Nucleosome placement and polymer mechanics explain genomic contacts on 100kbp scales

John Corrette[a], Jiachun Li[b], Hanjuan Shao[b], Praveen Krishna Veerasubramanian[b], Andrew Spakowitz[c], Timothy L. Downing[a,b,*], and Jun Allard[a,d,*]

[a]Mathematical, Computational and Systems Biology, University of California Irvine
[b]Department of Biomedical Engineering, University of California Irvine
[c]Department of Chemical Engineering, Stanford University
[d]Department of Mathematics, Department of Physics & Astronomy, University of California Irvine
[*]To whom correspondence should be addressed

Abstract

The 3d organization of the genome — in particular, which two regions of DNA are in contact with each other — plays a role in regulating gene expression. Several factors influence genome 3d organization. Nucleosomes (where $\sim$ 100 basepairs of DNA wrap around histone proteins) also bend, twist and compactify chromosomal DNA, altering its polymer mechanics. How much does the positioning of nucleosomes between gene loci influence contacts between those gene loci? And, to what extent is polymer mechanics responsible for this? To address this question, we combine a stochastic polymer mechanics model of chromosomal DNA including twists and wrapping induced by nucleosomes with two data-driven pipelines. The first estimates nucleosome positioning from ATACseq data in regions of high accessibility. Most of the genome is low-accessibility, so we combine this with a novel image analysis method that estimates the distribution of nucleosome spacing from electron microscopy data. There are no free parameters in the biophysical model. We apply this method to IL6, IL15, CXCL9, and CXCL10, inflammatory marker genes in macrophages, before and after immune stimulation, and compare the predictions with contacts measured by conformation capture experiments (4C-seq). We find that within a 500 kilo-basepairs genomic region, polymer mechanics with nucleosomes can explain 71% of close contacts. These results suggest that, while genome contacts on 100kbp-scales are multifactorial, they may be amenable to mechanistic, physical explanation. Our work also highlights the role of nucleosomes, not just at the loci of interest, but between them, and not just the total number of nucleosomes, but their specific placement. The method generalizes to other genes, and can be used to address whether a contact is under active regulation by the cell (e.g., a macrophage during inflammatory stimulation). Importantly, our findings suggest that gene function may have evolved through selective pressures that co-opted contact-mediated regulatory mechanisms reliant largely on polymer mechanics.

Pre-print server: biorxiv

Introduction

The 3d organization of the genome — in particular, which two regions of DNA are in contact with each other — plays a role in regulating gene expression [1–3]. Many complex, interdependent factors influence genome 3d organization [2, 4–9], beginning with the polymer mechanics of chromosomal DNA, and including cohesins (in conjunction with DNA binding factors such as CTCF), lamins, phase separation, and nucleosomes, to name just a few. How DNA contacts are formed and how those contacts support chromatin accessibility through nucleosome remodeling for the binding of regulatory factors remains largely unknown. The full set of biophysical factors that influence genome organization presents a formidable challenge to study. Therefore, to disentangle this complexity, we ask, how much can be explained by nucleosome positioning and polymer mechanics alone?

Nucleosomes are genomic structures comprising ~ 100 basepairs of DNA wrapped around octameric assemblies of histone proteins [7]. While one established role of nucleosomes is to directly control access to a locus [1], nucleosomes also bend, twist and compactify chromosomal DNA, altering its polymer mechanics [7]. This raises two possibilities: first, that the number of nucleosomes in between two loci could modulate their contact; Second, that specific placement of these nucleosomes, even at approximately the same total number of nucleosomes (i.e., nucleosome frequency), could induce either the formation of new contacts or the loss of existing contacts, as shown in Figure 1. Furthermore, nucleosome positions change rapidly (within minutes to hours) in response to cellular stimuli, for example during macrophage exposure to pro-inflammatory stimuli [10, 11]. We asked, to what degree are polymer mechanics and nucleosome positioning predictive of contact after a cell undergoes a functional change (e.g., inflammatory activation in macrophages)? And could these dynamics in nucleosome occupancy coordinate long-range (> 10 kilo-basepair, kbp) contacts between putative regulatory regions associated with active nearby gene promoters?

We address the first possibility by using a stochastic polymer model. Addressing the second possibility requires the polymer model be combined with knowledge of the specific placement of nucleosomes in a particular genomic region of interest. We achieve this with two data-driven pipelines. The first estimates nucleosome positioning from ATAC-seq data in regions of high accessibility [12, 13]. Most of the genomic regions are low-accessibility, so we combine this with a novel image analysis pipeline that estimates the distribution of nucleosome spacing from in vivo electron microscopy data [14]. With the aid of these two data-drive pipelines, there are no free parameters in the biophysical model.

We apply this method to gene loci for CXCL9, CXCL10, IL6 and IL15, which are inflammatory marker genes that are dynamically regulated in human macrophages following immune stimulation [15]. We compare model predictions with contacts measured experimentally by viewpoint-targeted chromatin conformation capture followed by sequencing (4C-seq). We find that within a 500 kilo-basepairs genomic region, polymer mechanics with nucleosome positions can explain 71% of close contacts ($p = 4 \times 10^{-7}$). Notably, we also find that the formation of polymer-mechanics-predicted contacts also harboring actively bound regulatory factor binding motifs likely precedes the formation of new sites of genome accessibility as captured through classical ATAC-seq experimental data. These results suggest that, while genome contacts on 100kbp-scales are multifactorial, they may be amenable to mechanistic, physical explanation.

Results

Varying nucleosome positioning within the physiologically-observed parameter range leads to order-of-magnitude changes in contact probability

We use an existing model of the polymer mechanics of chromosomal DNA [7] to ask, how much do contacts change when nucleosome positioning is varied within the physiologically-observed parameter range? This model

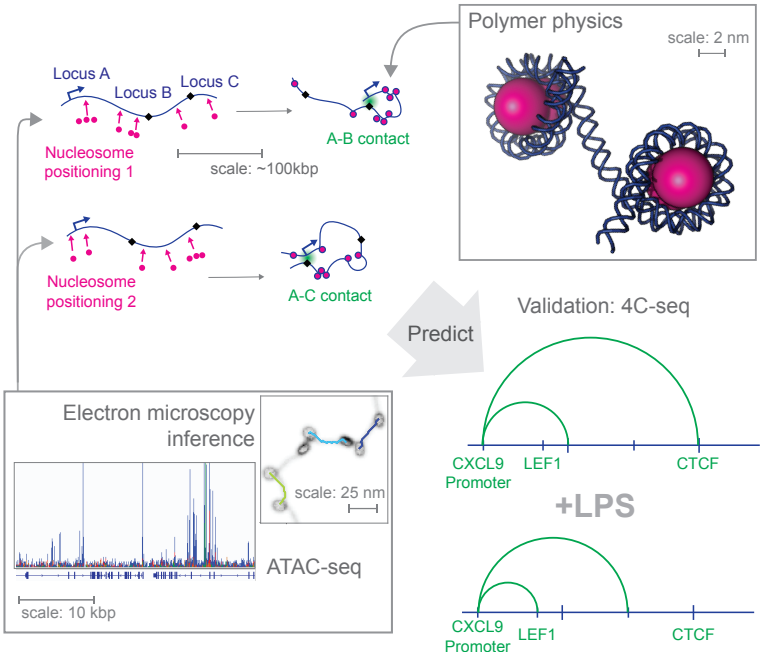

Figure 1: Nucleosomes (pink) between loci introduce "kinks" in chromosomal DNA, potentially driving contacts between distant genomic loci. We find that this is sufficient to significantly modulate the contact of genomic loci on $\sim$ 100 kilobasepair scales, and specifically that the positioning of nucleosome, even while holding constant the total number of nucleosomes, can introduce or reduce long-range contacts. Schematic of data driven mechanistic model pipeline: We combine ATAC-seq [18] and electron microscopy [14] data with a mechanistic model of the polymer physics of chromosomal DNA [7] to predict contacts between loci, and validate these predictions with 4C-seq. (Example EM image is synthetic test data.) We find that contact with some transcription factor binding motifs (e.g., LEF1) are gained and some (e.g., CTCF) are lost during inflammatory response in macrophages ("+LPS").

accounts for twisting and bending mechanics of DNA, as well as the entry and exit angles as DNA interacts with a nucleosome, taken from known molecular structure of nucleosomes from crystallography and cryo-EM [16] and mechanical properties of DNA from experiments including optical tweezer experiments [17] (see Beltran et al. [7] for full details).

Here, we repurpose this model to look at contact probabilities as a function of distance between two loci. For convenience, we refer to one locus as the locus of interest. Typically, we choose the transcriptional start site (TSS) or a nearby locus. We define the contact probability as the proportion of configurations with the two loci within 10nm. As shown in Supplemental Figure S1, our results are not sensitive to this choice.

Before exploring contact probabilities with nucleosome positions extracted directly from experimental data (next section), here in Figure 2, we perform an *ab initio* study with simplified assumptions for nucleosome positions. In this work, we specify nucleosome positions by specifying the sequence of nucleosome spacings, also called linker lengths, which is the number of basepairs between adjacent nucleosomes, and which does not include 147 basepairs assumed to be occupied by the nucleosome itself. We consider three scenarios: First, we assume nucleosomes are evenly spaced (A, D). Second, we assume nucleosomes are randomly spaced according to a uniform random distribution with a fixed mean, and a range of $\pm$10bp (B,E) following Beltran et al. [7]. Finally, we assume nucleosomes are spaced according to an exponential distribution (C, F). This exponential distribution could arise from a spatially-independent (Poisson) process, or the equilibrium of a 1-dimensional lattice gas.

In the simulations where there is no random variation in nucleosome spacing (Figure 2A), contact probability is largely controlled by the twist persistence length of DNA. This contact probability increases to a local maximum as the nucleosome spacing increases every 10.5 bp.

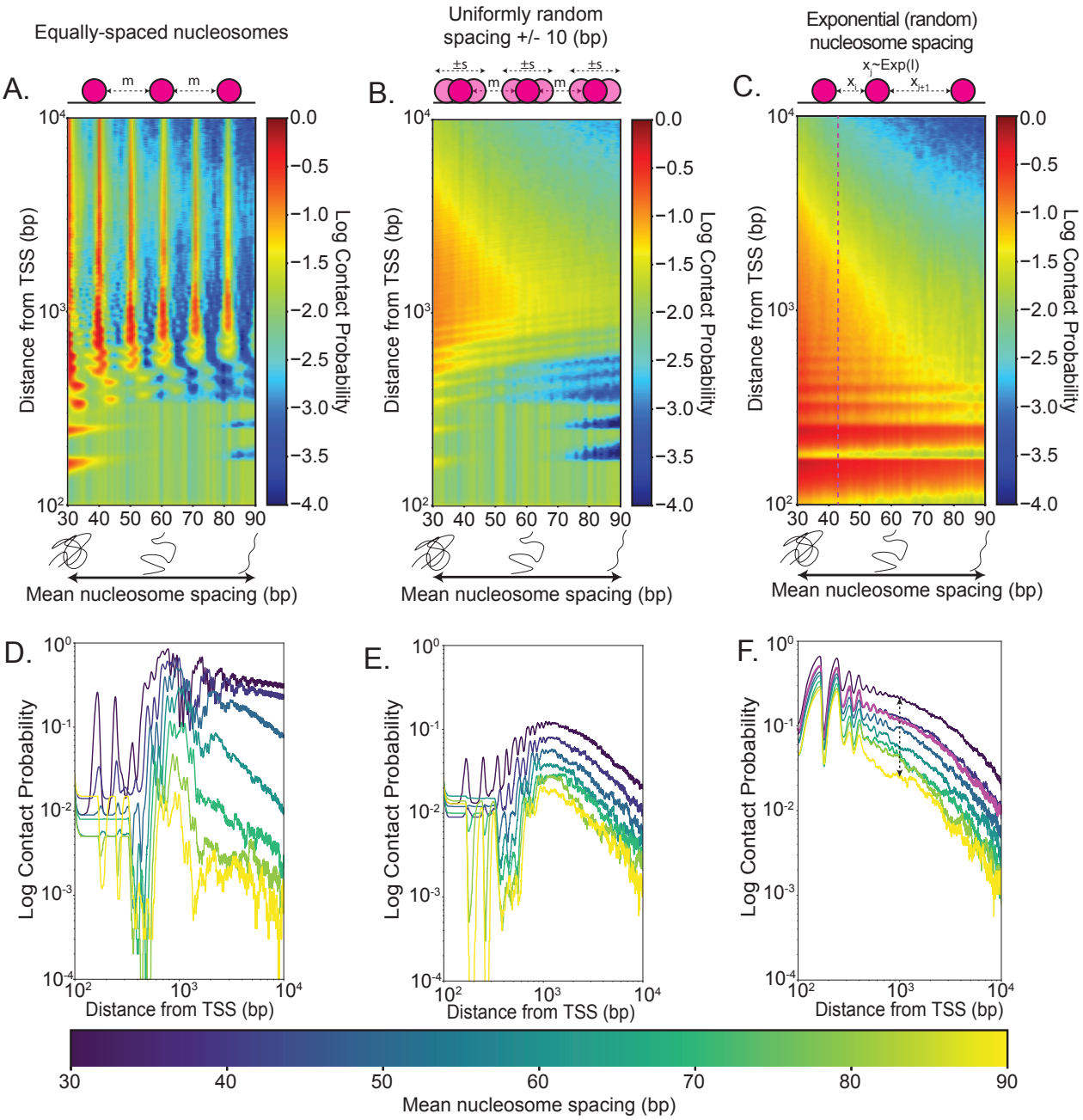

Figure 2: DNA polymer contact probability of different nucleosome spacing (i.e., linker length) models, defined as the proportion of configurations with the two loci within 10nm, shown as heatmaps (A-C) and families of curves (D-F). **A,D** Evenly-spaced nucleosomes with no variation result in a repeated pattern of high contact (red ridges) every 10.5 bp, consistent with the twist persistence of DNA. **B.** Introducing variation into the uniform model results in the loss of the 10.5 bp spaced ridges, and reveals larger regions of increased and decreased contact probability. **C.** Exponentially-distributed nucleosome spacings (i.e., a spatial Poisson process or lattice gas) result in a smoother contact probability profile. Contact probabilities are highly sensitive to changes in nucleosome spacing, for example changing from 30 bp to 90 bp results in a 88% decrease in contact probability 1 kbp away from the TSS in the Poisson process contact probability model (F, dashed black arrows). The magenta dashed vertical line in C and curve in F indicate the best-fit estimate from our analysis of electron microscopy data (Figure 3).

Next, we allow variation where nucleosome spacing is distributed uniformly with a mean ± 10 basepairs (Figure 2B). Here, the local maxima that arose from the twist persistence are "averaged-out", resulting in a more smoothed-

out contact probability profile. We see a local maximum $\sim$ 1 kbp. This maximum occurs when the distance is approximately equal to the persistence length, at which the chromosomal DNA can loop back on itself most easily.

Finally, we allow variation where nucleosome spacing is distributed exponentially (Figure 2C). Many of the qualitative features of the previous nucleosome placement models are preserved, but with a larger region of increased contact around the locus of interest.

Taken together, these results demonstrate that contact probabilities are highly sensitive to changes in nucleosome spacing. For example, changing from 30 bp to 90 bp results in a decrease in contact probability 1 kbp away from the TSS of 88%, almost ten-fold (Figure 2F, dashed black arrows).

Analysis of existing electron microscopy images allows learning nucleosome spacing in low-ATAC-coverage regions

To put the previous results into the context of a specific gene, in a cell undergoing regulated changes in gene expression, we require a method to extract nucleosome positions from data. This is possible using the DNA accessibility sequencing method ATAC-seq [18] and algorithms such as NucleoATAC [12] or NucPosSimulator [13]. However, we found that most of the genomic landscape (e.g. $\approx 98\%$ for within ± 200 kbp of IL6 TSS) has insufficient ATAC-seq data coverage, defined as any region that of the ATAC-seq data that is not identified as a peak region by MACS2 peak calling [39]. Therefore, for these low-ATAC-coverage regions, we develop a nucleosome spacing learning algorithm that uses electron microscopy (EM) images of chromosomal DNA, combined with existing EM images [14].

The full method is described in the Methods. Briefly, we first binarize this data to a threshold such that only the dark core of the nucleosome remains, and perform 3D connected components analysis to identify each distinct candidate nucleosome. We then apply a novel iterative pathfinding algorithm on the original, unbinarized image. This algorithm weighs all paths by both their path-length and their intensity under EM (where darker, i.e., denser regions are weighted less), and returns the minimal path between candidate nucleosomes identified from the binarized image. This proved advantageous: A method that simply computes shortest paths between detected nucleosomes would ignore DNA curvature, while a method that only considers EM intensity fails on many nucleosome pairs where the path is disjoint (due to imaging noise).

We test the method on synthetic data with a known ground-truth distribution of nucleosome spacings (Figure 3A-D), finding that the learned distribution is similar to the ground truth in terms of its mean and interquartile range, with larger tails in the learned distribution (Figure 3C,D). Testing on synthetic data is further elaborated in Methods and Figures S2-S6.

We then apply the method to EM images from Ou et al. [14], taken from heterochromatin in U2OS cells. We use a region of $328 \times 328 \times 119.72$ nm, resulting in ≈ 1400 nucleosomes (Figure 3E). We find a nucleosome spacing distribution that is well-approximated as exponential, with a mean of ≈ 43 bp, as shown in Figure 3G and indicated in Figure 2C,F. This learned distribution of exponential with mean of 43 bp is in the range of values predicted and/or assumed in previous work [8, 19, 20].

To validate the learned nucleosome spacing distribution, we simulated the polymer mechanics model at the learned distribution and parameters. We see small local clusters of nucleosomes forming (Figure 3I), similar to the observations made in Ou et al. [14]. We quantitatively validate by repeating the erosion analysis in Ou et al. [14] on both the real data (Figure 3H) and synthetic data (Figure 3J). We find that both linear fits to the erosion curve share similar axis intercepts.

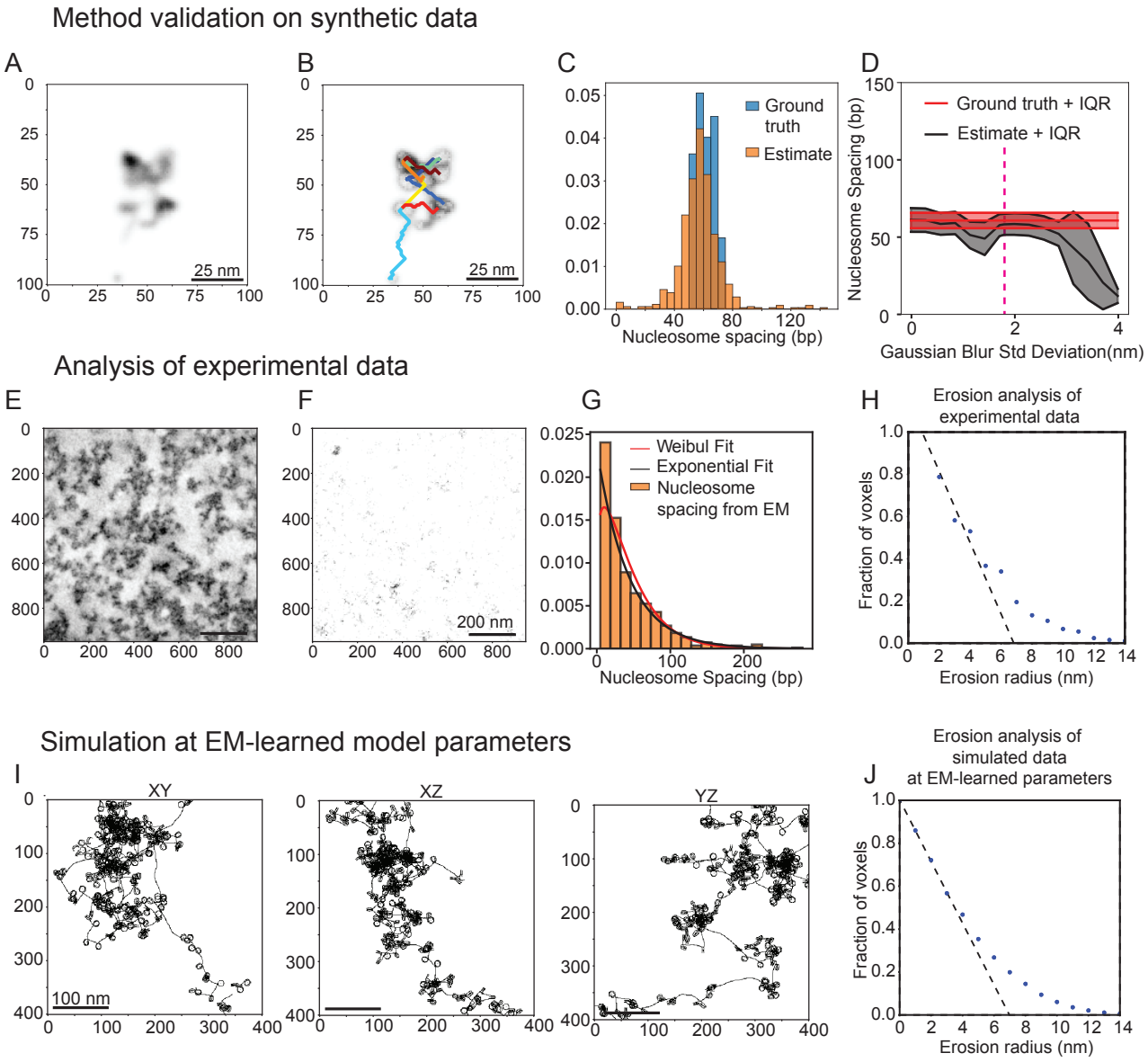

Figure 3: Learning nucleosome spacing distribution from electron microscopy data. **A-D** Validation of image analysis pipeline on synthetic data. **A.** Sample conformation of synthetic DNA with basepair spacing distributed uniform randomly with mean 60 bp $\pm$20bp, shown as a synthetic EM image with varying voxel intensity. **B.** Sample nucleosome spacing paths estimated by novel Dijkstra's algorithm pipeline, overlayed on image after estimated nucleosome cores are subtracted (see Methods). **C.** Comparison of ground truth distribution of nucleosome spacing and learned distribution for 100 random synthetic EM images. **D.** The learned nucleosome spacing, summarized by the mean and inter-quartile range (IQR), for various Gaussian blurs. Pink vertical dashed line indicates the Gaussian blur estimated from the data of Ou et al. [14]. Vertical magenta line indicates the estimated noise level that is seen in real ChromEMT data. **E.** Sample slice of 3D EM microscopy data from Ou et al. [14], **F.** Same as E, after binarization for analysis. **G.** Learned distribution of nucleosome spacings. Data fits an exponential distribution with a mean of 43 bp. We also fit a Weibull distribution to the data, where we find our scale parameter to be 1.196, which indicates that our distribution is close to exponential. **H.** Erosion analysis on EM data, following method described by Ou et al. [14] **I.** Sample conformation of synthetic DNA using our learned nucleosome spacing distribution. Data exhibits similar qualitative "clumping" of chromosomal DNA that was identified in Ou et al. [14]. **J.** Erosion analysis, same as H, on synthetic data.

An ATAC/EM-informed polymer model to understand the effects of nucleosome positioning and polymer mechanics on long-range contact probabilities and loci of high-contact for specific genes

The above method allows the study of nucleosome-driven contacts for a specific gene. We first apply the method to CXCL9, a gene that is upregulated in response to immune stimulation [15]. The full pipeline for simulation uses ATAC-seq to place nucleosomes on regions of high-ATAC-coverage with the NucleoATAC algorithm [12], and then uses the EM-informed model presented above for regions of low-ATAC-coverage, resulting in a complete map of nucleosome placements around the CXCL9 TSS. Our algorithm for sampling nucleosome placements from NucleoATAC in regions of high-ATAC-coverage is described in Methods. In Supplemental Figures S7-S9 we show that our nucleosome placement algorithm is consistent with the NucleoATAC data. We can then perform the same stochastic polymer mechanics calculation as in Figure 2 but with data-informed nucleosome positions. Using published ATAC-seq data from Zhang et al. [11], we perform this simulation for ATAC-seq data from both unstimulated macrophages, and from macrophages stimulated with bacterial lipopolysaccharide (LPS) for 1 hour. Crucially, there are no free parameters in the biophysical model.

The configurations, as viewed by radius of gyration and root-mean-square distances, show discrepancy from the 0.5 power-law predicted by the simplistic Gaussian chain assumption (Supplemental Figure S10). We plot the data-informed contact probabilities in Figure 4C. For comparison, we overplot the contact probabilities predicted for exponentially-distributed nucleosome spacing distribution estimated from EM data alone (so, same as Figure 2F), which approximately obeys Gaussian chain statistics. The discrepancies between these two nucleosome placement assumptions demonstrates that real nucleosome placement can indeed significantly modulate contact, both quantitatively and in introducing new qualitative features, e.g., a sudden reduction in contact around -25kbp.

The raw contact probabilities show significant differences between the contact probabilities before and after LPS stimulation (blue and red curves in Figure 4C). One example locus is shown in contact with the transcriptional start site in Figure 4D. This contact probability quadruples upon LPS-stimulation, suggesting that nucleosome placement alone is sufficient to significantly alter contact probabilities. Note that this is not simply a consequence of increasing or decreasing the number of nucleosomes, since the total number of nucleosomes varies less than 2% between the two conditions.

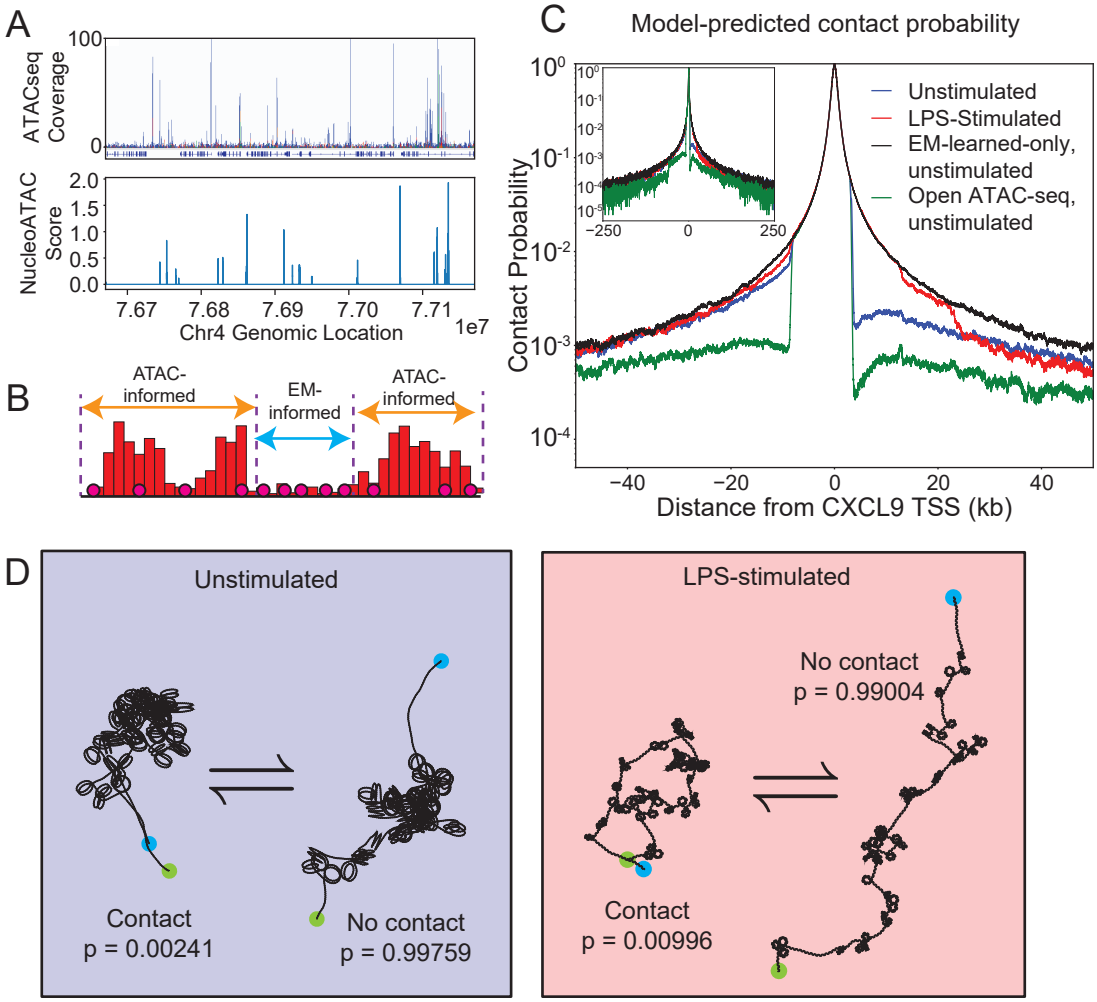

Figure 4: Simulation of polymer mechanics of CXCL9 with nucleosomes placed according to ATAC-seq and EM data. **A.** Example ATAC-seq coverage (top) and NucleoATAC output probability density function (bottom) centered at CXCL9 TSS ± 250 kbp. Peaks correspond with a higher probability of nucleosome presence in transcriptionally active DNA. ATAC-seq and NucleoATAC maps are sparse and require EM data to fill in low coverage regions. **B.** The MACS2-identified high-coverage regions are analyzed using NucleoATAC [12], which gives a probability distribution of nucleosome center positions in all peak regions (orange). We sample these regions for nucleosomes, then fill the remaining EM-informed regions (blue) with nucleosomes according to the learned distribution (described in Figure 3). This list of nucleosome spacings for the genomic region of interest of DNA is provided to the polymer mechanics model [7], and we perform Monte Carlo sample of the 3D conformation of DNA in thermal equilibrium. **C.** Contact probability profiles for unstimulated (pre-LPS; blue) and 1-hour post-LPS stimulated macrophages (red) within a ± 50kbp window around the transcriptional start site of CXCL9. EM-only (black) and open ATAC-seq (green) is shown for comparison. Open ATAC-seq corresponds to not placing any nucleosomes in MACS2-identified high-coverage regions as seen in B. The EM-only probability profile approximately behaves as a Gaussian chain. Inset is the same data but plotted on a ± 250kbp window to demonstrate the genomic domain of simulation. **D.** Sample conformations of CXCL9 before and after LPS-stimulation. One example locus is shown in contact with the transcriptional start site; this contact probability quadruples upon LPS-stimulation, suggesting that nucleosome placement alone is sufficient to significantly alter contact probabilities.

The ATAC/EM-informed polymer model is predictive of long range genomic contacts observed by 4C-seq

We next sought to identify specific loci that we could categorize as in "close contact" with the locus of interest, namely the CXCL9 transcriptional start site. We perform background removal, smoothing, and normalization to produce a contact signal (see Methods and Supplemental Figure S11). We then use wavelet peak finding [21] to identify "peaks" in the contact signal, which we then refer to as predicted close contacts.

Note that while the biophysical model that predicted the contact probabilities shown in Figure 4C has no free parameters, there are 2 hyperparameters in the wavelet peak finding algorithm. To avoid overfitting, we optimize these hyperparameters on 4C data before LPS stimulation (Figure 5ACE, "Unstimulated", which serves as the training data set of LPS-treated human macrophages at zero hours), and then use the same peak-finding hyperparameters on one hour post-LPS-treated cells (Figure 5BDF, "Stimulated", which therefore serves as the validation data set).

A straightforward way to assess the predictive power of this model is via two quantities: the proportion of polymer-mechanics-predicted contacts that are within a tolerance distance of a 4C contact (that is, the accuracy of the model); and the proportion of 4C contacts that are within a tolerance distance of a polymer-mechanics-predicted contact (that is, the precision of the model, specifically $1 -$ FDR where FDR is the false discovery rate). These are shown in Figure 5EF for various tolerance distances. At high tolerance, prediction becomes perfect, as expected. At the genomic distance of 10 kbp, roughly the size of larger gene regulatory elements [22–24], the accuracy is 0.83 and precision is 0.83. This value has a p-value of 0.0001 compared to randomly guessing the locations of the same number of peaks.

This striking result indicates that polymer mechanics and nucleosome placement, alone, can account for a statistically significant proportion of contacts between two loci. Furthermore, we performed a similar analysis using one 4C replicate to predict another 4C replicate, and found that the average predictive power of 4C is 0.63 (Figure 6B and Supplemental Figure S12 and S13), which is below the predictive power of our model. This suggests that the predictive test of our model is not limited by the model itself, but the variability of the 4C data.

Note that we report a p-value in Figure. 5C, but this should not be interpreted as a hypothesis test since it is part of the training set. Note also that we use a false discovery rate instead of a false positive rate, since the latter is less meaningful for identifying ~ 10 contacts on a region of ~ 100 kbps.

Beyond CXCL9, we next sought to study other inflammatory genes, IL6, IL15 and CXCL10. Crucially, here we do not retrain the peak-finding hyperparameters (and recall there are no free parameters in the biophysical model), reusing those we found for CXCL9. The resulting tracks are shown in Figure 6A. We find that for all genes together, our model yields a metric of 0.71, corresponding to a p-value 4.3076×10^{-7}, and comparable with that average predictive power of 4C as seen in Figure 6B (red) and Supplemental Figures S12 and S13.

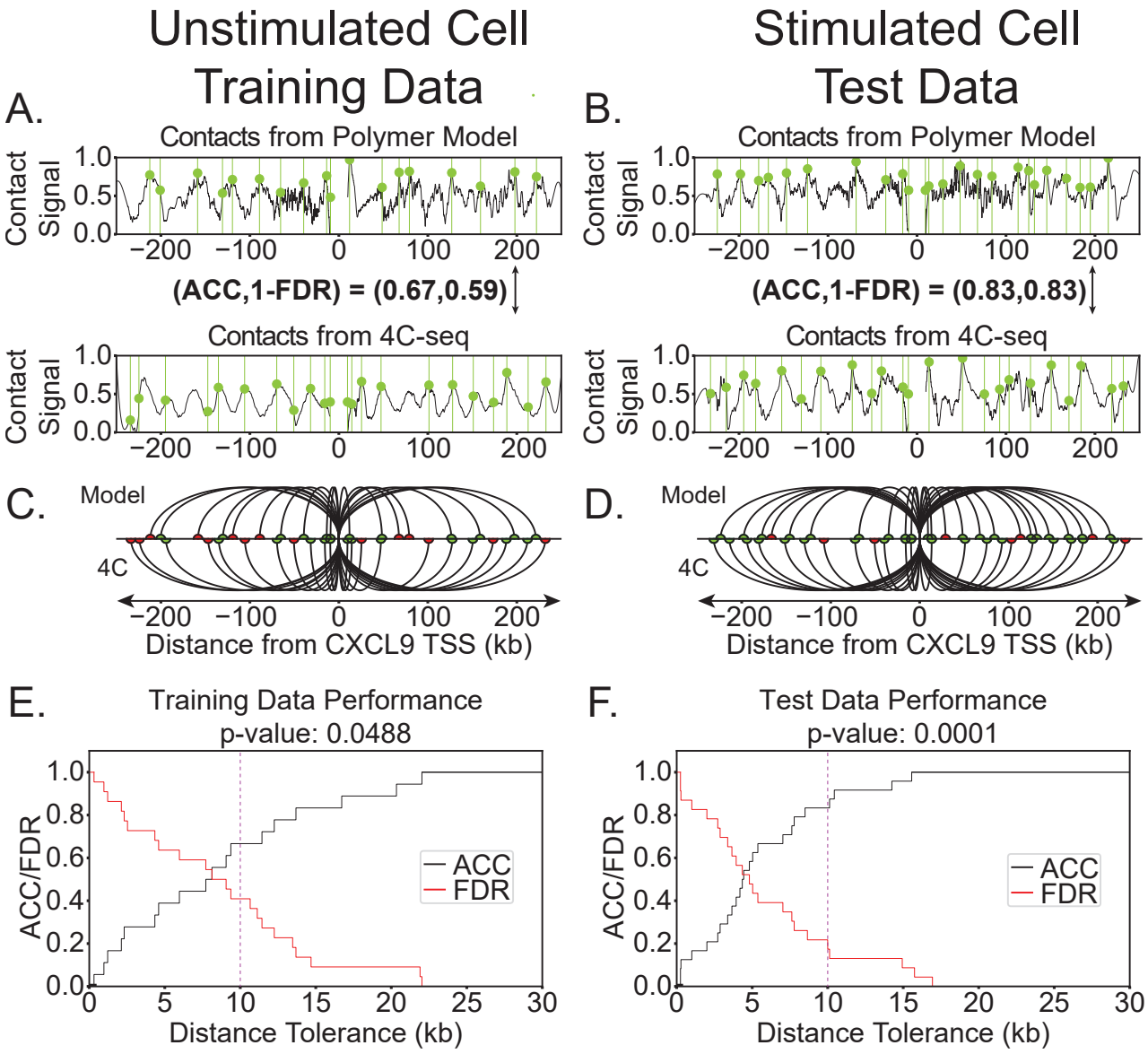

Figure 5: ATAC-EM-informed polymer model predicts close contacts observed in 4C-seq. **A-B.** Contact signals for unstimulated, pre-LPS human macrophage (left) and post-LPS 1hr stimulated macrophage (left). Tracks from the ATAC-EM-informed model (top) and 4C (bottom) are shown with peaks overlayed in green. **C-D.** Predicted contact agreement between our model and 4C. Arcs represent and identified contact. Semicircles represent a 10kb contact, and are colored green if both 4C and the model are in agreement, and are colored red otherwise. **E.** Training set performance. The pre-LPS dataset was used only to train the peak-finding algorithm (all biophysical parameters were fixed). Vertical dashed line shows where accuracy and precision were measured for hyperparameter optimization. (The p-value 0.0488 is shown but should not be interpreted as a hypothesis test, since it is on the training set.) **F.** Validation set performance. Within a threshold of 10 kbp, the model predicts peaks with an accuracy and a precision (1-FDR) of 0.83 (vertical dashed line). This corresponds to a p-value ≈ 0.0001 compared to randomly guessing the locations of the same number of peaks.

The range of validity of nucleosome-polymer-mechanics modeling

As mentioned, many other factors influence genome organization besides polymer mechanics and kinks due to nucleosome placement. The scaling behavior of contacts exhibits different regimes at different genomic length scales [6, 25]. From these previous studies, we speculate that the simple model developed here would lose validity around 400kbp, where other effects such as fractal globule effects, and topological domains organized by regulatory complexes such as cohesins, may dominate. Therefore, we repeated the analysis, including polymer simulation, for a range of genomic domain sizes, and show the metric in Figure 6B, for all 4 genes and 2 stimulation conditions. We compare this to the ability of one 4C replicate to predict another 4C replicate (red curve). As mentioned above, at this level of predictive performance, we expect the polymer model to be limited by the reproducibility of the 4C data. We also compare to the average predictive power of randomly placed contacts (dashed line; at this predictive performance, the p-value would be 0.5). We find that the model performs significantly better than random guessing, and comparable to the limitation of the 4C data. Furthermore, this appears to hold throughout a range of genomic distances, to 250kbp, suggesting that the nucleosomes continue to have a significant impact up to 250kbp. Computational limitations prevented us from simulating larger genomic distances.

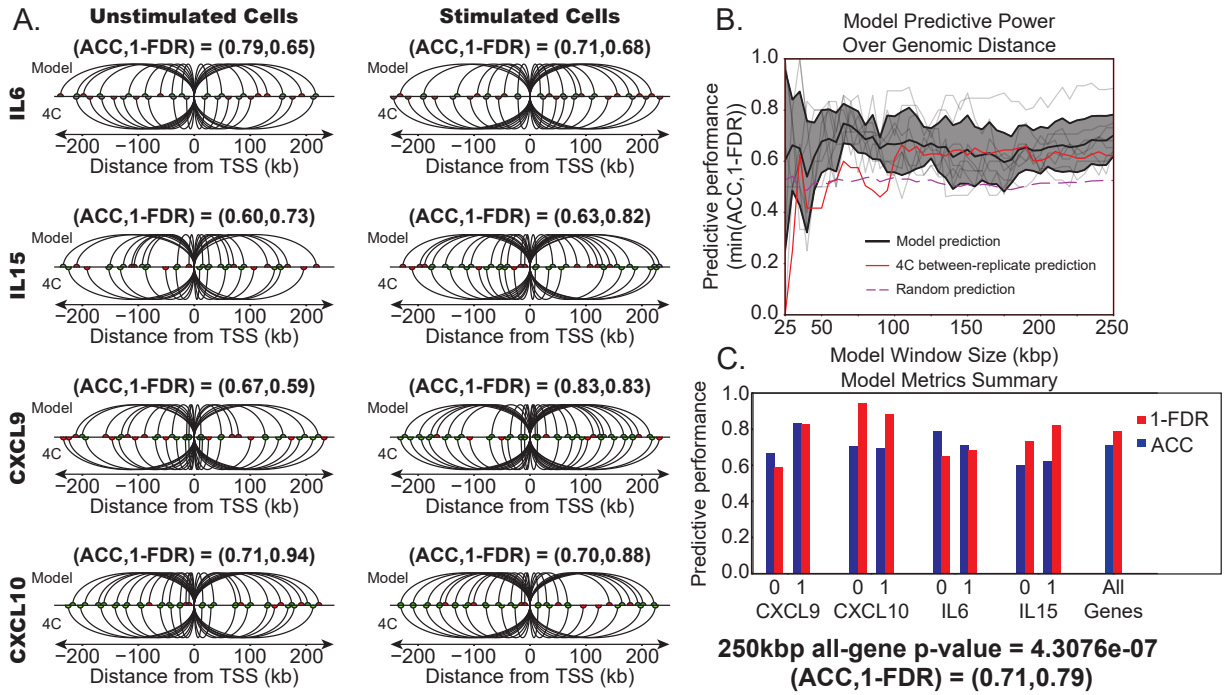

Figure 6: ATAC-seq is predictive of genomic contact across multiple genes and genomic length scales. **A.** Predicted contacts for 4C and our polymer model for IL6, IL15, CXCL9, and CXCL10. All contacts were predicted using CXCL9 hyperparameters as described in Figure 5. Contact signals, such as those in 5AB are plotted for all four genes in Figure S14. **B.** Performance over genomic distance for all genes. Gray band about model prediction depicts one standard deviation from the average prediction metric. Average 4C metric and random guessing metric are shown for reference. Model overall shows little trend over genomic distance, we find a slight downtrend for small window sizes. At large genomic distance, model performs significantly better than the random predictor with a p-value 4.3076×10^{-7}. **C.** Summary of accuracy and precision metrics for each gene, for both unstimulated (0 hour) and stimulated (1 hour) human macrophages, and the combination of all four genes being predicted simultaneously.

Polymer-mechanics-predicted contacts harbor actively engaged regulatory factor DNA-binding motifs and show dynamic coordination with genome accessibility

Given that long-range genomic contacts are widely thought to contribute to active gene regulation through the co-localization of regulatory factor binding hubs with distantly located gene promoters, whereby this co-recruitment can, e.g., boost or "enhance" expression, we next examined the frequency of regulatory binding factor motifs located within said contacts. To do so, we devised a strategy to identify genomic regions that overlapped as both experimentally observed (based on 4C-seq conserved across two biological replicates) and polymer-mechanics-predicted contacts across unstimulated (0hr) and stimulated (1hr) conditions. Using this strategy, we identified genomic contact regions that were either dynamically regulated (gained or lost) or static during macrophage stimulation (Figure 7A). We next used the MotifMap database to quantify the abundances of transcription factor (TF) binding motifs present within our genomic contact regions and found that these regions harbored a total of 243 motifs spanning 40 distinct transcription factors, with the top 3 most abundant being LEF1, YY1, and IRF8. Notably, LEF1 (lymphoid enhancer binding factor 1) was the most abundant motif within our contacts and is known to regulate gene expression through DNA looping mechanisms [26] and its role in 3d genome organization has particularly been implicated in immune regulation of myeloid lineages [27]. We also observed overlap with the binding motif of CCCTC-binding factor CTCF, an architectural protein known to mediate intrachromosomal interactions [28] within a contact that was lost upon inflammatory stimulation.

To better understand whether contacts were related to nucleosome remodeling events that occurred directly at the site of contacts or at locations between contacts, we partitioned contacts into four categories according to the status of ATAC-seq peak behavior that they were associated (directly overlapped) with (Figure 7B). Thus we annotated our experimentally confirmed, polymer-mechanics-predicted contacts that were gained, lost, or remained static during 1hr stimulation (green, red, and blue, respectively), and annotated these by whether they overlap with ATAC-seq peaks that were gained, lost, and/or static or that showed no peak during 1hr stimulation. At this resolution, we found that of the 8 contact regions captured within the CXCL9 locus, 1 region lacked overlap with any ATAC peaks (no peak), 5 regions overlapped with regions harboring static ATAC peaks, 1 region showed overlap with ATAC peaks that were lost during stimulation, and 1 region overlapped with ATAC peaks that were both gained and lost during stimulation. The one contact region associated with ATAC peaks that were both gained and lost during stimulation was associated with 1 of 3 static contacts within the locus. The other 2 static contacts overlapped with ATAC peaks that were static. We also observed that all 3 of the gained contacts experienced scenarios where ATAC peaks were both static as well as completely absent. These findings suggest that contacts change largely independently from the observed ATAC behavior, where all three categories of contact dynamics (gained, lost, static) were observed, and that contact formation could occur independent from (or is not preceded by) ATAC peak formation. While only 2 out of the 8 contact region experienced a loss in ATAC peaks, we found that 1 of these contacts overlapped with 1 of the 2 contact regions that were also lost during stimulation. These findings leave the possibility that the stability of some ATAC peaks may be contact-dependent. However, additional analysis of multiple gene loci are needed to broaden the interpretation of these findings.

Finally, to further investigate the biological significance of our motif analysis, we compared the genomic positions of the motifs contained within our contact regions with publicly available data from genome-wide TF-binding assays (ChIP-seq) for TFs LEF1 and CTCF (Figure 7C). Indeed we find that several LEF1 motifs and the CTCF motif captured within our contact regions are occupied by actively bound transcription factor proteins directly in macrophages for CTCF (GSM2544247) or cells of closely related lineage origin to our THP-1 cell line (specifically K562 lymphoblast cells) for LEF1 (GSM2828684).

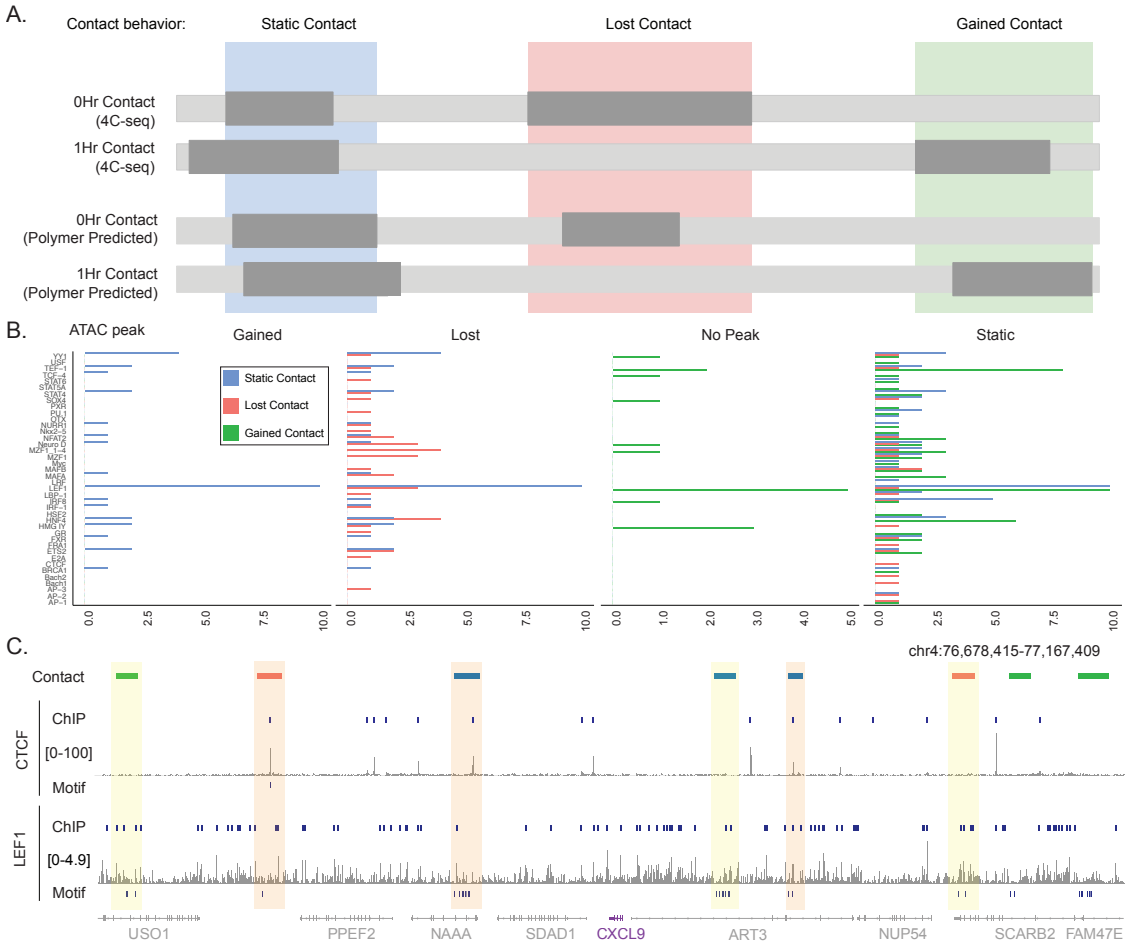

Figure 7: Polymer-mechanics-predicted contacts harbor actively engaged regulatory factor DNA-binding motifs and show dynamic coordination with genome accessibility peaks. **A.** Strategy to identify polymer-mechanics-predicted and experimentally validated regions of viewpoint contacts at the CXCL9 locus. **B.** The frequency of TF binding motifs located within genomic contacts defined in (A) according to contact dynamics (gained: green; lost: red; static: blue) and divided according to the behavior of ATAC-seq peaks observed during 1hr stimulation (gained, lost, static, or no peak) overlapping with each respective contact region. **C.** Genome browser track showing the CXCL9 locus indicating genomic contacts identified in (A) along with overlapping TF-binding motif for CTCF and LEF1 and the respective ChIP-seq tracks for CTCF binding within macrophages and LEF1 binding within blood lineage-derived K562 lymphoblast cells. Shaded regions indicate contact regions where CTCF and LEF1 ChIP-seq peaks coexist (orange) or individually (yellow).

Discussion

Besides their role at individual genomic loci, nucleosomes between loci can significantly modulate their long-range contact. Changing nucleosome density within the physiologically-observed range can lead to order-of-magnitude changes in contact probability. More surprisingly, we find that the specific placement of these nucleosomes between loci, even at approximately the same total number, can introduce or reduce contacts.

The specific contacts predicted by the polymer model, after training on ATAC-seq and electron microscopy data, were statistically significant and on-par with the difference between replicates of 4C-seq. Indeed, 71% of close contacts within four 500kbp genomic regions, before and after inflammation, were correctly predicted by the model. This is surprising, since many other factors are known to influence 3d genome organization, but is consistent with prior work suggesting that this change is stronger than that due to the presence or absence of cohesins [8], and that cohesins are not always necessary to predict structure [29]. These results suggest that, while genome contacts on 100kbp-scales are multifactorial, they may be amenable to mechanistic, physical explanation.

The image analysis and pathfinding algorithm developed herein was necessary, since ATAC-seq coverage is low in most of the genome. Our method, which uses a novel pathfinding algorithm that weighs paths by both their intensity under EM and their path-length, yielded a distribution of nucleosome spacings consistent with previous work [8, 19, 20]. It would be intriguing to apply this method to EM data on specific regions of the nucleus, and to cells in different regulatory states.

This model assumed that causality runs from nucleosome placement, to polymer mechanics, to conformation and close contacts. While the predictive success of this model demonstrates this is sufficient to explain a significant fraction of contacts, other mechanisms are possible. For example, it is possible that 3d conformation influences nucleosome placement, since histones may bind more easily to certain DNA conformations. Moreover, the purely causal, mechanistic assumptions of the model may also be a detriment if the goal were singularly to predict 4C data from ATAC data — something that could be better achieved by incorporating more information or removing physical constraints implied by the model.

We emphasize that the granularity of this polymer mechanics model is optimal for this 100-kilo-basepairs-scale inquiry. The configurations are computed using efficient Monte Carlo sampling where analytical expressions [30, 31] are used for the mechanics of unbound DNA, greatly increasing the efficiency of the code. Models with more atomistic detail would be computationally prohibitive, and models on larger genomics scales would miss the effect of individual nucleosomes. This mesoscale approach, and the specific model and pipeline presented here, generalize to other genes, and can be used to address other cases of cells undergoing a state change.

Throughout evolution, the sequence structure of the genome is undoubtably shaped through selective pressures associated with physical polymeric properties of the genome [32]. However the intricacies of this co-evolution of sequence and physical property remain largely undefined. Notably, there are long-standing observations that TF binding sites actually occur in relatively high-density clusters (or hotspots) throughout the genome [33, 34]. Furthermore, the positioning of nucleosomes is heavily guided by genetic sequence yet still subject to precise nucleosome eviction events dynamically regulated according to cell state [35]. Our findings present a new paradigm regarding how evolutionary pressures contribute to selection of genetic architectures that favor contact- (or loop-) formation between “motif-rich” regulatory binding site hotspots with gene promoters through sites of conditional nucleosome placement or polymer mechanics, which may assist in shaping the motif environment to influence transcription factor search dynamics [36].

Methods

Computational simulation of contacts from the twistable worm-like model

We use the twistable worm-like chain model with nucleosomes, as described previously in Beltran et al. [7] and `github.com/ajspakow/wlcstat`. This model requires as input the nucleosome positions.

For the uniform model, we take a mean spacing μ and a spacing variability σ, and place a nucleosome every $\mu + X$ bp, where $X \sim U(-\sigma, \sigma)$ and U is a uniform random distribution. For the Poisson process, we randomly place nucleosomes every Y basepairs where $Y \sim \text{Exp}(\lambda)$, where our exponential distribution is defined such that the expected value $E[Y] = \lambda$.

Once we randomly sample nucleosomes, we compute the lengths of exposed DNA between nucleosome to generate a list of nucleosome spacings. This list of nucleosome spacings is then provided as input to a chain growth algorithm to compute the 3D conformation of DNA, with nucleosome wrapping, under random thermal fluctuations.

For each conformation, we count each basepair within 10 nm of the locus of interest (usually the TSS). We repeat this a minimum of 150,000 randomly generated polymer conformations and average these scores together. The computation takes approximately 2 hours on an Intel Xeon Gold 6336Y 24-core CPU at 2.4GHz.

Nucleosome spacing from electron microscopy data

We develop an algorithm for learning nucleosome spacings from EM microscopy data of heterochromatin, specifically here using data from Ou et al. [14]. We first determine the centroids of the nucleosomes in the EM data. To do so, we binarize the 3D EM image with a noise threshold of 95% of the maximum voxel value. We find that naked DNA is fully eroded and the nucleosomes cores remain intact in the image. We then use the `cc3d` python package and the function `connected_components` to compute the unique connected 3D components in the image to cluster and compute the centroids of each disjointed candidate nucleosome. With a list of centroids in hand, we return to our non-EM image and use Dijkstra's algorithm [37] to compute nearest neighbor distances between all unique centroids. A centroid included in a previous pair is excluded from future pairs. The Dijskstra algorithm weighs all paths by both their intensity under EM and their path-length, and returns the minimal path. We found that this method is the most robust: A method that simply computes shortest paths between detected nucleosomes would ignore DNA curvature, while a method that only considers EM intensity fails on many nucleosome pairs where the path is disjoint.

To validate our method, we simulate DNA in 100 randomly generated 100 kbp windows via our ATAC-free algorithm with a uniform nucleosome spacing of 30, 60, and 90 ± 20 bp (slightly wider distributions than those presented in Figure 2A). We then apply binary dilation with a diameter of 2.5 nm to each 1 cubic nm voxel containing a bp long segment of the polymer, as well as a 5.5 nm dilation to the known positions of the nucleosome cores, so as to match know diameter of DNA and half the diameter of nucleosomes [16, 38]. We seek to apply the same noise and uncertainty to the synthetic data as in the real data, which we estimate has a Gaussian blur (point-spead function) width of 1.8nm, and a contrast ratio of 2.5 between nucleosomes and DNA (see Figure S2). We apply Poisson random noise to the synthetic EM image and Gaussian noise ranging from a standard deviation of 0.0 nm to 4.0 nm. We then perform the nucleosome spacing learning algorithm as described above, visually described in Figures S3, S4, and S5. Finally, we compare the resultant learned nucleosome spacing distribution to the simulated ground truth in Figure 3 and Figure S6. These tests demonstrate that the method is able to correctly learn the mean nucleosome spacing for the noise parameters identified from the real data.

Nucleosome Spacing from combined ATAC-seq and EM-informed model

For our polymer model of chromosomal DNA, we wish to inform our nucleosome placement from ATAC-seq data. First, we detect peak regions in the ATAC-seq data using MACS2 [39] with a window size of 1 kbp. This is to ensure that the regions in ATAC-seq that we determine nucleosome positioning from are of sufficient sequencing depth. We refer to these as high-ATAC-coverage regions. We then use NucleoATAC [12] to determine the probability distribution of nucleosome placement. Briefly, NucleoATAC uses a model fit on *S. cerivisiae* to create estimates of the positions of nucleosomes along the genomic body using ATAC-seq fragment size and locations distributions as input. This model works by running a V-plot, a learned distribution of fragment size relative to the position of a dyad, across the genome, and cross correlates with the genomic spatial fragment size distribution. Thus the output of this model can be interpreted as a probability density function for the locations of nucleosomes across the genome. A higher cross-correlation score coincides with a higher probability with a nucleosome being centered at that genomic position. We verify that our algorithm is behaving as stated by examining the nucleosome densities compared to the NucleoATAC score in Figure S7, as well as verify that the distribution of local maxima in NucleoATAC signal are similar to nucleosome placement probabilities in Figures S8 and S9. Once we have the NucleoATAC probability distribution for a given region of DNA that is high-ATAC-coverage, we sequentially sample the probability distribution for nucleosome center positions until we can no longer place a nucleosome center without being within ± 146 bp (the size of DNA wrapped around the nucleosome) of any previously placed nucleosome center. We then take the remaining regions of DNA that are not high-ATAC-coverage, and fill them with nucleosomes according to a Poisson process (i.e., exponential distribution) with mean 43 bp, as learned from the EM data, as described in the previous section.

Inflammatory stimulation of THP-1-derived macrophages, 4C-seq library preparation and sequencing

4C-seq was performed according to the protocol described by Krijger et al. [40], with minor modifications. Briefly, 10 million THP-1 cells were seeded in 10 cm dishes and treated with 20 nM phorbol 12-myristate 13-acetate for 48 hours, followed by stimulation with 100 ng/ml ultrapure lipopolysaccharide (Invitrogen, Cat. tlrl-3pelps). Cells were fixed with 2% paraformaldehyde for 10 minutes at room temperature, quenched with 1 M glycine, and lysed using cell lysis buffer comprising 1M Tris-HCl, NP-40, Triton X-100, 5M NaCl, 0.5M EDTA, and 100X proteinase inhibitor. The supernatant was removed, and 10 ml of cold 0.13 M glycine in PBS was added to quench the reaction, followed by shaking on ice for 10 minutes. Cells were lysed in 5 ml of cold cell lysis buffer, comprising 1M Tris-HCl, NP-40, Triton X-100, 5M NaCl, 0.5M EDTA, and 100X proteinase inhibitor. The mixture was shaken on ice for 15 minutes, then scraped and shaken for another 5 minutes. The nuclei were collected by centrifugation at 500g for 5 minutes at 4°C, resuspended in 1.2X restriction enzyme (RE) buffer, and treated with 0.3% SDS and 2.5% Triton X-100, with incubation at 37°C between steps.

Chromatin was fragmented by subjecting isolated nuclei to primary restriction enzyme digestion using the 4-bp cutter Csp6I (ThermoFisher, Cat. ER0211) to achieve optimal contact-mapping resolution, followed by in situ ligation in a reaction mix containing ligase buffer and ligase, incubated overnight at 16°C. Post-ligation, cross-links were reversed by overnight incubation with Proteinase K at 65°C. DNA was purified using SPRI beads and subjected to a second restriction enzyme digestion using NlaIII (NEB, Cat. R0125L) to generate sticky ends for circularization. The digested DNA was then ligated again, and the resulting 3C templates were purified using AMPure XP beads (Beckman Coulter, Cat. A63881), and quantified using the Qubit dsDNA BR Assay Kit. Next, two rounds of PCR amplification were conducted: an inverse PCR to amplify captured regions, followed by an indexing PCR to generate sequencing-ready libraries. Libraries were purified using AMPure XP beads and sequenced on an Illumina MiSeq platform. Data processing, including demultiplexing, trimming, and mapping to the reference genome (hg38), was performed using the pipe4C pipeline as described by Krijger et al. [40]. The PCR primers used were:

Universal forward primer for indexing (5' – 3'): AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTA-

CACGACGCTCTTCCGATCT

Viewpoint gene locus	Forward primer for inverse PCR (5' – 3')	Reverse primer for inverse PCR (5' – 3')
TNF (region 1)	TACACGACGCTCTTCCGATCTTGGGGGTAGGGTTAGTAC	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTACACTTAGTGAGCACCTTC
TNF (region 2)	TACACGACGCTCTTCCGATCTTGGGGGTAGGGTTAGTAC	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCACTCCATACACACTTAGT
IL-6 (region 1)	TACACGACGCTCTTCCGATCTCCTTCCCTGCCCCAGTAC	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTCGAGGGCAGAATGAGC
IL-6 (region 2)	TACACGACGCTCTTCCGATCTCCTTCCCTGCCCCAGTAC	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCTCGAGGGCAGAATG
IL-6 (region 3)	TACACGACGCTCTTCCGATCTCCTTCCCTGCCCCAGTAC	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCTTTCGTTCCCGGTGGG
CXCL9 (region 1)	TACACGACGCTCTTCCGATCTTTTGGATAACTGGTCATG	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGACTGCTACACAATGT
CXCL10 (region 1)	TACACGACGCTCTTCCGATCTCACAGCATTCAATAGTAC	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTTATGCAATGAAGTTCTTTTT
IL-15 (region 1)	TACACGACGCTCTTCCGATCTAGGACTCGATGGAGGTAC	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGGTCTCCTCCTCTACC

Table 1: Viewpoint Gene Primers

Contact signal processing

To compare our ATAC-informed and experimental 4C-seq contact signals, we seek to identify locations of close contact with the locus of interest. We first perform monotonic regression and subtract the resulting fit from the original data. Monotonic, or isotonic, regression is a technique that fits a piecewise-constant function to data, while enforcing monotonicity. Since we expect contact to peak at the transcriptional start site of each gene and to decay to the left and to the right, we fit two separate monotonic regression models for each side of the site and enforce that they converge to the same peak value at the locus of interest. We do not consider data within ± 10 kbp of the locus of interest. We then use a 3rd order Savitzky-Golay filter with a window size of 25 kbp. Background removal and smoothing of this nature leads to the data exhibiting heteroscedasticity, with local variance increase further away from the locus of interest (Figure S11B), which can in turn lead to errors in peak finding. Thus, to address this we normalize our data with a rolling z-score with a window size of 25 kbp, consistent with our choice of filter window size.

Contact detection, hyperparameter selection and statistical significance

With data processed as above, use wavelet peak finding with the SciPy function `find_peaks_cwt` [21, 41]. This method explores wavelengths in the interval $[\alpha, \beta]$, where α,β are hyperparameters of the model. We repeat this procedure for both the 4C-seq and the ATAC-informed contact signals.

Although the biophysical model has no free parameters, the wavelet peak finding algorithm has 2 hyperparameters.

To fit hyperparameters α, and β, we first determine the 4C-seq contact signal parameters by fitting peaks to two replicates of the 4Cseq data, and predicting each other for the 0hr LPS data only. To quantify the predictive performance, for each choice of hyperparameters, we compute the proportion of peaks within 10 kbp of replicates one and two respectively, which we term the (ACC) and precision, or positive predictive value (PPV, alternatively 1-FDR), and take the minimum of these as our metric.

In other words, we optimize hyperparameters to be

$$[\hat{\alpha}, \hat{\beta}] = \underset{\alpha,\beta}{\operatorname{argmax}}\Big(\min\big(ACC(\alpha,\beta), PPV(\alpha,\beta)\big)\Big) = \underset{\alpha,\beta}{\operatorname{argmax}}\Big(\min\big(ACC(\alpha,\beta), 1 - FDR(\alpha,\beta)\big)\Big) \quad (1)$$

Once we compute $\hat{\alpha}$ and $\hat{\beta}$ on the 4C data, we use these parameters to compute the peak locations on the average of the 4C LPS 0hr replicates to use as the ground truth.

We then repeat the above optimization where we have the ATAC-informed contact map predicting this 4C-seq ground truth in the same manner.

To estimate the statistical significance of our results, we compare the results of our ATAC-informed model to a null model in which we assume to be uniform random placement of N peaks, where N is the number of peaks in our estimated 4C ground truth. We sample M different random peak placements, compute the performance metric for each of these, and count the proportion of these placements where this statistic is larger than the performance of the polymer model. In other words, we count the number of random placements with

$$\min\big(ACC(\alpha,\beta), PPV(\alpha,\beta)\big) \geq \min\big(ACC(\hat{\alpha}_{\text{ATAC}}, \hat{\beta}_{\text{ATAC}}), PPV(\hat{\alpha}_{\text{ATAC}}, \hat{\beta}_{\text{ATAC}})\big). \quad (2)$$

Note that, doing so with the training data (unstimulated cells) is not a valid hypothesis test, since this is the result of an optimization procedure,. We then repeat p-value procedure on our test data (LPS-stimulated cells), without refitting our hyperparameters, giving a true statistical p-value.

Acknowledgements

We thank Elizabeth Read (UCI) for valuable discussion. This work is funded by NSF DMS 2052668, NSF Emerging Frontiers URoL 2022182, NIH T32 GM136624, NSF DMS 1763272, and the Simons Foundation (594598, QN).

References

[1] Charles P. Fulco, Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, Rockwell Anyoha, Benjamin R. Doughty, Tejal A. Patwardhan, Tung H. Nguyen, Michael Kane, Elizabeth M. Perez, Neva C. Durand, Caleb A. Lareau, Elena K. Stamenova, Erez Lieberman Aiden, Eric S. Lander, and Jesse M. Engreitz. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12):1664–1669, 2019. ISSN 1061-4036. doi: 10.1038/s41588-019-0538-0.

[2] Rajarshi P. Ghosh and Barbara J. Meyer. Spatial Organization of Chromatin: Emergence of Chromatin Structure During Development. *Annual Review of Cell and Developmental Biology*, 37(1):1–34, 2021. ISSN 1081-0706. doi: 10.1146/annurev-cellbio-032321-035734.

[3] Xuesong Li, Ruigong Zhu, Hong Jiang, Quanwen Yin, Jiaming Gu, Jiajing Chen, Xian Ji, Xuan Wu, Haiping Fu, Hui Wang, Xin Tang, Yuanqing Gao, Bingjian Wang, Yong Ji, and Hongshan Chen. Autophagy enhanced by curcumin ameliorates inflammation in atherogenesis via the TFEB–P300–BRD4 axis. *Acta Pharmaceutica Sinica B*, 12(5):2280–2299, 2022. ISSN 2211-3835. doi: 10.1016/j.apsb.2021.12.014.

[4] Leonid A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1):37–51, 2011. ISSN 0967-3849. doi: 10.1007/s10577-010-9177-0.

[5] Ivan Junier, Ryan K. Dale, Chunhui Hou, François Képès, and Ann Dean. CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the β-globin locus. *Nucleic Acids Research*, 40(16):7718–7727, 2012. ISSN 0305-1048. doi: 10.1093/nar/gks536.

[6] Adrian L. Sanborn, Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P. Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K. Stamenova, Eric S. Lander, and Erez Lieberman Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):E6456–E6465, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1518552112.

[7] Bruno Beltran, Deepti Kannan, Quinn MacPherson, and Andrew J. Spakowitz. Geometrical Heterogeneity Dominates Thermal Fluctuations in Facilitating Chromatin Contacts. *Physical Review Letters*, 123(20):208103, 2019. ISSN 0031-9007. doi: 10.1103/physrevlett.123.208103.

[8] Aymen Attou, Tilo Zülske, and Gero Wedemann. Cohesin and CTCF complexes mediate contacts in chromatin loops depending on nucleosome positions. *Biophysical Journal*, 2022. ISSN 0006-3495. doi: 10.1016/j.bpj.2022.10.044.

[9] Stephanie Portillo-Ledesma, Suckwoo Chung, Jill Hoffman, and Tamar Schlick. Regulation of chromatin architecture by transcription factor binding. *eLife*, 12:RP91320, 2024. doi: 10.7554/elife.91320.

[10] Da Lin, Weize Xu, Ping Hong, Chengchao Wu, Zhihui Zhang, Siheng Zhang, Lingyu Xing, Bing Yang, Wei Zhou, Qin Xiao, Jinyue Wang, Cong Wang, Yu He, Xi Chen, Xiaojian Cao, Jiangwei Man, Aikebaier Reheman, Xiaofeng Wu, Xingjie Hao, Zhe Hu, Chunli Chen, Zimeng Cao, Rong Yin, Zhen F. Fu, Rong Zhou, Zhaowei Teng, Guoliang Li, and Gang Cao. Decoding the spatial chromatin organization and dynamic epigenetic landscapes of macrophage cells during differentiation and immune activation. *Nature Communications*, 13 (1):5857, 2022. doi: 10.1038/s41467-022-33558-5.

[11] Ping Zhang, Harindra E. Amarasinghe, Justin P. Whalley, Chwen Tay, Hai Fang, Gabriele Migliorini, Andrew C. Brown, Alice Allcock, Giuseppe Scozzafava, Phalguni Rath, Benjamin Davies, and Julian C. Knight. Epigenomic analysis reveals a dynamic and context-specific macrophage enhancer landscape associated with innate immune activation and tolerance. *Genome Biology*, 23(1):136, 2022. ISSN 1474-7596. doi: 10.1186/s13059-022-02702-1.

[12] Alicia N. Schep, Jason D. Buenrostro, Sarah K. Denny, Katja Schwartz, Gavin Sherlock, and William J. Greenleaf. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Research*, 25(11):1757–1770, 2015. ISSN 1088-9051. doi: 10.1101/gr.192294.115.

[13] Robert Sch opflin, Vladimir B. Teif, Oliver Müller, Christin Weinberg, Karsten Rippe, and Gero Wedemann. Modeling nucleosome position distributions from experimental nucleosome positioning maps. *Bioinformatics*, 29(19):2380–2386, 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt404.

[14] Horng D. Ou, Sébastien Phan, Thomas J. Deerinck, Andrea Thor, Mark H. Ellisman, and Clodagh C. O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357 (6349), 2017. ISSN 0036-8075. doi: 10.1126/science.aag0025.

[15] Vijaykumar S. Meli, Praveen K. Veerasubramanian, Hamza Atcha, Zachary Reitz, Timothy L. Downing, and Wendy F. Liu. Biophysical regulation of macrophages in health and disease. *Journal of Leukocyte Biology*, 106(2):283–299, 2019. ISSN 0741-5400. doi: 10.1002/jlb.mr0318-126r.

[16] Amber R. Cutter and Jeffrey J. Hayes. A brief review of nucleosome structure. *FEBS Letters*, 589(20PartA): 2914–2922, 2015. ISSN 0014-5793. doi: 10.1016/j.febslet.2015.05.016.

[17] Zev Bryant, Michael D. Stone, Jeff Gore, Steven B. Smith, Nicholas R. Cozzarelli, and Carlos Bustamante. Structural transitions and elasticity from torque measurements on DNA. *Nature*, 424(6946):338–341, 2003. ISSN 0028-0836. doi: 10.1038/nature01810.

[18] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019. ISSN 1471-0056. doi: 10.1038/s41576-018-0089-8.

[19] Wakim and Spakowitz. Model of euchromatin clustering resulting from local nucleosome interactions. 5 2024.

[20] Andrew Routh, Sara Sandin, and Daniela Rhodes. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proceedings of the National Academy of Sciences*, 105(26):8872–8877, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0802336105.

[21] Pan Du, Warren A. Kibbe, and Simon M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl355.

[22] Anil Panigrahi and Bert W. O'Malley. Mechanisms of enhancer action: the known and the unknown. *Genome Biology*, 22(1):108, 2021. ISSN 1474-7596. doi: 10.1186/s13059-021-02322-1.

[23] Warren A. Whyte, David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*, 153(2):307–319, 2013. ISSN 0092-8674. doi: 10.1016/j.cell.2013.03.035.

[24] Sebastian Pott and Jason D Lieb. What are super-enhancers? *Nature Genetics*, 47(1):8–12, 2015. ISSN 1061-4036. doi: 10.1038/ng.3167.

[25] Johan H. Gibcus, Kumiko Samejima, Anton Goloborodko, Itaru Samejima, Natalia Naumova, Johannes Nuebler, Masato T. Kanemaki, Linfeng Xie, James R. Paulson, William C. Earnshaw, Leonid A. Mirny, and Job Dekker. A pathway for mitotic chromosome formation. *Science*, 359(6376), 2018. ISSN 0036-8075. doi: 10.1126/science.aao6135.

[26] Kangsun Yun, Jae-Seon So, Arijita Jash, and Sin-Hyeog Im. Lymphoid Enhancer Binding Factor 1 Regulates Transcription through Gene Looping. *The Journal of Immunology*, 183(8):5129–5137, 2009. ISSN 0022-1767. doi: 10.4049/jimmunol.0802744.

[27] Qiang Shan, Xiang Li, Xia Chen, Zhouhao Zeng, Shaoqi Zhu, Kexin Gai, Weiqun Peng, and Hai-Hui Xue. Tcf1 and Lef1 provide constant supervision to mature CD8+ T cell identity and function by organizing genomic architecture. *Nature Communications*, 12(1):5863, 2021. doi: 10.1038/s41467-021-26159-1.

[28] Chin-Tong Ong and Victor G. Corces. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics*, 15(4):234–246, 2014. ISSN 1471-0056. doi: 10.1038/nrg3663.

[29] Simona Bianco, Darío G. Lupiáñez, Andrea M. Chiariello, Carlo Annunziatella, Katerina Kraft, Robert Schöpflin, Lars Wittler, Guillaume Andrey, Martin Vingron, Ana Pombo, Stefan Mundlos, and Mario Nicodemi. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nature Genetics*, 50(5):662–667, 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0098-8.

[30] Shafigh Mehraeen, Bariz Sudhanshu, Elena F. Koslover, and Andrew J. Spakowitz. End-to-end distribution for a wormlike chain in arbitrary dimensions. *Physical Review E*, 77(6):061803, 2008. ISSN 1539-3755. doi: 10.1103/physreve.77.061803.

[31] Andrew J. Spakowitz and Zhen-Gang Wang. End-to-end distance vector distribution with fixed end orientations for the wormlike chain model. *Physical Review E*, 72(4):041802, 2005. ISSN 1539-3755. doi: 10.1103/physreve.72.041802.

[32] Shelley Sazer and Helmut Schiessel. The biology and polymer physics underlying large-scale chromosome organization. *Traffic*, 19(2):87–104, 2018. ISSN 1398-9219. doi: 10.1111/tra.12539.

[33] Maria I. Arnone and Eric H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, 1997. ISSN 0950-1991. doi: 10.1242/dev.124.10.1851.

[34] Martin C. Frith, Michael C. Li, and Zhiping Weng. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research*, 31(13):3666–3668, 2003. ISSN 0305-1048. doi: 10.1093/nar/gkg540.

[35] Kevin Struhl and Eran Segal. Determinants of nucleosome positioning. *Nature Structural & Molecular Biology*, 20(3):267–273, 2013. ISSN 1545-9993. doi: 10.1038/nsmb.2506.

[36] Iris Dror, Remo Rohs, and Yael Mandel-Gutfreund. How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays*, 38(7):605–612, 2016. ISSN 0265-9247. doi: 10.1002/bies.201600005.

[37] Dijkstra and W. E. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959. ISSN 0029-599X. doi: 10.1007/bf01386390.

[38] Marshal Mandelkern, John G. Elias, Don Eden, and Donald M. Crothers. The dimensions of DNA in solution. *Journal of Molecular Biology*, 152(1):153–161, 1981. ISSN 0022-2836. doi: 10.1016/0022-2836(81)90099-1.

[39] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008. ISSN 1465-6906. doi: 10.1186/gb-2008-9-9-r137.

[40] Peter H.L. Krijger, Geert Geeven, Valerio Bianchi, Catharina R.E. Hilvering, and Wouter de Laat. 4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis. *Methods*, 170:17–32, 2020. ISSN 1046-2023. doi: 10.1016/j.ymeth.2019.07.014.

[41] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1 0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A Nicholson, David R Hagen, Dmitrii V Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G Young, Gavin A Price, Gert-Ludwig Ingold, Gregory E Allen, Gregory R Lee, Hervé Audren, Irvin Probst, Jörg P Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A Brodtkorb, Perry Lee, Robert T McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J Pingel, Thomas P Robitaille, Thomas Spura, Thouis R Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. ISSN 1548-7091. doi: 10.1038/s41592-019-0686-2.