distribution of S_n depends only on the value of S_{n-1} and not on any base pairs further away, so that

$$(4) \quad P(S_n | \bigcap_{i=1}^{n-1} S_i) \approx P(S_n | S_{n-1})$$

If we apply this assumption, we obtain the model of Segal *et al.*

For the model to make predictions, it needs to be parameterized. Segal *et al.* and follow-up work (1–3) produced experimental thermodynamic ensembles of sequences with high affinity for nucleosomes. The probability of a given sequence in such an ensemble should be described by the model above, so one counts the prevalences of the dinucleotides and mononucleotides at every nucleosomal position in this average to produce the probability distributions needed to inform the model.

We here repurpose this model for a somewhat different endeavor. Another common approach to investigating nucleosome affinity is to model the energetics of the nucleosome directly. This can be done with a DNA model such as the Rigid Base Pair model (4) and a suitable model for the nucleosome. We have made use here of the nucleosome model presented in (5). This model can also be used to predict nucleosome affinity, based on the local elastic properties of base pair steps. Unfortunately, this model is computationally very expensive and cannot be used to analyze large numbers of sequences, such as entire genomes.

Such a model can, however, in a reasonable amount of time, be used to generate sequence ensembles of the same kind as employed to parameterize the Segal *et al.* model and follow-ups. Using a recently published computational method (Mutation Monte Carlo, (5)) we were able to generate ensembles large enough that probability distributions of mono-, di- and even trinucleotides could be calculated. When we plug those distributions into the Segal *et al.* model, we find that we have a good approximation of the predictions made by the full underlying nucleosome model, which is computationally far less expensive and allows us to analyze whole genomes.

We finally note that we used not the dinucleotide-based model of Segal *et al.*, but we have extended it to trinucleotides:

$$(5) \quad P(S) = P(S_1)P(S_2|S_1) \prod_{n=3}^{147} P(S_n|S_{n-1} \cap S_{n-2})$$

In this case, we make the assumption that the probability of S_n depends on the values of S_{n-1} and S_{n-2} . This assumption on the correlations between base pairs is less stringent than that of the dinucleotide model and should therefore provide a better approximation. The downside is that many more probability values need to be calculated, and a correspondingly larger sequence ensemble is required. However, we found that we were able to create a large enough ensemble (10^7 sequences) that the trinucleotide model provided a significant improvement over the dinucleotide model. When predicting the affinities of all 147-base-pair subsequences of the first chromosome of *S. cerevisiae*, the trinucleotide model came to a root-mean-square deviation of 0.85 kT when comparing its predictions with those of the underlying energetic model. The dinucleotide model yielded a deviation of 1.08 kT, so the trinucleotide model reduces the deviation by about 20%.

For the underlying nucleosome model, we chose the same model presented in (5). However, we have made an important alteration to the model in order to perform the analyses presented here. Previously, a hybrid parameterization was chosen for the Rigid Base Pair Model (6) that underlies the nucleosome model presented there. In this hybrid parameterization (7), the intrinsic deformations of the base pair steps are derived from crystal-structure data, and the stiffnesses of the steps from all-atom molecular dynamics simulations.

This hybrid model had previously been found to approximate reality best by Becker *et al.* (7). Those authors, however, used only short sequences to test the different parameterizations. Hence they primarily tested the local accuracies of the parameterizations, for which the correct oscillatory behavior of the predicted energy with the helical repeat of DNA is most important.

However, we are interested not in the local changes in affinity, but in long-range effects on the order of tens of helical repeats. For this purpose, we found that the hybrid parameterization yields unsatisfactory results. Although it gives correctly phased dinucleotide probability distributions, the average abundances of AT-rich dinucleotide steps in high-affinity sequences are overestimated with respect to those of GC-rich steps. It is known that high GC content correlates with high affinity, but the hybrid model ascribes higher affinity to AT-rich sequences. See Fig. S1. The result is that the model is unable to detect the nucleosome-depleted regions in *S. cerevisiae* promoters.

We find that when using a parameterization where both the intrinsic deformations and the stiffnesses are derived from crystal-structure data (4), the model does correctly ascribe high affinity to high GC content. See Fig 1b. When using this pure parameterization for our model, we find we do detect the NDR in yeast.

We speculate that the two parameterizations can fulfill complementary roles. The hybrid model may be most accurate when considering local changes in affinity, but its performance in detecting long-range effects is lacking. Conversely, the pure crystallography parameterization may not be as realistic locally (7), but it is able to capture long-range effects much more accurately. For this work we therefore applied the pure parameterization.

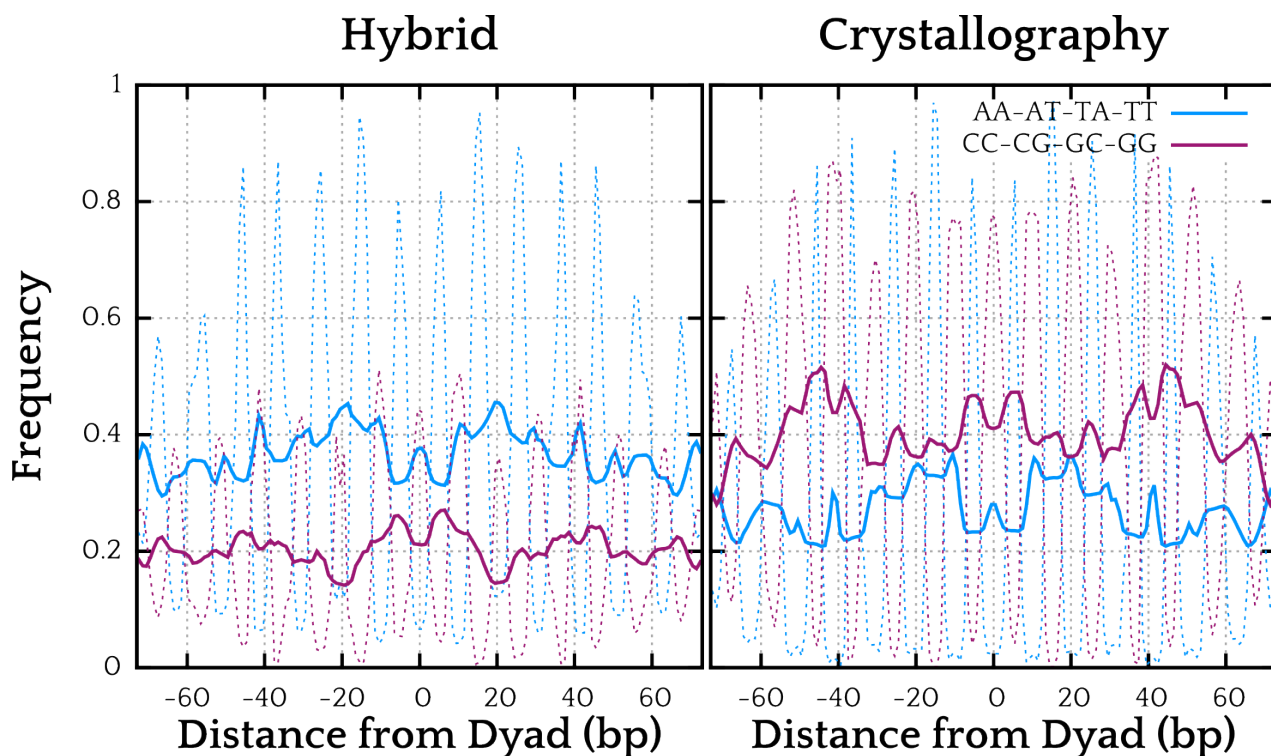


Fig. S1: Dinucleotide step frequencies and their 11-bp averages in high-affinity nucleosome ensembles. Left: Using the hybrid parameterization, AT-rich dinucleotide steps are enriched, while GC-rich steps are depleted. Right: In the pure parameterization, GC-rich steps are enriched, in line with experimental evidence.

Supplementary Results

Full set of mechanical signals: In this section we supply the full set of nucleosome positioning signals centered on transcription start sites. The signals are plotted in Figs. S2-S7, with organisms grouped together under a number of headings.

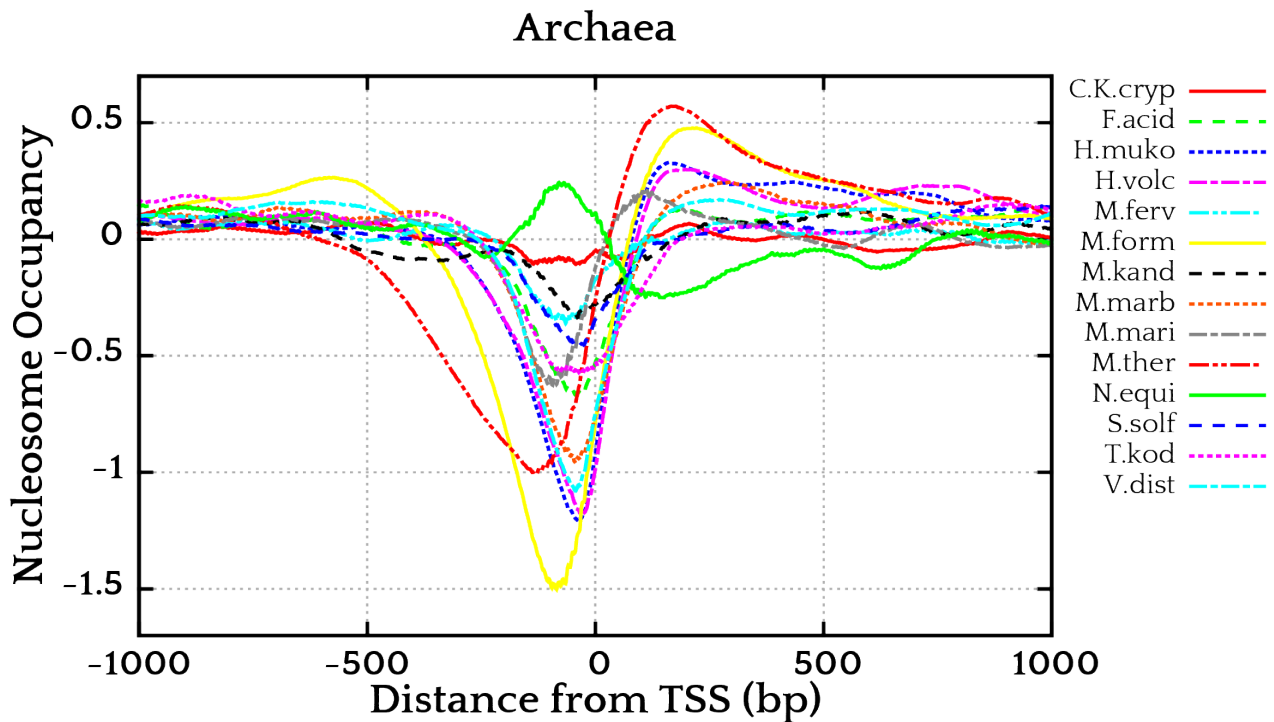


Fig. S2: Nucleosome positioning signals in the promoter regions of a number of Archaea.

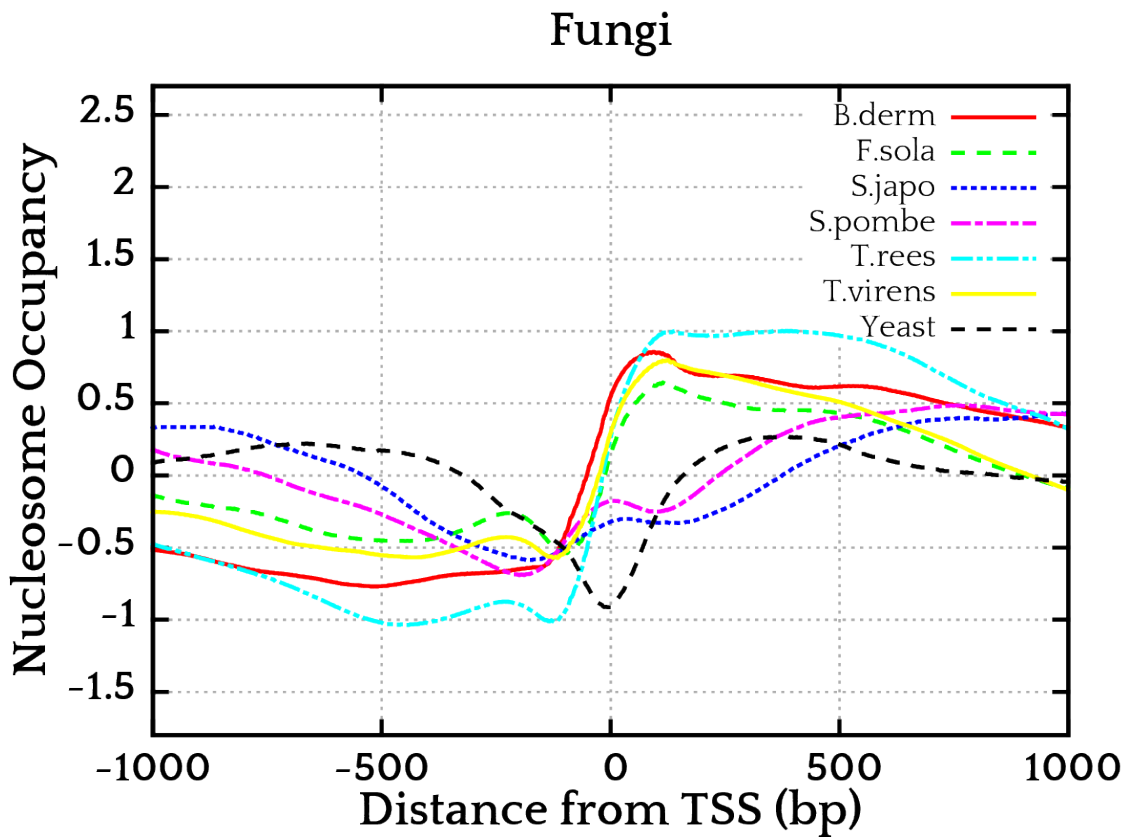


Fig. S3: Nucleosome positioning signals in the promoter regions of a number of fungi.

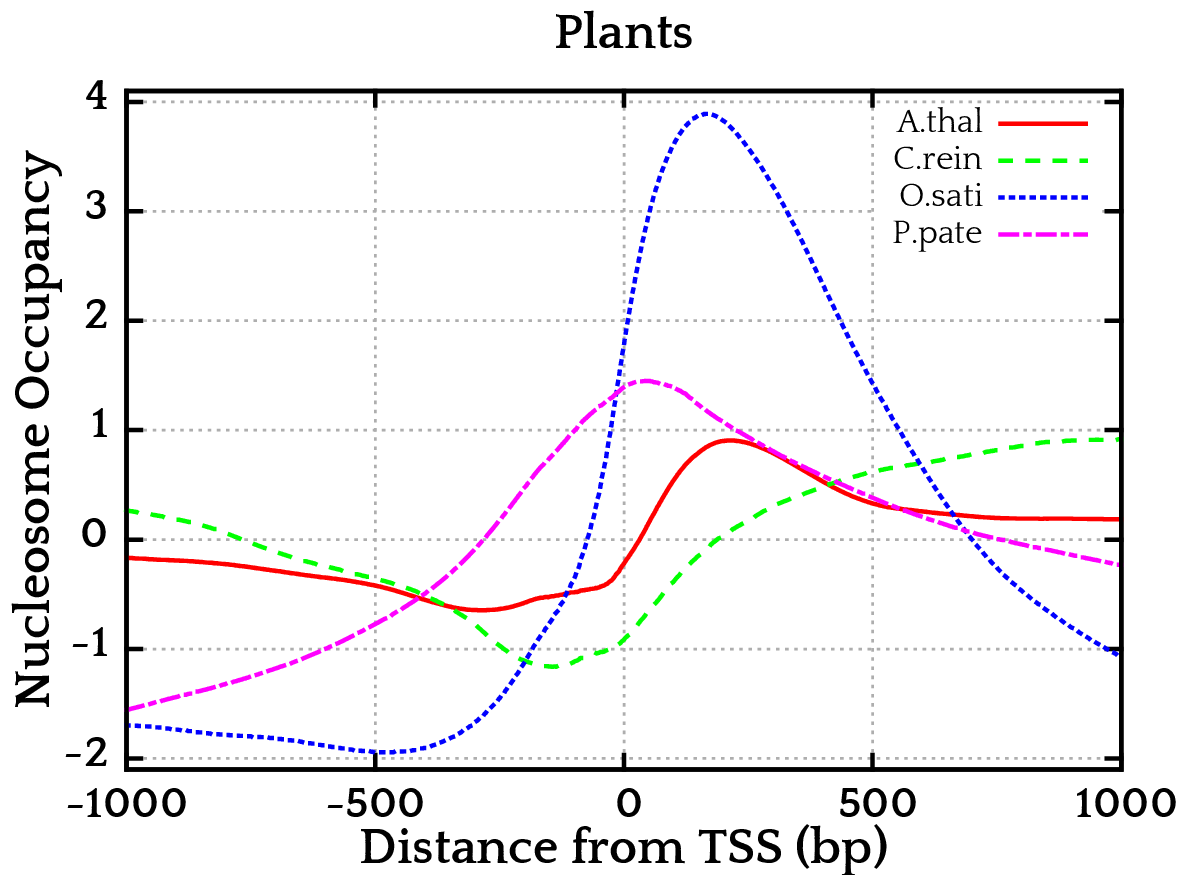


Fig. S4: Nucleosome positioning signals in the promoter regions of a number of plants.

Worms

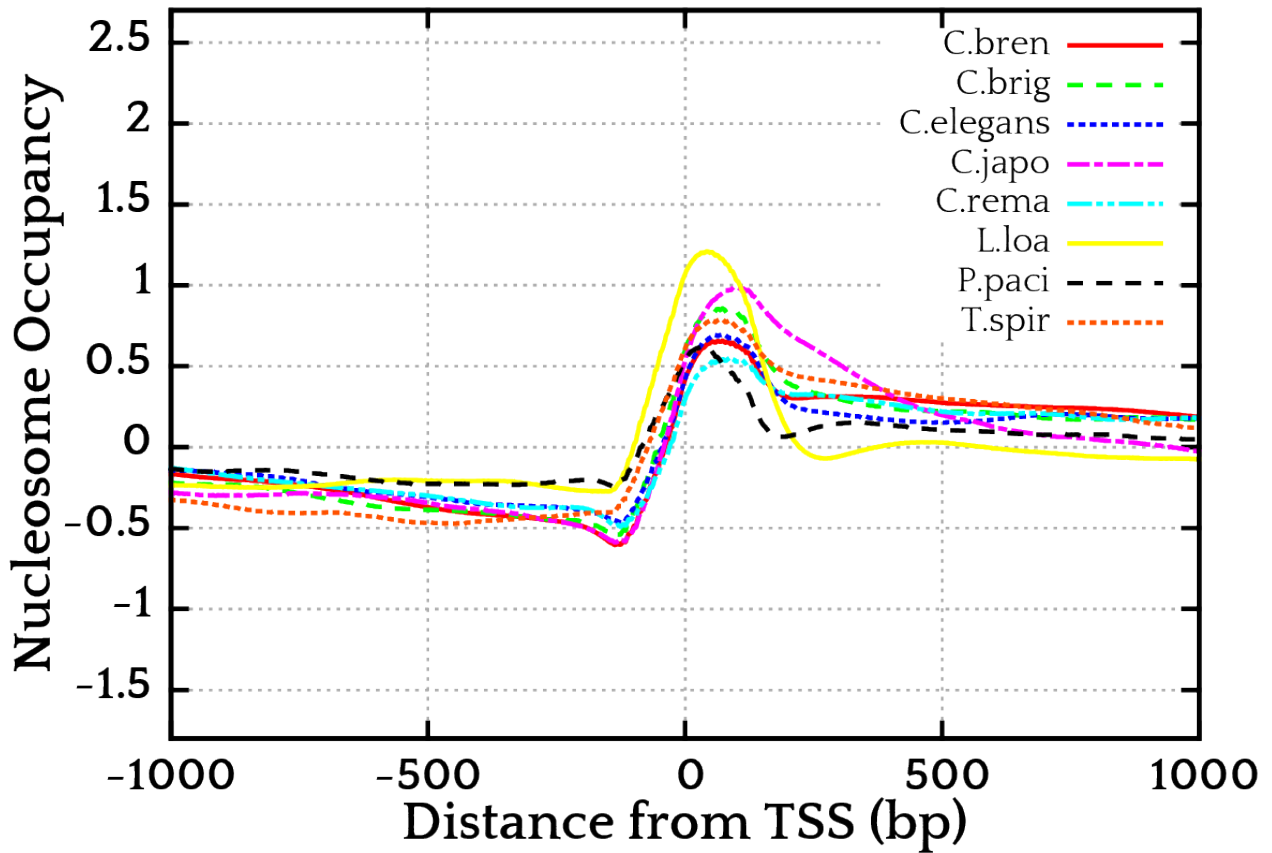


Fig. S5: Nucleosome positioning signals in the promoter regions of *C. elegans* and a number of other nematodes.

Flies

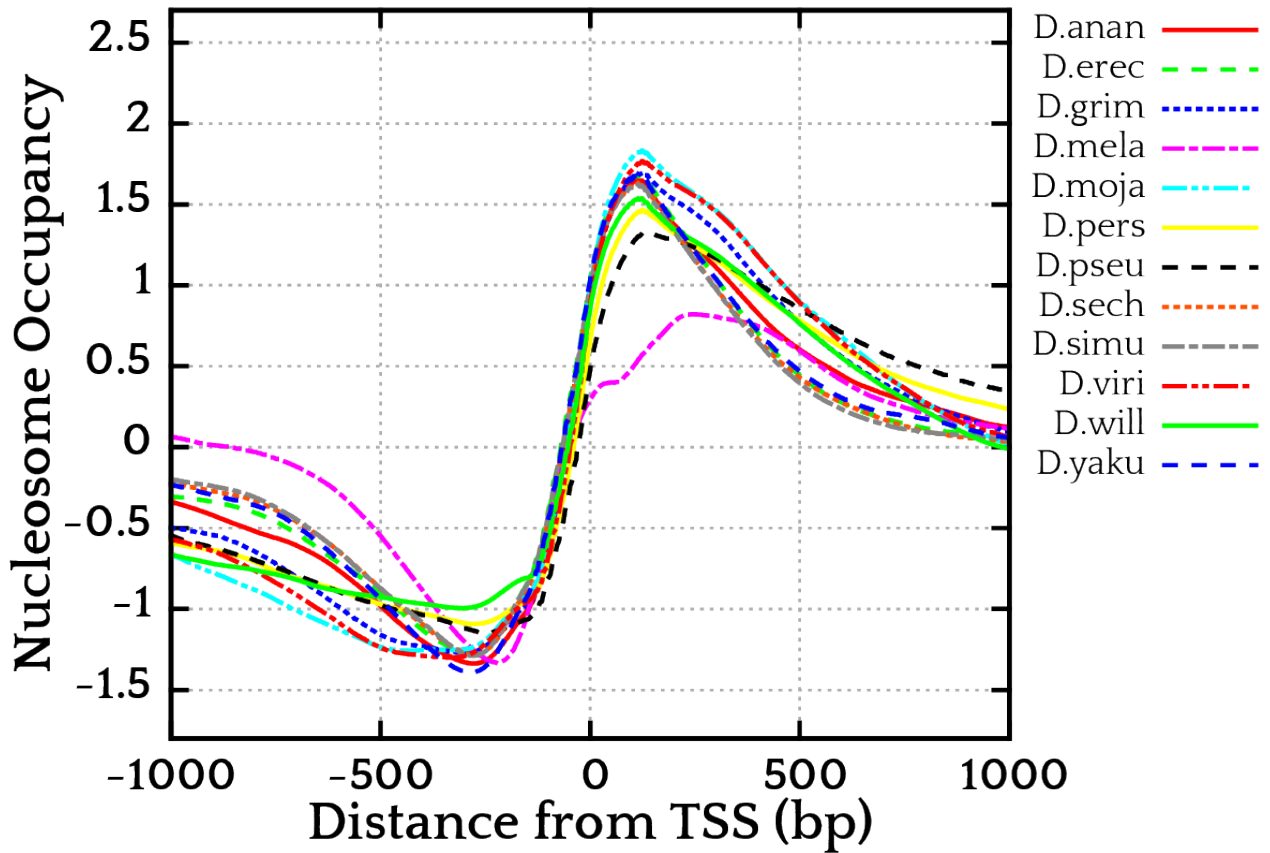


Fig. S6: Nucleosome positioning signals in the promoter regions of *D. melanogaster* and a number of other flies.

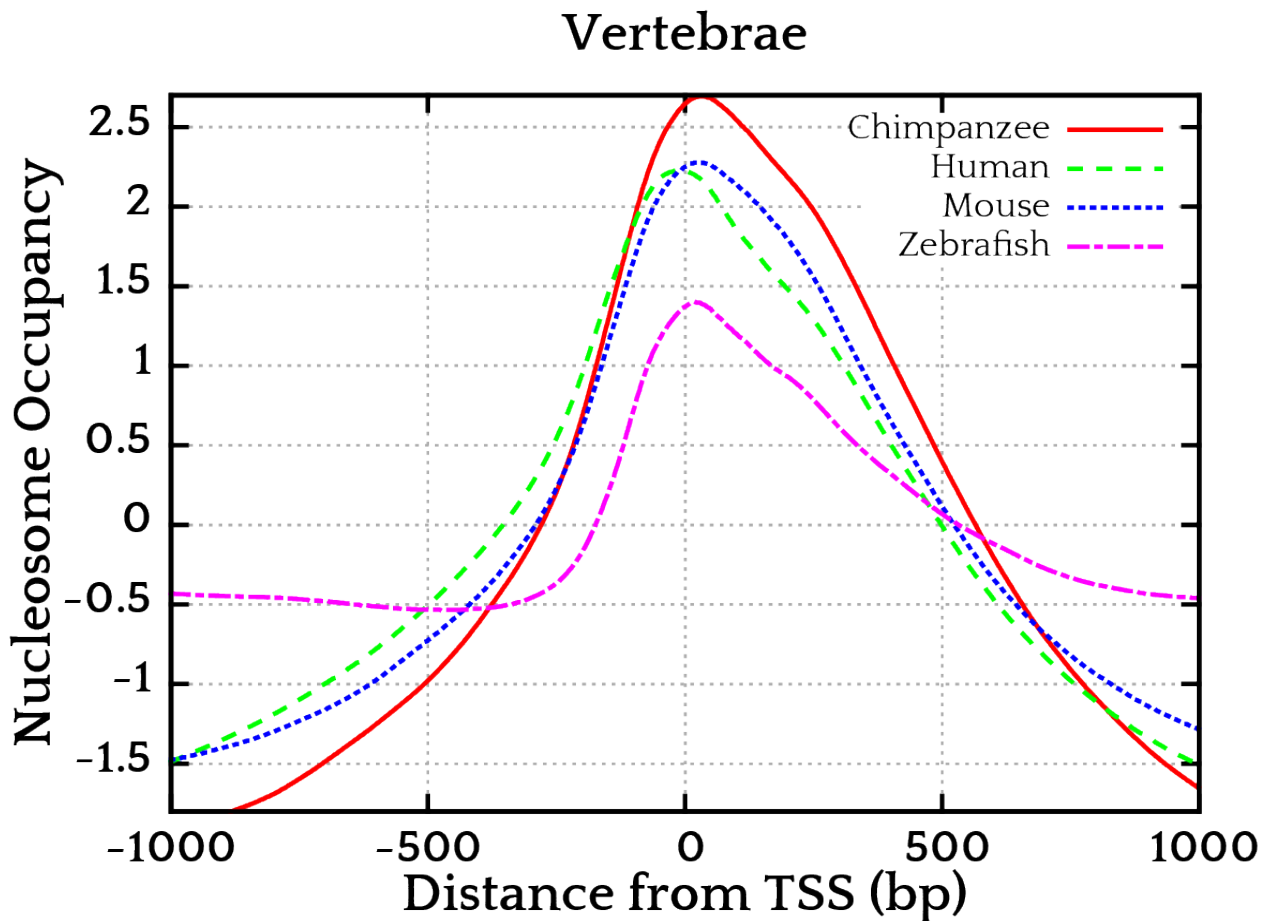


Fig. S7: Nucleosome positioning signals in the promoter regions of human, chimpanzee, mouse and zebrafish genomes.

GC content as signal predictor: Finally we wish to note that, in terms of classifying these signals as we have done in Fig. 2 in the main manuscript, one might also look at the signals in the GC content, which are depicted in Fig. S8. The visual similarity with Fig. 2 is of course striking.

We would warn against relying on GC content alone for the purpose for which we have applied our model here. The first reason is that, obviously, GC content in itself does not tell us anything about the numerical values of the nucleosome occupancy without some sort of calibration. Our model, on the other hand, has no free parameters, and is built on physical principles.

Secondly, we have also found that, using the Mutation Monte Carlo method with the Eslami-Mossallam nucleosome model [5], we can create sequences with very different mechanical properties by only changing the order of the sequence, while keeping GC content fixed, which shows that GC content is only part of the story.

That said, statistically, signals in GC content in promoter regions may also be a fruitful way to classify organisms. This will require further study.

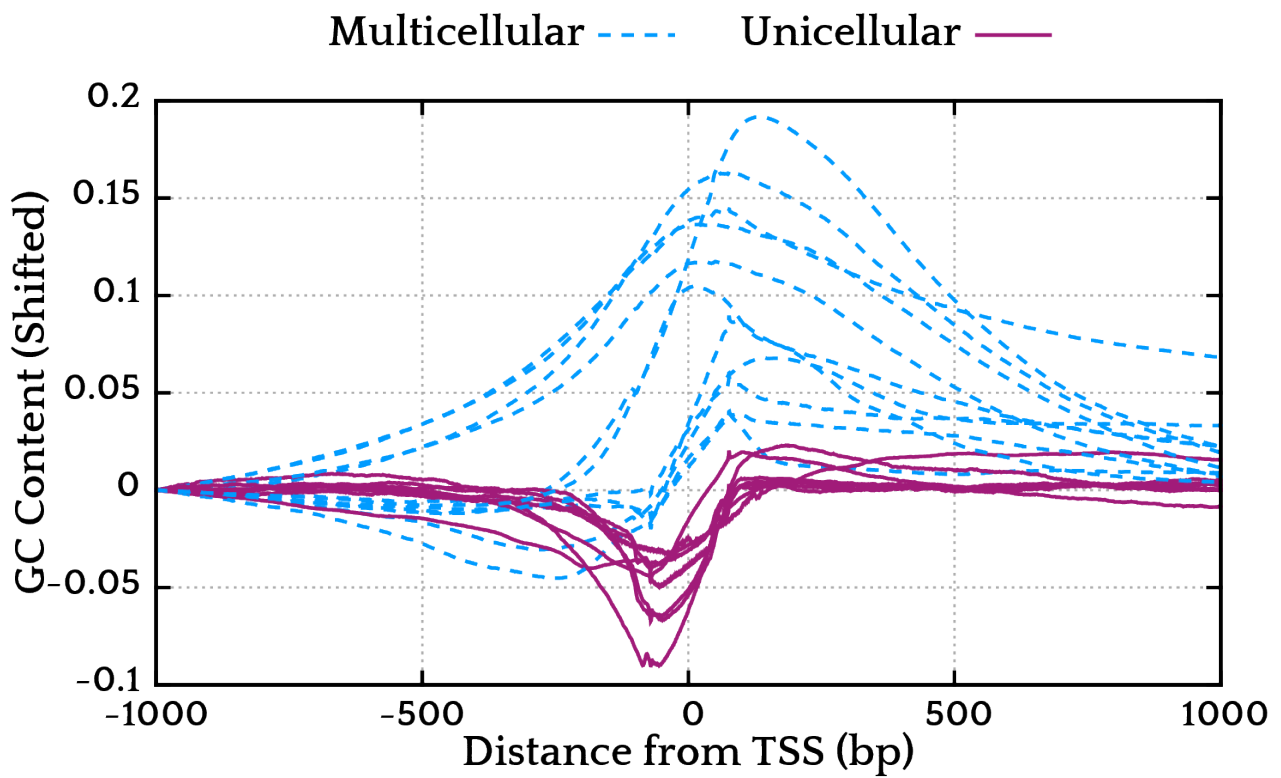


Fig. S8: Average GC content around the transcription start sites for the same organisms as presented in Fig. 2. Curves have been shifted such that the value at -1000 is zero, and have been smoothed using a 147-bp running average.

1. Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, J.-P.Z. Wang, and J. Widom. 2006. A genomic code for nucleosome positioning. *Nature*. 442: 772–8.
2. Field, Y., N. Kaplan, Y. Fondufe-Mittendorf, I.K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* 4.
3. Kaplan, N., I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, E.M. Leproust, T.R. Hughes, J.D. Lieb, J. Widom, and E. Segal. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*. 458: 362–366.
4. Olson, W.K., A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.* 95: 11163–11168.
5. Eslami-Mossallam, B., R.D. Schram, M. Tompitak, J. van Noort, and H. Schiessel. 2016. Multiplexing Genetic and Nucleosome Positioning Codes: A Computational Approach. *PLoS One*. 11: e0156905.
6. Calladine, C.R., and H.R. Drew. 1984. A base-centred explanation of the B-to-A transition in DNA. *J. Mol. Biol.* 178: 773–782.
7. Becker, N.B., L. Wolff, and R. Everaers. 2006. Indirect readout: Detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Res.* 34: 5638–5649.