

Euchromatin forms condensed domains with short active regions on the surface

In the format provided by the
authors and unedited

Contents

1	Micro-C data processing	2
1.1	Alignment and filtering	2
1.2	Mapping reads to expected nucleosome center positions	2
1.3	Removal of likely unligated reads	2
1.4	ICE balancing and interpolation of regions with low coverage and low capture probe coverage	2
1.5	Neighbor balancing protocol	3
1.6	Contact map smoothing	4
1.7	Processing of human cell lines data	4
1.8	Neighbor balancing on genome-wide micro-C maps	4
2	View of neighbor balancing as correction for competition between ligation events	4
3	Details of maximum entropy inversion	4
4	Comparison to chromatin tracing	5
5	Comparison to ChromSTEM	5
6	Analysis of 3D structures	6
6.1	Assignment of packing domains using DBSCAN	6
6.2	Assignment of contiguous nucleosome clutches using insulation scores	6
6.3	Estimating the density of domains	7
6.4	Assessing the enrichment of regulatory elements in protrusions	7
7	Epigenetics data	7
8	Note on absolute densities in HCT116 cell simulations	7
9	Supplementary Figures	11

Micro-C data processing

Alignment and filtering

Reads were aligned to the mm39 genome using bowtie2 with options “-reorder -local -very-sensitive-local -maxins 1 -minins 1000000” [1]. These constraints on the insert size are designed to be unsatisfiable, resulting in no “concordantly” aligning reads. This removes a bias towards short-range inward-oriented alignments [2]. The read alignments are then transferred to pairs format using “pairtools parse” with options “-walks-policy mask -add-mapq -min-mapq 2” [3]. Duplicate reads are then removed using “pairtools dedup” with option “-max-mismatch 1”. Only reads where both ends align uniquely to a capture region are retained.

Mapping reads to expected nucleosome center positions

A custom analysis script is used to shift the assigned positions to the most probable nucleosome center positions according to the read pair orientation, i.e. the strand to which each read end aligns, as described below and illustrated in Figure S1a. We acknowledge that not every nucleosome protect the same length of DNA and that entities other than nucleosomes, such as transcription factors, which might protect different lengths of DNA, are likely the source of some reads. In any case, the shifted position is a better guess for the center position than the 5' end of the read.

- Upstream +/ Downstream -: “inward”, result of ligation between the 3' end of an upstream nucleosome and the 5' end of a downstream nucleosome, shift first + protected / 2, shift second - protected / 2
- Upstream - / Downstream +: “outward”, result of ligation between the 5' end of an upstream nucleosome and the 3' end of a downstream nucleosome, shift first - protected / 2, shift second + protected / 2
- Upstream - / Downstream +: “tandem entry”, result of ligation between the 5' end of an upstream nucleosome and the 5' end of a downstream nucleosome, shift first - protected / 2, shift second - protected / 2
- Upstream +/ Downstream +: “tandem exit”, result of ligation between the 3' end of an upstream nucleosome and the 3' end of a downstream nucleosome, shift first + protected / 2, shift second + protected / 2

The number of base pairs protected by a nucleosome was set to 131 as this value maximizes the concordance between the P(s) curve for the different read orientations (Figure S1b). This is inconsistent with the textbook view that 147 bps are protected by each nucleosome, however, a shift of 131 maximizing agreement between read orientations is in agreement with previous work [4].

The genomic region we plan to simulate is divided into 200 bp windows. We assume that a nucleosome is positioned in the center of each of these bins. For each read, we assign each end to a nucleosome corresponding to its aligned genomic position, after applying the shift described above. Read pairs are converted to a contact map using cooler [5].

Removal of likely unligated reads

After shifting, we found that the abundances and distance-dependence of all read directions was concordant, except for an overabundance of “inward” reads between adjacent nucleosomes (Figure S1b). These “inward” reads are in the expected orientation for unligated DNA fragments, and previous work suggests that these reads arise from unligated DNA fragments that escape exonuclease mediated degradation [6]. To correct for the overabundance of close-range “inward” reads, we discarded “inward” reads assigned to i+1 or i+2 nucleosomes and multiplied the remaining counts by 4/3. This operation assumes that true inward ligation events occur at approximately the same rate as other types of junctions. This step, along with shifting the reads according to their read direction, dramatically alters the P(s) curve at short range (Figure S1c).

ICE balancing and interpolation of regions with low coverage and low capture probe coverage

ICE balancing is used to account for variations in coverage across the genome [7]. We made several modifications to the approach to make it more amenable for use with region-capture data.

First, we use modified filtering rules to explicitly mask pixels with low capture probe coverage. The entirety of the region of interest is not covered by capture probes to avoid including probes with sequences that occur multiple

places in the genome, as this would lower the enrichment rate for the region of interest. This poses a problem for ICE balancing. ICE balancing assumes that biases in read counts are multiplicative, e.g. if a region contains twice as many free DNA ends after MNase digestion, two times as many contacts are observed for all pairs including the region. However, low capture probe coverage is instead additive as a ligated section of DNA can be captured by either end. If one end of a pair has no coverage, it can still be captured efficiently if the other end has capture probe coverage, but if the other end does not have capture probe coverage, then it will rarely be captured. This results in problematic gaps in coverage for pairs with no coverage in either end. These are particularly common for $i+1$ contacts as uncovered regions often extend greater than 200 bps, introducing artifacts in the neighbor contact frequencies used in neighbor balancing. These regions are not always filtered out with standard masking protocols as they can still contain a large number of reads. To deal with this we explicitly mask pixels where the total capture probe coverage is below 20%. This removes the problematic gaps in coverage and stabilizes the estimated neighbor contact probabilities (Figure S2b).

Second, we modify the ICE balancing algorithm to account for a boundary effect due to only considering a subset of the genome. About half of the contacts for nucleosomes at the very ends of the region of interest are to out-of-region nucleosomes and thereby not considered in ICE balancing. Applying ICE balancing thereby artificially doubles the contact frequencies for these nucleosomes. To account for this, we modify the protocol for computing marginal contacts in the ICE balancing protocol. For each position, we infer the out-of-region contacts assuming that the missing values have the average contact frequency for their corresponding genomic separation (Figure S2d).

Finally, we integrate an interpolation step to prevent masking from perturbing the contact density in the simulations. The standard practice is to treat masked values as zeros during ICE balancing. However, in our simulations (and in reality) these masked contacts occur with non-zero frequency. Thereby nucleosomes with a high number of masked contacts get inflated contact densities when these masked contacts are added on top of the unmasked values (Figure S2c). These artificial variations in contact density are of about the same magnitude as the true variations. To address this issue, we employ a three phase ICE balancing protocol. The first phase is performed as usual. In the second phase, between each iteration, masked values are interpolated. Interpolation is performed across each diagonal of the contact map to avoid perturbing the relationship between genomic separation and contact frequency. The third stage is standard ICE balancing with the interpolated values included but not re-computed between each stage. The first phase is necessary as we found that interpolating from the beginning can result in the algorithm not converging. The third step is necessary to ensure that all rows sum to the same value as interpolating between each iteration can result in small deviations.

Neighbor balancing protocol

The neighbor balanced contact frequencies $C^{(\text{neighbor})}_{ij}$ are computed from the ICE balanced contact frequencies C as

$$C^{(\text{neighbor})}_{ij} = \frac{C_{ij}}{\sqrt{\frac{1}{2}(C_{i,i-1} + C_{i,i+1})\frac{1}{2}(C_{j-1,j} + C_{j,j+1})}}$$

The ICE balanced contact frequencies are smoothed using a Gaussian kernel smoother with a standard deviation of 1 bin when computing the denominator of this expression.

Applying neighbor balancing to an ICE balanced contact map, as opposed to the raw contact counts, yields more robust results but does not remove (low frequency) variation in contact density. If neighbor contacts were constant over a region, then neighbor balancing will exactly update the neighbor contact frequencies to be one, i.e. if $C_{i-1,i} = C_{i,i+1} = C_{i+1,i+2}$ then $C^{(\text{neighbor})}_{i,i+1} = 1$. However, in the raw contact map, the neighbor contacts are far from uniform and the balancing fails to make the neighbor contact frequencies close exactly one. We considered performing iterative updates to bring the neighbor contact frequencies closer to one, but this gave unstable results with rapidly oscillating contact densities. ICE balancing makes the neighbor contacts smoother (but not completely uniform) allowing the balancing to succeed. Importantly, this does not decrease low frequency variation in contact density. For instance, if a region has contact densities higher than average by a given factor, then ICE balancing will decrease contact frequencies involving this segment by this same factor to make the contact densities uniform. This would include the neighbor contact frequencies, resulting in neighbor balancing increasing the contact frequencies in this region by the correct factor. In the future, it may be possible to develop a procedure that explicitly optimizes for smooth contact densities and neighbor contact frequencies close to 1, combining the ICE and neighbor balancing objectives.

Contact map smoothing

The contact map is smoothed prior to being used in simulations. Despite the fact that RCMC produces much higher-resolution contact maps than typical micro-C experiments, we found that at 200 bp resolution some smoothing is required to obtain maps that are not visibly pixelated. Pixels far from the diagonal have particularly high variance as they generally have lower contact frequencies and thereby fewer reads to estimate the contact frequency. Therefore, we smooth the matrix more aggressively far from the diagonal and not at all near the diagonal. We smooth the map by grouping squares of pixels and assigning all pixels the average value. The squares are doubled in size at farther distances from the diagonal to maintain a consistent average number of reads per coarsened pixel.

Processing of human cell lines data

In the RCMC data for human cell lines from Hong et al. [8], 150 bp was sequenced for each read end, as opposed to 40 bp in the mESC data from Goel et al. [9]. This results in many of the read ends spanning the ligated junction between different genomic regions. Our original pipeline does not correctly handle these chimeric reads, so we made three changes. First, the reads were aligned using ‘bwa-mem2 mem -SP’ instead of our original ‘bowtie’ command to capture multiple alignment positions for chimeric reads. Second, we ran pairtools with options ‘-walks-policy 5any -max-molecule-size 200’. This treats fragments shorter than 200 bp as a single nucleosome protected region and reports multiple contacts per read if there are multiple junctions. Third, since the read ends often span the ligated junction, we often know precisely how long the protected region. We therefore inferred the nucleosome center position as the midpoint of the alignment instead of using a constant shift.

Neighbor balancing on genome-wide micro-C maps

Genome-wide micro-C data are generally too sparse to view at nucleosome resolution (200 bps), so a protocol was developed to apply neighbor balancing at lower resolution. We first compute ICE balanced contact maps at both 200 bp resolution and the desired resolution. Second, we compute the neighbor contact frequencies in the 200 bp map, as described above. Third, for each lower resolution bin, we average the respective 200 bp resolution neighbor contact frequencies. Finally, each row and column is divided by the square root of these average neighbor contact frequencies.

View of neighbor balancing as correction for competition between ligation events

In this work, we assume that the probability of a ligation occurring between two nucleosomes (or, more accurately, two fragments of nucleosome protected DNA) is proportional only to the distance between the two nucleosomes. However, in a single cell, each nucleosome can only ligate productively to one other nucleosome. Therefore, a more accurate model would be to assume that nucleosomes are in competition with one another to be ligated. In effect, nucleosomes surrounded by many nucleosomes will have a lower probability of ligating to any given nucleosome than a nucleosome in a more dilute environment.

Neighbor balancing approximately acts to convert a contact map generated under this competitive process to one generated purely based on pairwise distances assumed in the rest of this study. A nucleosome in a dense environment will have lower neighbor contact frequencies as these interactions are competed for by many other nucleosomes. Consistent with this, neighbor balancing increases the contact density for nucleosomes with low neighbor contacts.

Details of maximum entropy inversion

In MEI, the energy coefficients are typically solved for using Newton’s method [10]. However, considering large regions at 200 bp resolution makes the Hessian matrix too large to fit into memory as its size scales with the number of monomers to the fourth. To enable simulations, we instead use gradient descent:

$$\alpha_{ij}^{t+1} = \alpha_{ij}^t + \gamma(T_{ij} - \hat{S}_{ij}^t)$$

where $\hat{S}^t$ is an iteration averaged matrix of simulated contact frequencies, T is the matrix of target contact frequencies, and γ is the learning rate.

$\hat{S}^t$ is defined as an exponential moving average of the contact maps for each iteration $\hat{S}^t$

$$\hat{S}_{ij}^{(t)} = (1 - \beta)S_{ij}^{(t)} + \beta\hat{S}_{ij}^{(t-1)}$$

This moving average is not used in the first five rounds of optimization as the contact frequencies change rapidly, so the contact frequencies from previous rounds are not informative. In all simulations, $\gamma = 0.3$ and $\beta = 0.75$.

For each iteration, 16 simulations are performed for 11,000,000 steps each, a frame is stored every 10,000 steps, and the first 1,000,000 steps are excluded when computing contact maps. This procedure is run until the simulated P(s) curve closely matches the reference. The simulated contact maps were visually assessed for key features, such as matching contact densities, bulk insulation scores, and presence of microcompartment (Figure S5).

In the first iteration, the initial conformation is created by placing each nucleosome on a Hilbert curve, which provides a dense conformation with no knots. Subsequent iterations are initiated from the final frame of the previous iteration to reduce the required equilibration time.

Comparison to chromatin tracing

Coordinates determined by chromatin tracing contain measurement noise presenting a major challenge for comparison with simulations. Adding noise to coordinates tends to inflate the distances between them, especially if the true distances are of comparable magnitude to the measurement noise. We therefore needed to account for measurement noise to accurately assess whether our simulations reproduce the distance scaling observed in experiments.

To estimate the measurement noise distribution in ORCA data for the *Nanog* locus [11], we used control data in which one region was imaged twice per chromatid. Assuming that the imaged region does not move, the distances between these subsequent measurements represent the difference in position between two independent samples from the noise distribution. Therefore, we searched for an underlying noise distribution for which the difference between two samples matches the empirical distribution of distances between images. Initially, we expected the noise distribution to be normally distributed, but the empirical distribution has too heavy of a tail to be modeled with a normal distribution. We expect this reflects some of the measurement noise corresponding to mis-identification of the region in the image. We found that a scaled t-distribution, which provides an additional shape parameter controlling the weight of the tails, provided a better fit. We performed a grid search to find the scale and shape parameters maximizing the likelihood of the empirical distribution (Figure E2a). We fit parameters separately using only the x and y coordinates, as these appeared to have less measurement noise than the z coordinates (Figure E2b).

For the *Sox2* region [12], no control data is available. Instead, we solved for a noise distribution that, when added to the simulated coordinates, brought the distribution of distances between loci separated by 60 kb into agreement with the ORCA measurements (Figure E2c). This approach effectively assumes that the simulated distances at this scale are on average correct. Thereby, we cannot assess if the absolute distance values are correct. However, this approach allows meaningful assessment of the overall distance scaling trend and relative distances between specific loci. Comparison to live cell imaging data from Li et al. [13] provides support for the absolute scale of distances (Figure E2d).

Comparison to ChromSTEM

To compare quantitatively with ChromSTEM [14], we generated synthetic tomograms by extracting a 100 nm thick slice of a conformation, projecting the coordinates on the x-y plane, and representing each nucleosome as a Gaussian density with a standard deviation of 5 nm. This yields images that visually match the ChromSTEM images with broad “packing domains”, which themselves contain finer structures (Figure E3a,c).

We implemented the algorithm proposed by Li et al. [14] to identify and characterize packing domains (Figure E3b). In their approach, they identify packing domains by smoothing and enhancing the contrast of the image and then identifying local maxima. To characterize each domain they consider how the density changes in concentric circles away from the peak center. In the real data, as well as our simulated data, peaks generally begin with a power-law mass scaling that then decays when the radius exceeds the size of the domain. The radius of the domain is thereby identified by the radius at which the density deviates from the initial regime by 5%. Two additional criteria are also applied, though they rarely determine the domain size: an increase in density or increase of the mass scaling exponent above 3. Domain sizes in our structures closely match those from ChromSTEM with a peak between 50 and 100 nm and a tail extending to several hundred nm (Figure E3d).

The packing domains identified through this analysis are generally smaller and more numerous than those identified by DBSCAN (compare Figure 4h to Figure E3d). This analysis method identifies the shortest dimension of the domain (the minor axis) because when the shortest dimension is exceeded the density begins to drop. Additionally, this analysis places a domain on all local maxima, even if they are connected by continuously high density, so more domains are detected.

We also found that the initial scaling exponents (labeled “D” in Figure E3d) to be in agreement with experiments. The value of ~ 2.5 suggests nucleosomes inside the domains are more clustered than expected. Visual inspection of our simulations suggests that this internal structure is due to the packing domains being formed by nucleosome clutches.

Analysis of 3D structures

Assignment of packing domains using DBSCAN

DBSCAN [15], as implemented in scikit-learn [16], was used to decompose chromatin conformations into a set of packing domains and protrusions. DBSCAN has two tunable parameters: epsilon, which controls the maximum distance at which two points are considered neighboring, and min_samples, which controls the number of neighbors needed for a point to be considered a “core” point. Epsilon was set to 30 nm and min_samples was set to 14. This combination of parameters results in a nucleosome concentration of at least 205 μM being required at the core of a domain. The clusters identified by DBSCAN were considered packing domains and the remaining “noise” points are considered protrusions. Protrusions less than 3 nucleosomes long that exit and return to the same packing domain were relabeled as part of the packing domain. A two nucleosome protrusion was added between packing domains with no protrusion linking between them.

We chose the epsilon and concentration (min samples) parameters to classify visually coherent groups of nucleosomes as domains. Increasing the required concentration results in some groups that appear spatially coherent being broken into multiple domains, whereas lowering the concentration seems to merge visually separated domains (Figure S14a). Admittedly, this choice is subjective.

These parameters impact our quantitative results, but selecting different values does not alter our core conclusions. In Figure S14, we compare how changing the DBSCAN parameters impacts the results for the most (klf1) and least (fhn2) active regions we studied. First, increasing the concentration or radius results in more nucleosomes being called protrusions. Despite the changes in absolute protrusion probability, the values are correlated with each other and with ATAC-seq (Figure S14c,d). Second, increasing either parameter right shifts the peak of small domain sizes (Figure S14e, around 10 kb) and generally decreases the frequency of large domains. These small domain sizes have a similar distribution between the two regions in all cases, highlighting a consistent local structure. For all parameters, we also reach a similar conclusion that the fhn2 region more often forms large domains. Finally, the average protrusion probabilities change proportionately for the two systems for all parameters we considered (Figure S14b). In this way, regardless of the chosen parameters, we reach a similar conclusion that euchromatin regions are only modestly more likely to form protrusions.

Assignment of contiguous nucleosome clutches using insulation scores

We define clutches using insulation scores calculated from single-frame contact maps. Positions with fewer than half the median number of cross contacts are defined as clutch boundaries, and the segments between such boundaries are defined as clutches.

Clutches were assigned by considering single frame insulation scores. First, a contact map was computed using a capture radius of 40 nm. Then, insulation scores were computed through cooltools using a window size of ten nucleosomes, ignoring the first two diagonals [17]. Finally, positions with an insulation score less than half the median value across all systems were called as boundaries and the remaining contiguous blocks were defined as contiguous nucleosome clutches (Figure S10a).

By construction clutches are insulated from flanking regions, however, we find that the clutches are also generally not mixed with other regions (Figure 5c, Figure S11c). The precise choice of parameters used to define clutches, and thereby their precise sizes and boundary probabilities, is somewhat arbitrary, as hierarchically folded structure exists across a broad range of scales (Figure S9). However, there are several factors supporting that clutches, as defined here, are an important layer of chromatin organization. First, these are the largest contiguous segments that are mostly unmixed (Figure S11e). Second, their sizes coincide with a change in the end-to-end distance scaling exponent from $1/2$ to $1/3$ (Figure 5g). Finally, their size distribution overlaps the distribution of packing

domain sizes, suggesting that clutches represent an underlying feature that can either occur alone as small packing domains or group together into large packing domains (Figure 5f, Figure 6a).

Estimating the density of domains

Estimating the density of domains was surprisingly challenging. Their irregular shape and internal cavities makes it unclear how to define their surface and different approaches resulted in different density estimates. Here we summarize the methods we considered and comment on their results. First, we considered using the probe excluded volume, defined as the volume unreachable by a sphere with a radius of 20 nm. This method gave inflated estimates for small domains because the surface was drawn precisely around the nucleosomes, whereas it gave deflated estimates for large domains because the packing domains often contain unfilled cavities (Figure E4d,g). Second, we augmented this approach by padding the nucleosome diameters to prevent inflated estimates for small domains. Specifically, we padded the nucleosome radii to 11 nm, which is roughly half the average distance between nucleosomes inside of packing domains (Figure E4h). Third, we considered the maximum density inside of spheres centered on each nucleosome in a domain (Figure E4c). This yielded substantially higher densities than the other methods (Figure E4c,f). Fourth, we used the radius of gyration to characterize the size of packing domains. This approach does not provide a density, but it has no free parameters and does not appear to display any significant size dependent bias. Finally, we considered the relationship between the time a nucleosome spends on the surface and its average local density. These quantities are linearly related and if we extrapolate the trend to no time spent on the surface, we obtain an average local density of about 300 μM .

While these different approaches yield different absolute densities, the values are very similar between systems, after accounting for size dependent biases.

Assessing the enrichment of regulatory elements in protrusions

Whether protruding regions contain more than expected regulatory elements was assessed using a permutation test (Figure S15). Promoters were defined as RefSeq annotated gene starts with H3K4me3 signal p-value greater than 10. Regulatory elements more generally were defined as any position with ATAC-seq coverage greater than 50. For both features, we computed the risk ratio, defined as the probability of protruding monomers being a regulatory element divided by the probability of all other monomers being a regulatory element. For all regions, we found that both promoters and accessible regions generally are more likely to form protrusions (risk ratios ranging from 1.92 to 5.01). We used a permutation test to assess the statistical significance. To create a null distribution, we kept the regulatory annotations fixed but shifted the protrusion labels by a random shift, wrapping the entries going past the end of the array back to the start. This null distribution maintains correlations along the genome. We repeated this for 1000 trials and the observed risk ratio was higher than nearly all of these samples ($p < 0.005$ for all regions).

Epigenetics data

We used pybigwig [18] to load epigenetics tracks and UCSC liftover tool [19] to transfer the tracks to mm39 coordinates where appropriate. Where available, we used signal p-value tracks from ENCODE [20, 21].

Note on absolute densities in HCT116 cell simulations

Even small variations in the micro-C procedure can impact the effective capture radius. Ideally, one would calibrate the capture radius used in the simulations to chromatin tracing data for each new dataset. However, chromatin tracing data was not available in HCT116 cells for any of the simulated regions. Therefore, we simply used the same parameters as in the mESC simulations. Consequently, while our simulations support a similar relative density between active and inactive regions in HCT116 cells, they cannot assess if the absolute densities differ from mESCs.

References

- [1] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4): 357–359, April 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923. URL <https://www.nature.com/articles/nmeth.1923>. Publisher: Nature Publishing Group.
- [2] Houda Belaghzal, Job Dekker, and Johan H. Gibcus. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods*, 123:56–65, July 2017. ISSN 1046-2023. doi: 10.1016/j.ymeth.2017.04.004. URL <https://www.sciencedirect.com/science/article/pii/S1046202316304662>.
- [3] Open2C, Nezar Abdennur, Geoffrey Fudenberg, Ilya M. Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, and Sergey V. Venev. Pairtools: From sequencing data to chromosome contacts. *PLOS Computational Biology*, 20(5):e1012164, May 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1012164. URL <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012164>. Publisher: Public Library of Science.
- [4] Masae Ohno, Tadashi Ando, David G. Priest, Vipin Kumar, Yamato Yoshida, and Yuichi Taniguchi. Sub-nucleosomal Genome Structure Reveals Distinct Nucleosome Folding Motifs. *Cell*, 176(3):520–534.e25, January 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2018.12.014. URL <https://www.sciencedirect.com/science/article/pii/S0092867418316283>.
- [5] Nezar Abdennur and Leonid A Mirny. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, 36(1):311–316, January 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz540. URL <https://doi.org/10.1093/bioinformatics/btz540>.
- [6] Tsung-Han S. Hsieh, Assaf Weiner, Bryan Lajoie, Job Dekker, Nir Friedman, and Oliver J. Rando. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, 162(1):108–119, July 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.05.048. URL <https://www.sciencedirect.com/science/article/pii/S0092867415006388>.
- [7] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, October 2012. ISSN 1548-7105. doi: 10.1038/nmeth.2148. URL <https://www.nature.com/articles/nmeth.2148>. Publisher: Nature Publishing Group.
- [8] Clarice K. Y. Hong, Fan Feng, Varshini Ramanathan, Jie Liu, and Anders S. Hansen. Genome structure mapping with high-resolution 3D genomics and deep learning, May 2025. URL <https://www.biorxiv.org/content/10.1101/2025.05.06.650874v1>. Pages: 2025.05.06.650874 Section: New Results.
- [9] Viraat Y. Goel, Miles K. Huseyin, and Anders S. Hansen. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nature Genetics*, 55(6):1048–1056, June 2023. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-023-01391-1. URL <https://www.nature.com/articles/s41588-023-01391-1>.
- [10] Benoît Roux and Jonathan Weare. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *The Journal of Chemical Physics*, 138(8):084107, February 2013. ISSN 0021-9606. doi: 10.1063/1.4792208. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3598863/>.
- [11] Taihei Fujimori, Carolina Rios-Martinez, Abby R. Thurm, Michaela M. Hinks, Benjamin R. Doughty, Joydeb Sinha, Derek Le, Antonina Hafner, William J. Greenleaf, Alistair N. Boettiger, and Lacramioara Bintu. Single-cell chromatin state transitions during epigenetic memory formation. *bioRxiv*, page 2023.10.03.560616, October 2023. doi: 10.1101/2023.10.03.560616. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10592931/>.
- [12] Antonina Hafner, Minhee Park, Scott E. Berger, Sedona E. Murphy, Elphège P. Nora, and Alistair N. Boettiger. Loop stacking organizes genome folding from TADs to chromosomes. *Molecular Cell*, 83(9):1377–1392.e6, May 2023. ISSN 1097-2765. doi: 10.1016/j.molcel.2023.04.008. URL <https://www.sciencedirect.com/science/article/pii/S1097276523002526>.

- [13] Jieru Li, Angela Hsu, Yujing Hua, Guanshi Wang, Lingling Cheng, Hiroshi Ochiai, Takashi Yamamoto, and Alexandros Pertsinidis. Single-gene imaging links genome topology, promoter–enhancer communication and transcription control. *Nature Structural & Molecular Biology*, 27(11):1032–1040, November 2020. ISSN 1545-9985. doi: 10.1038/s41594-020-0493-6. URL <https://www.nature.com/articles/s41594-020-0493-6>. Publisher: Nature Publishing Group.
- [14] Yue Li, Vasundhara Agrawal, Ranya K. A. Virk, Eric Roth, Wing Shun Li, Adam Eshein, Jane Frederick, Kai Huang, Luay Almassalha, Reiner Bleher, Marcelo A. Carignano, Igal Szleifer, Vinayak P. Dravid, and Vadim Backman. Analysis of three-dimensional chromatin packing domains by chromatin scanning transmission electron microscopy (ChromSTEM). *Scientific Reports*, 12(1):12198, July 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-16028-2. URL <https://www.nature.com/articles/s41598-022-16028-2>. Publisher: Nature Publishing Group.
- [15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD’96*, pages 226–231, Portland, Oregon, August 1996. AAAI Press.
- [16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928. URL <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [17] Open2C, Nezar Abdennur, Sameer Abraham, Geoffrey Fudenberg, Ilya M. Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, Betul A. Oksuz, Sergey V. Venev, and Yao Xiao. Cooltools: Enabling high-resolution Hi-C analysis in Python. *PLOS Computational Biology*, 20(5):e1012067, May 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1012067. URL <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012067>. Publisher: Public Library of Science.
- [18] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, July 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw257. URL <https://doi.org/10.1093/nar/gkw257>.
- [19] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(suppl_1):D590–D598, January 2006. ISSN 0305-1048. doi: 10.1093/nar/gkj144. URL <https://doi.org/10.1093/nar/gkj144>.
- [20] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 1476-4687. doi: 10.1038/nature11247. URL <https://www.nature.com/articles/nature11247>. Publisher: Nature Publishing Group.
- [21] Benjamin C. Hitz, Jin-Wook Lee, Otto Jolanki, Meenakshi S. Kagda, Keenan Graham, Paul Sud, Idan Gabdank, J. Seth Strattan, Cricket A. Sloan, Timothy Dreszer, Laurence D. Rowe, Nikhil R. Podduturi, Venkat S. Malladi, Esther T. Chan, Jean M. Davidson, Marcus Ho, Stuart Miyasato, Matt Simison, Forrest Tanaka, Yunhai Luo, Ian Whaling, Eurie L. Hong, Brian T. Lee, Richard Sandstrom, Eric Rynes, Jemma Nelson, Andrew Nishida, Alyssa Ingersoll, Michael Buckley, Mark Frerker, Daniel S. Kim, Nathan Boley, Diane Trout, Alex Dobin, Sorena Rahmanian, Dana Wyman, Gabriela Balderrama-Gutierrez, Fairlie Reese, Neva C. Durand, Olga Dudchenko, David Weisz, Suhas S. P. Rao, Alyssa Blackburn, Dimos Gkoutaroulis, Mahdi Sadr, Moshe Olshansky, Yossi Eliaz, Dat Nguyen, Ivan Bochkov, Muhammad Saad Shamim, Ragini Mahajan, Erez Aiden, Tom Gingeras, Simon Heath, Martin Hirst, W. James Kent, Anshul Kundaje, Ali Mortazavi, Barbara Wold, and J. Michael Cherry. The ENCODE Uniform Analysis Pipelines, April 2023. URL <https://www.biorxiv.org/content/10.1101/2023.04.04.535623v1>. Pages: 2023.04.04.535623 Section: New Results.
- [22] Longzhi Tan, Dong Xing, Chi-Han Chang, Heng Li, and X. Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, August 2018. doi: 10.1126/science.aat5641.

URL <https://www.science.org/doi/10.1126/science.aat5641>. Publisher: American Association for the Advancement of Science.

- [23] Leonid A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1):37–51, January 2011. ISSN 1573-6849. doi: 10.1007/s10577-010-9177-0. URL <https://doi.org/10.1007/s10577-010-9177-0>.

Supplementary Figures

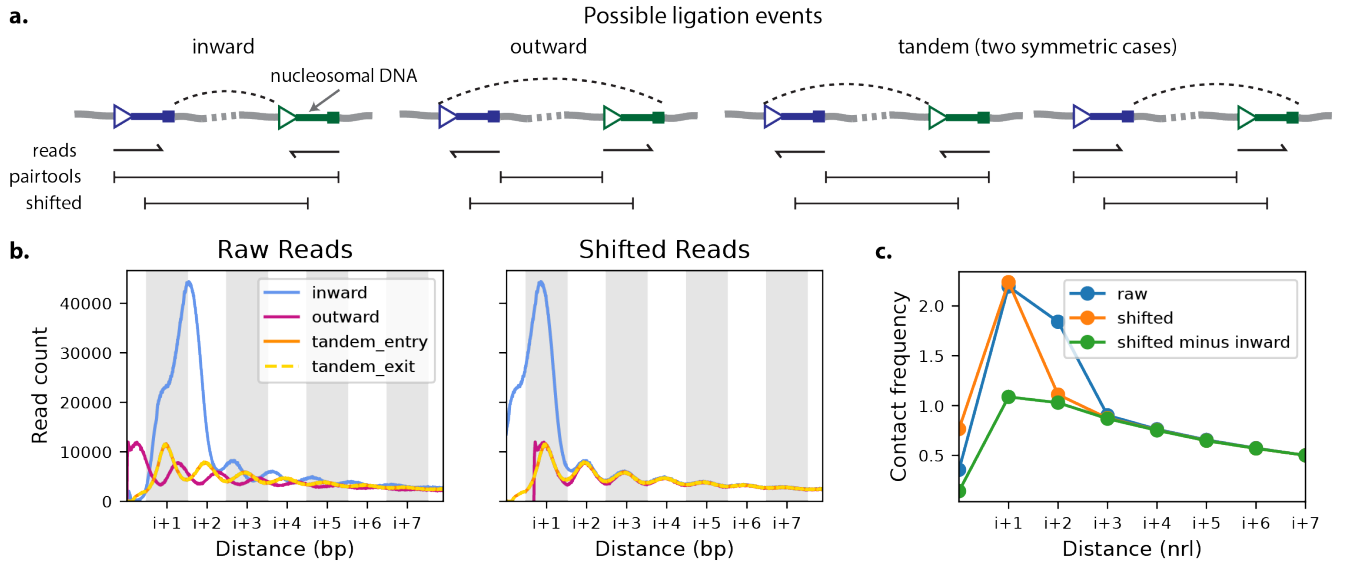


Fig. S1: Impact of micro-C data processing on local interaction patterns. (a) A cartoon depicting micro-C read orientations and the strategy for estimating nucleosome center positions (shifted). (b) Base resolution $P(s)$ curves separated by read orientation. Using the 5'-ends of the reads the curves are out of phase with one another (raw reads), but after shifting to estimated nucleosome center positions they are mostly concordant. However, a disproportionate number of short range inwards oriented reads are present. (c) Nucleosome resolution $P(s)$ curves with different processing protocols. For the raw data, the large peak in $i+1$ and $i+2$ reads stems from the large number of short range inwards reads. Shifting the reads to estimated nucleosome center positions moves this peak to the $i+1$ position. Removing inwards reads from the $i+1$ and $i+2$ bins and rescaling the remaining counts by $4/3$ leaves a smooth curve (shifted minus inward).

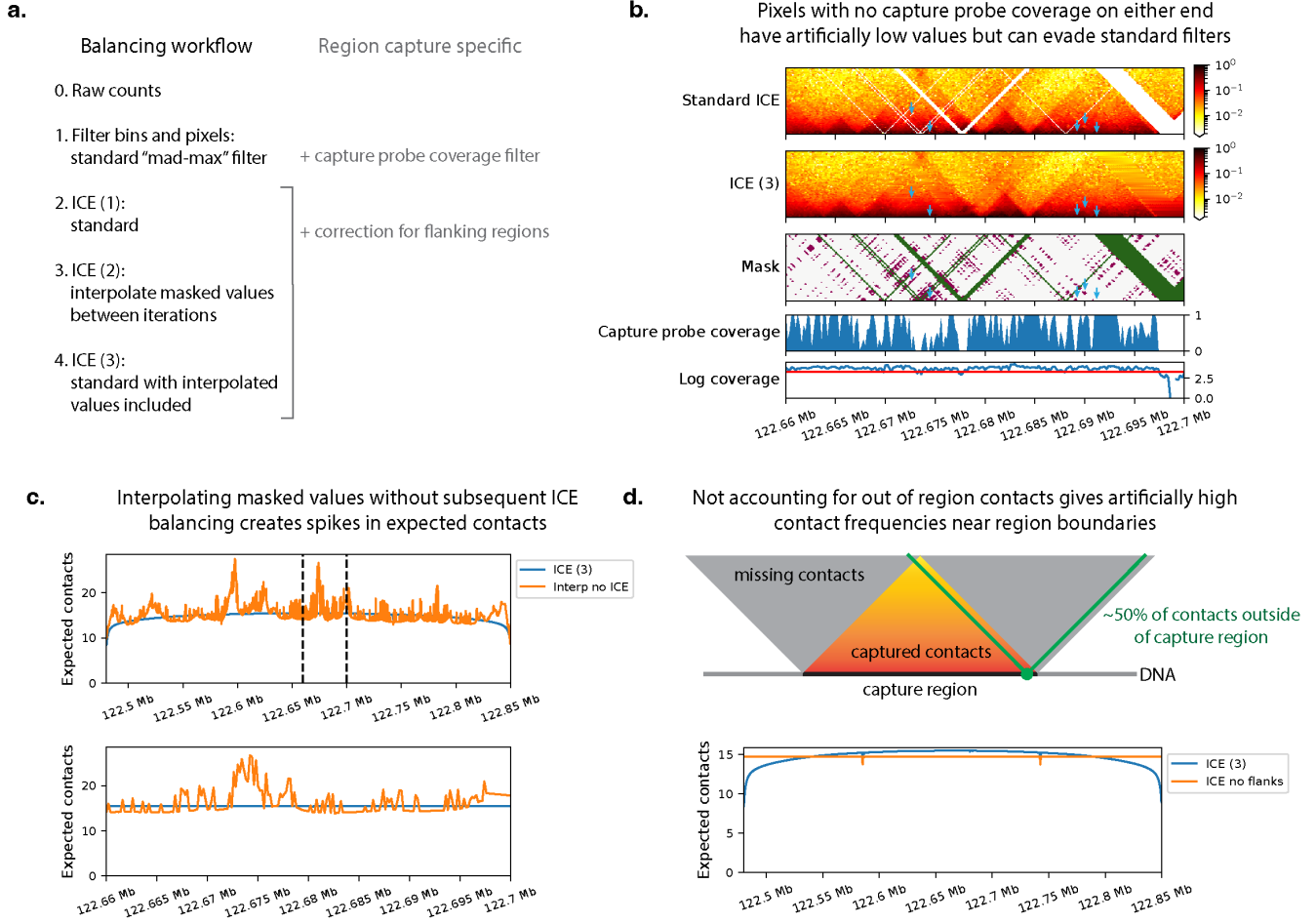


Fig. S2: An augmented iterative correction balancing workflow. (a) Outline of our workflow for micro-C data processing. (b) Lack of capture probe coverage can result in artifacts not accounted for using standard processing approaches. The capture probe coverage track shows the fraction of 200-bp windows that are covered by capture probes. The mask image shows pixels filtered based on read coverage in green and additional pixels filtered based on capture probe coverage in magenta. These pixels without capture probe coverage, highlighted by blue arrows, can result in holes in the contact map using standard ICE balancing. These holes are removed using our procedure. (c) Typically, masked pixels of the contact map are treated as zeros in ICE balancing. However, this results in artificial variation in contact density if these masked values are subsequently interpolated because some windows interact more closely with masked positions. This is problematic as our simulations effectively interpolate missing values. Our augmented ICE balancing procedure overcomes this issue. (d) Enforcing that the capture region has uniform coverage creates artificially high interaction frequencies near the boundaries. We account for this by filling in the missing contacts (gray) with the average value for the respective genomic separation.

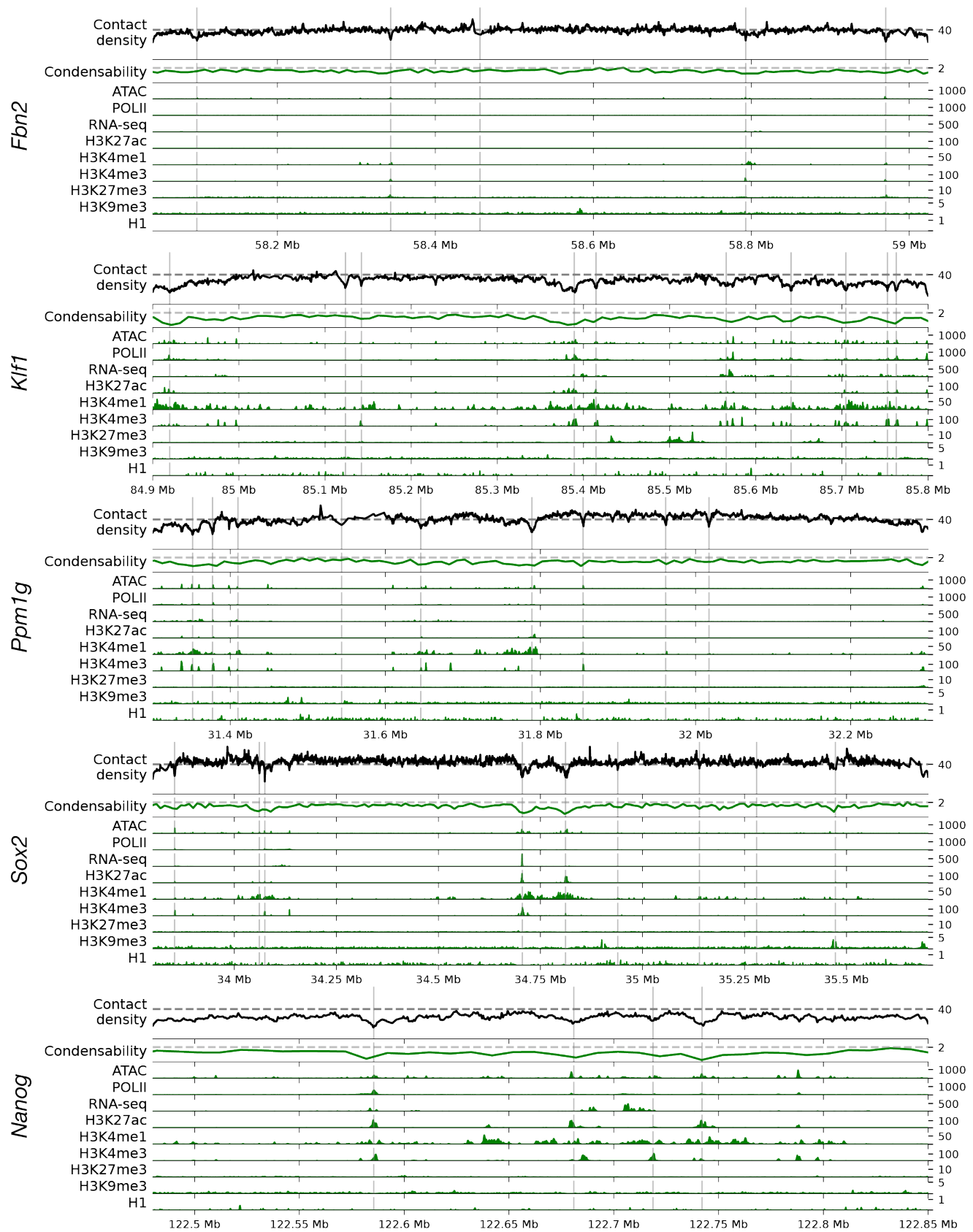


Fig. S3: Contact densities predicted by neighbor balancing are correlated with epigenetics marks Equivalent to Figure 1c but showing all five regions and a larger repertoire of epigenetics tracks.

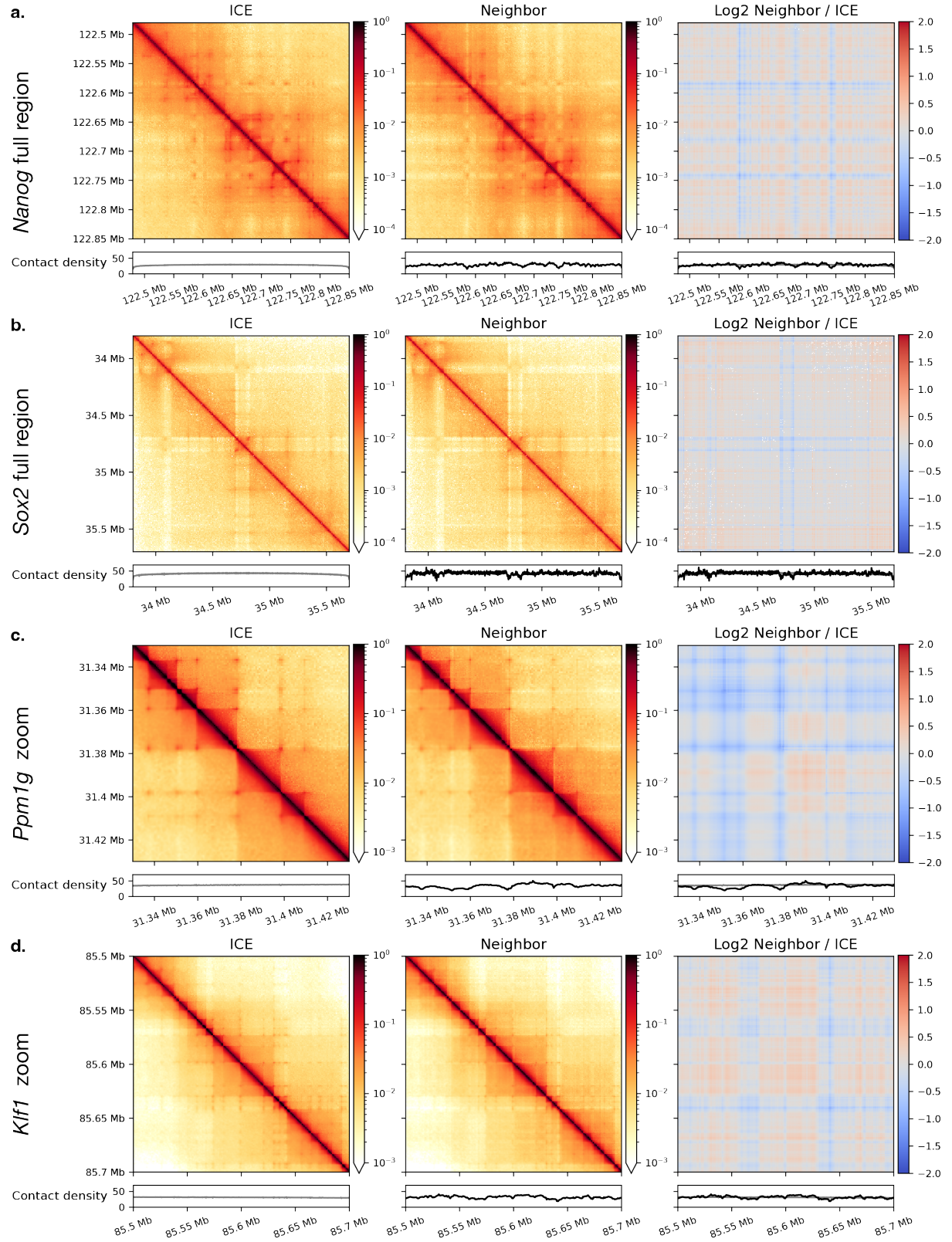


Fig. S4: **Enhancer–promoter interactions are reduced by neighbor balancing** (a, b) Contact maps for the full regions show that neighbor balancing generally reduces the strength of enhancer–promoter interactions. (c, d) Zoom ins of microcompartment containing regions show that neighbor balancing quantitatively decreases the strength of the interactions, but they remain clearly visible in the contact map.

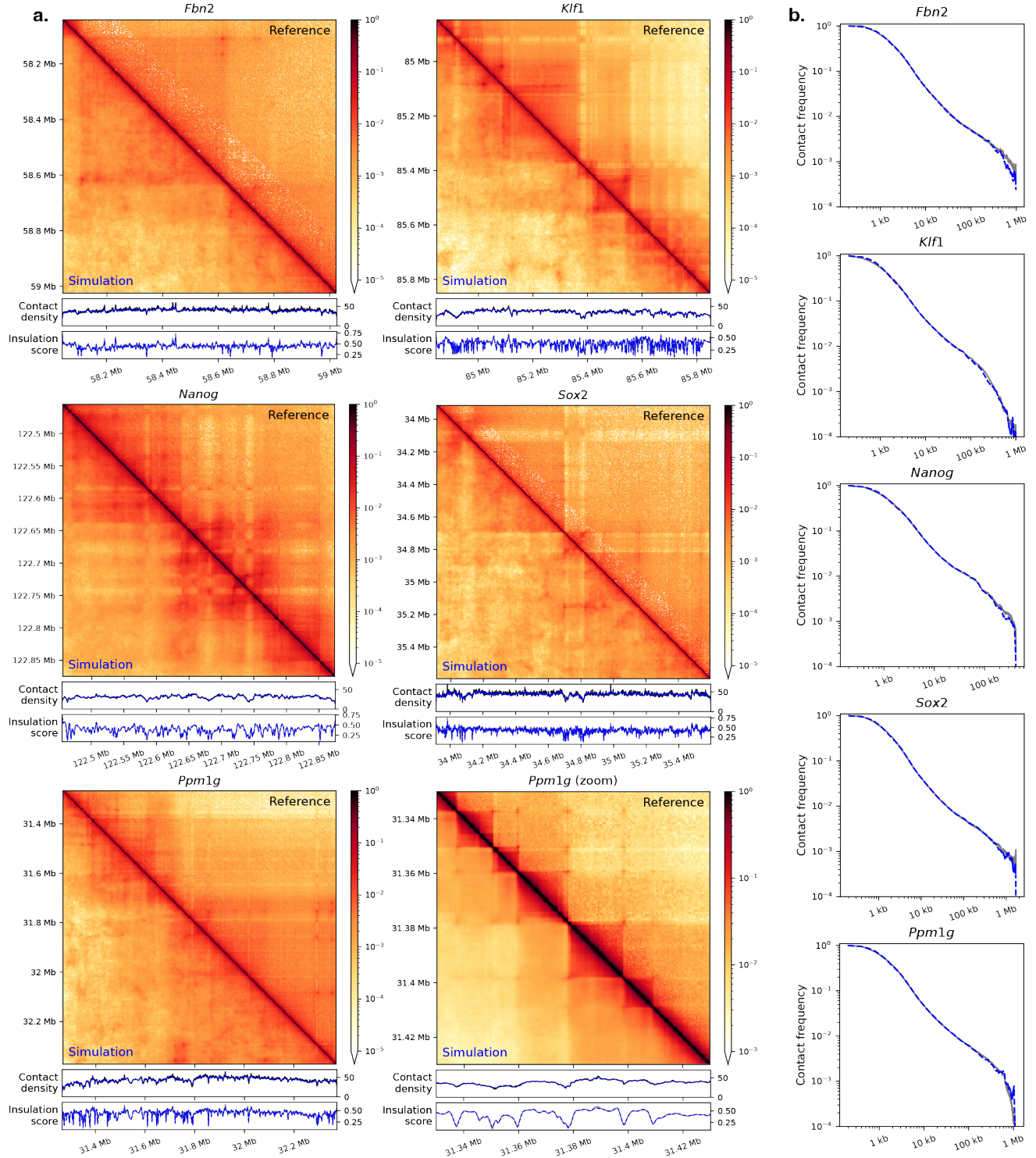


Fig. S5: **Simulations successfully reproduce experimental contact maps.** (a) Simulated contact maps are shown in the lower triangle and reference RCMC contact maps are shown in the upper triangle. Below, the contact densities and insulation scores using a window size of 10 nucleosomes (as used in our definition of clutches) are shown for simulation (blue) and reference (black). The bottom right plot shows a zoom in of the *Ppm1g* region highlighting that our simulations capture fine-scale microdomains and microcompartments. (b) P(s) curves for simulation (blue) and reference RCMC maps (gray).

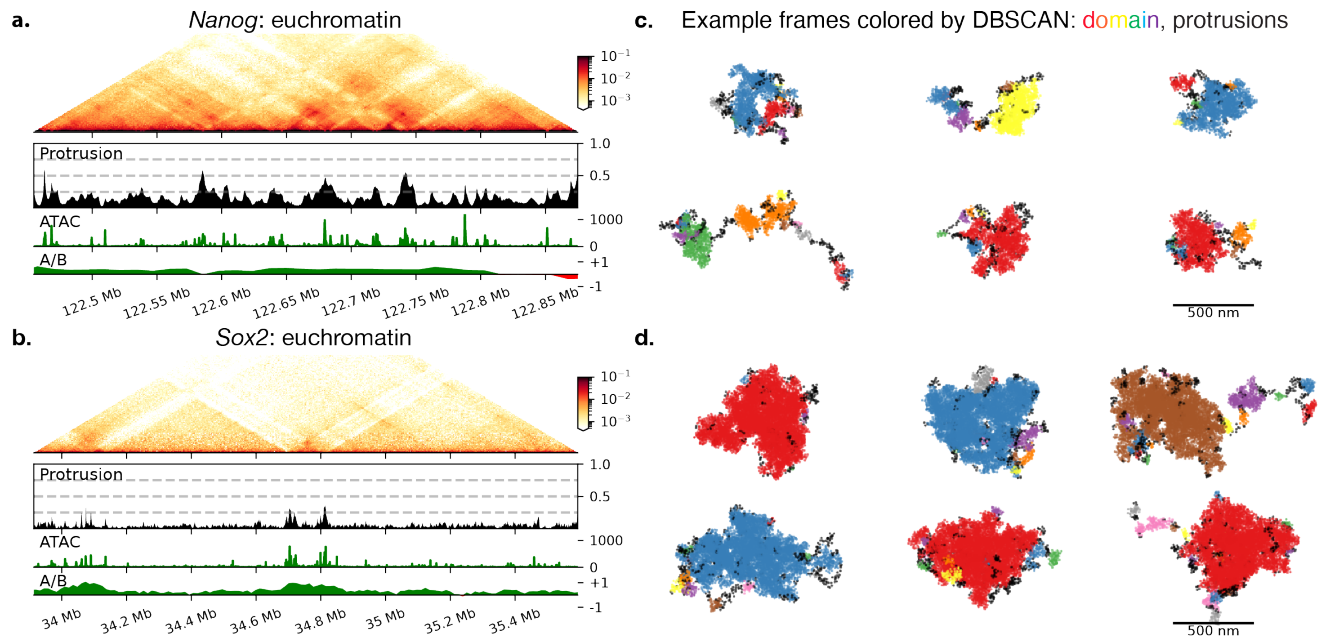


Fig. S6: **Supporting both heterochromatin and euchromatin are largely comprised of condensed domains.** Equivalent to Figure 3a-f for the remaining two systems. Note that the *Sox2* system is 1.6 Mb and the *Nanog* system is 0.37 Mb, whereas the remaining systems are approximately 1 Mb.

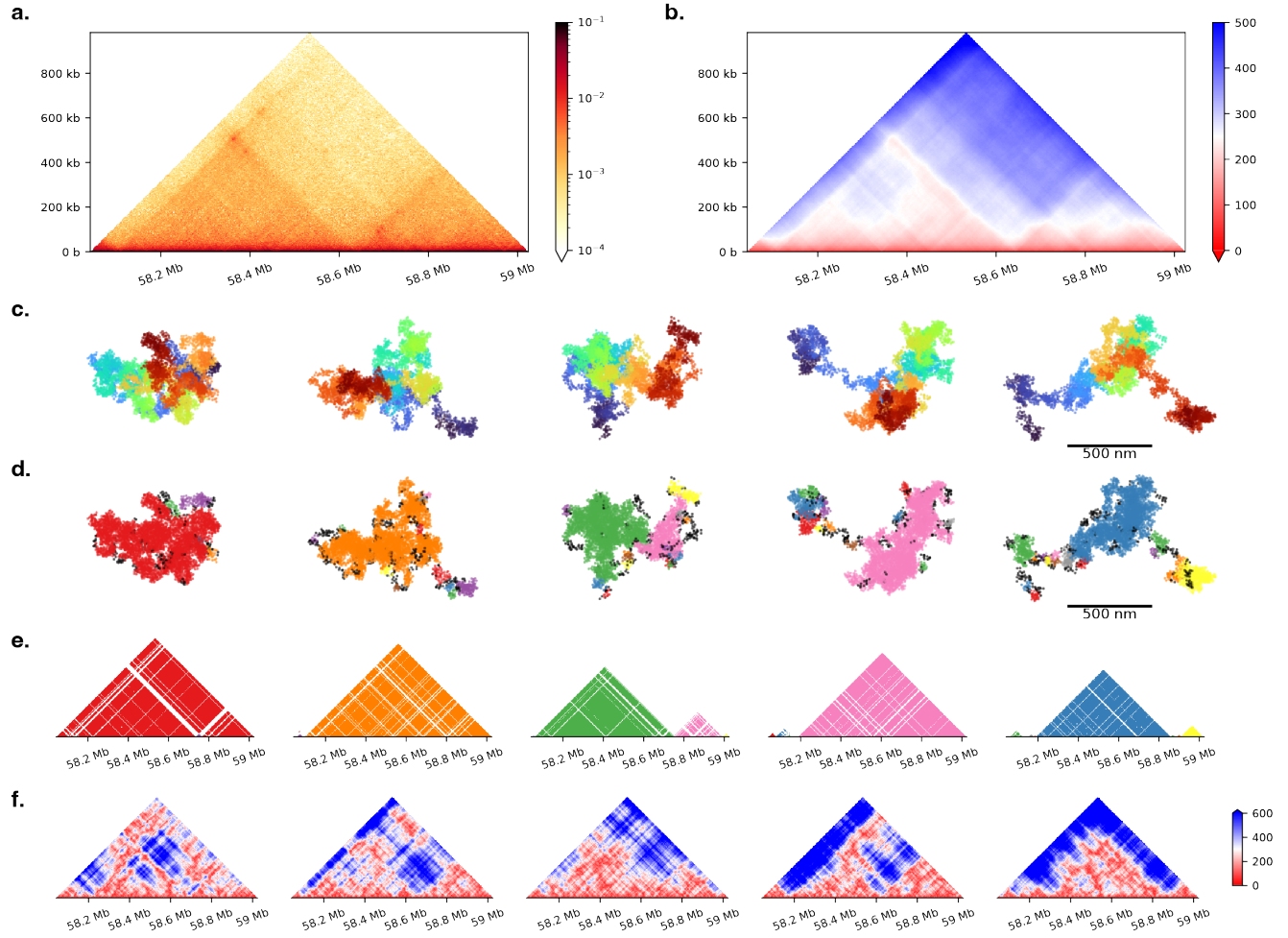


Fig. S7: Simulations capture the stochastic character of packing domains. (a) The reference contact map for the *Fbn2* region, shows that the region is decomposed into three TADs: one complete and two partial. (b) Median simulated distances show that nucleosomes within TADs are indeed on average closer to each other than nucleosomes in different TADs. However, in individual simulation frames, the allotment of this region into packing domains is highly varied. Representative simulation frames colored by chain index (c) and packing domain assignments (d). (e) An illustration of the genomic content of each packing domain. (f) Per-frame distance maps.

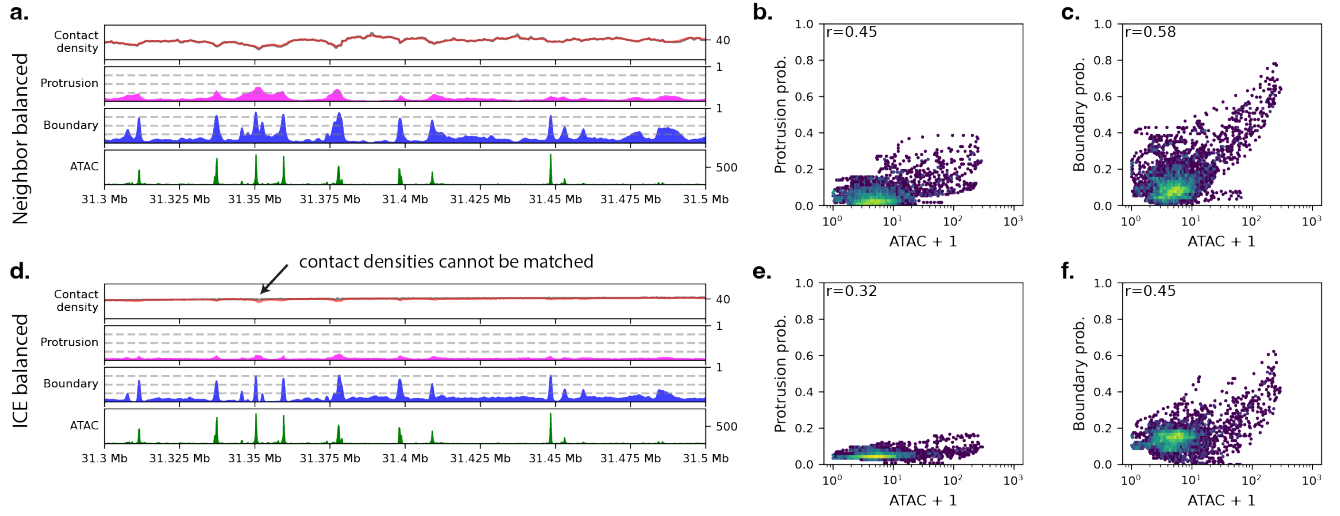


Fig. S8: **Simulations using ICE balanced contact maps.** (a, d) Reference (gray) and simulated contact densities, protrusion probabilities, clutch boundary probabilities, and ATAC-seq coverage for a portion of the *Ppm1g* region from simulations using a neighbor balanced and ICE balanced contact map, respectively. (b, e) Correlations between ATAC-coverage and protrusion probability. (c, f) Correlations between ATAC-coverage and clutch boundary probability. When using ICE balanced contact maps, the contact densities are largely constant and accordingly, there is little variation in the protrusion probability. There is still some variation because neighbor contact frequencies cannot be optimized (as the bond length is largely fixed) and ICE balanced data contains substantial variation in the neighbor contact frequencies. In particular, ATAC peaks generally have higher values in the first off-diagonal, leading to deviations between the reference contact density and the simulated local density.

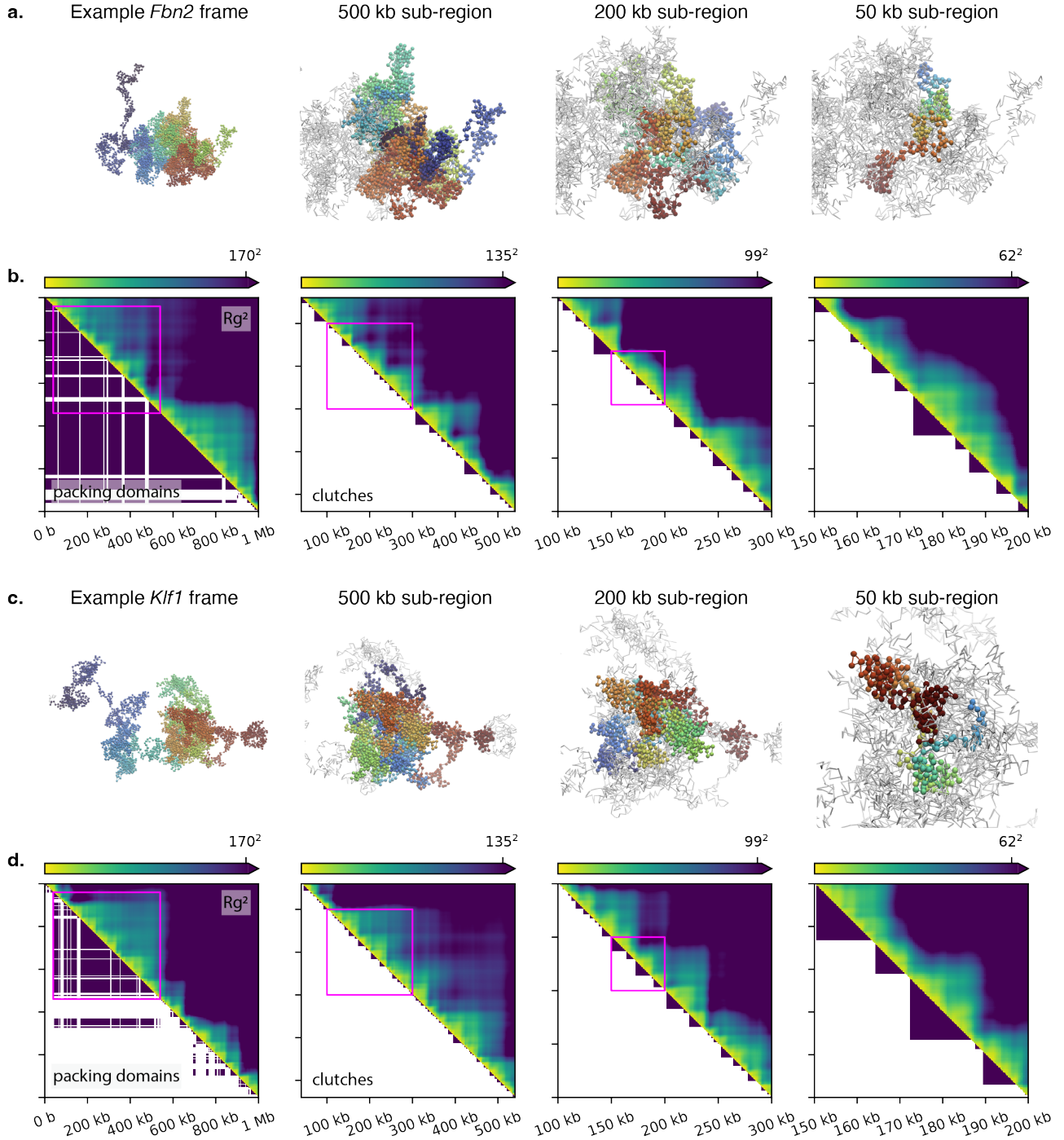
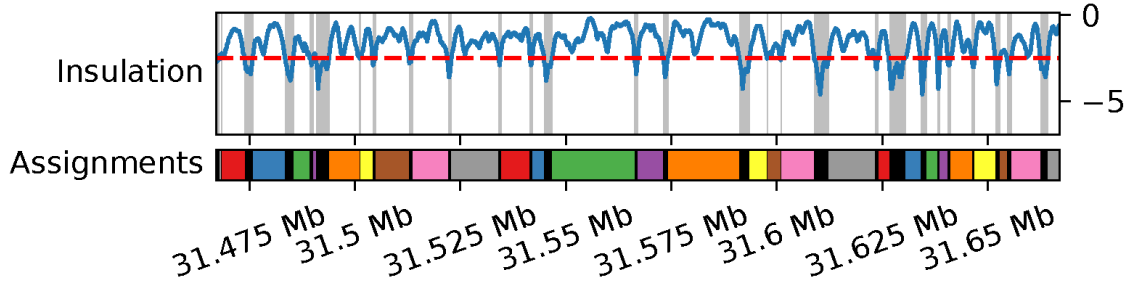


Fig. S9: **Chromatin has hierarchical structure across a broad range of scales.** (a) Example simulation frame for the optimized *Fbn2* system. Each panel shows increasingly smaller sub-regions in rainbow colors with the beginning of the region in red and end in purple and the remainder of the system in white. (b) Comparison of the squared radius of gyration (Rg^2) to packing domain and clutch assignments for the simulation frame shown above. Each pixel of the upper triangle of each panel represents the Rg^2 for the corresponding segment of the system, i.e. the ij th pixel is the Rg^2 for nucleosomes i through j . The lower triangle shows the packing domain assignments (left panel) or clutch assignments (remaining panels). Self-interacting domains are present at all scales. Clutches, though defined using insulation scores, not Rg , represent a slice of this hierarchical structure. (c, d) Equivalent for an example simulation frame for the *Klfl* system. This figure was inspired by Figure 2F from Tan et al. [22]. They observe a fractal globule at greater than Mb scales [23]. Our model similarly shows self-interacting domains down to the kilobase scale. However, at small scales, sub-regions expand rapidly within packing domains.

a. Example single frame insulation scores and domain assignments (black -> boundary)



b. Boundary probability at an ATAC peak v. featureless region with different thresholds

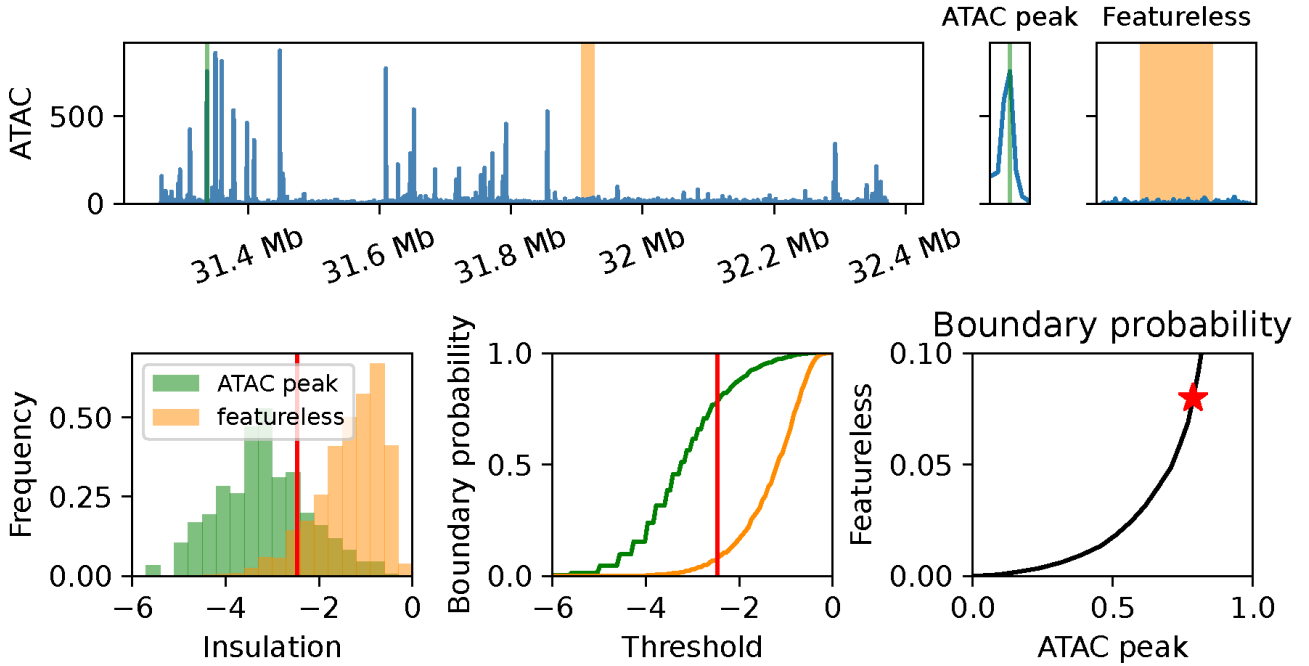


Fig. S10: **Impact of threshold choice in insulation analysis.** (a) Illustration of clutch assignments. (b) Comparison of boundary probabilities at an ATAC peak to those in a largely featureless region. Insulation scores at the ATAC peak are generally lower than in the featureless region—consistent with this region being more locally extended. However, there is significant overlap between the distributions, supporting that meaningful clutch boundaries also occur spontaneously in featureless regions.

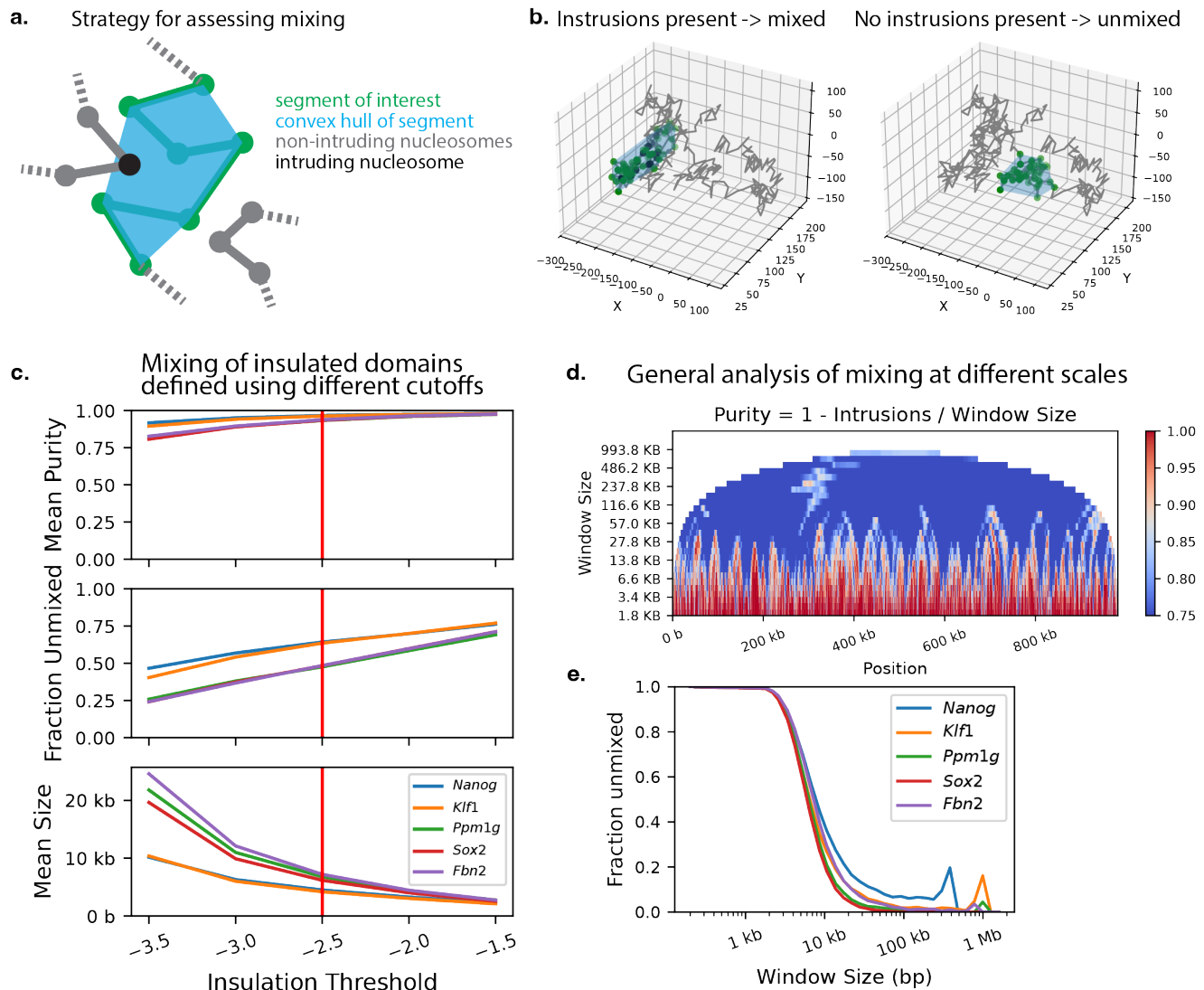


Fig. S11: **Clutches are largely unmixed.** (a) The extent to which a segment of interest is mixed with the rest of the system was assessed by counting the number of other nucleosomes inside the convex hull of the segment of interest. (b) A segment is considered mixed if any such intrusions are present and unmixed if no intrusions are present. We additionally define the “purity” of a segment as one minus the number of intrusions divided by the size of the segment of interest. (c) Clutches have a high purity (top) and are unmixed roughly half of the time (middle). Choosing a stricter insulation score threshold would have resulted in bigger (bottom) and more mixed clutches (middle). (d) The purities for segments of various sizes centered on each nucleosome. (e) The fraction of unmixed segments decays with increasing segment size. The window size at which the fraction unmixed is 0.5 is about the size of clutches.

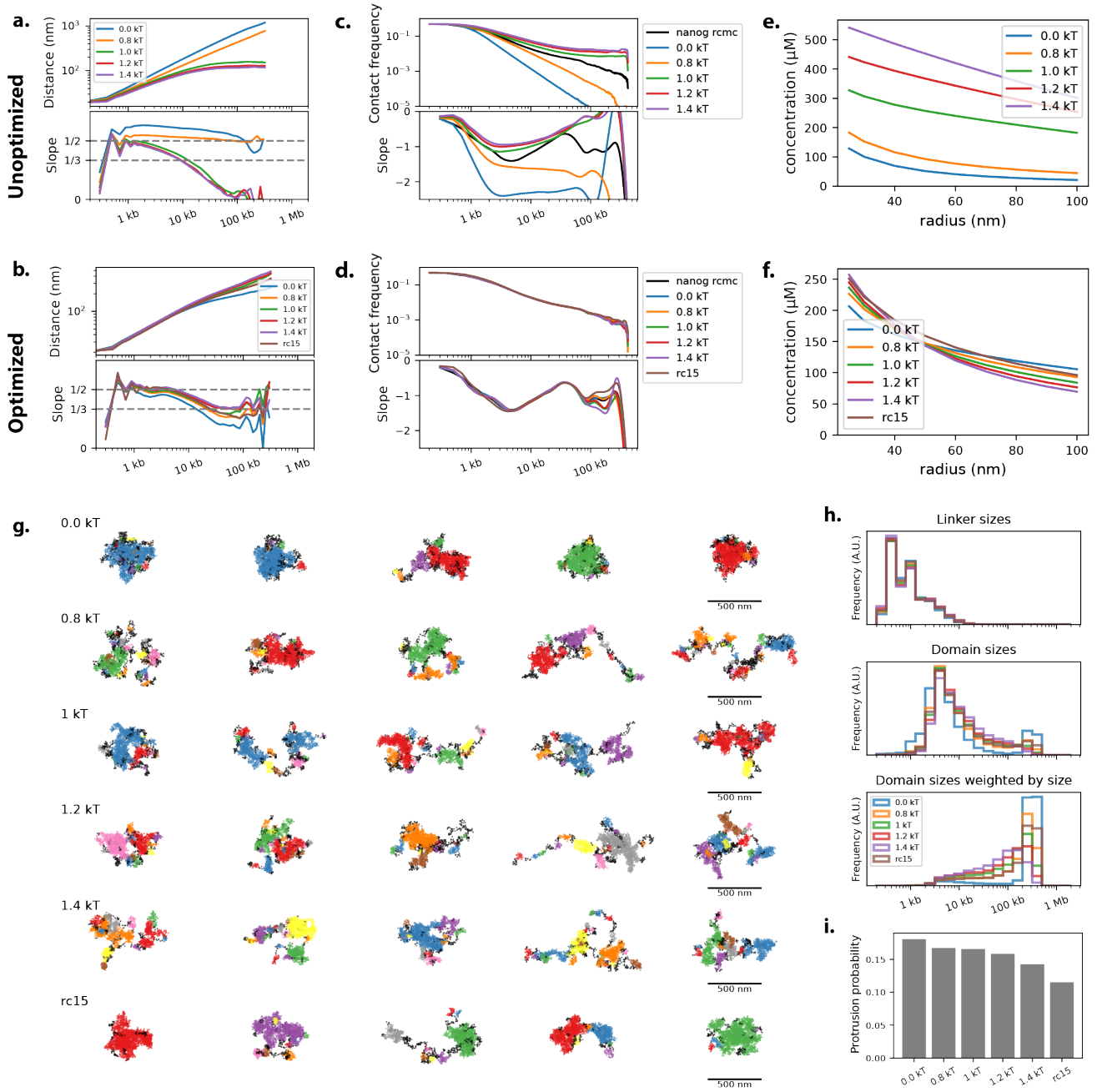


Fig. S12: Altering the base nucleosome interaction energy changes scaling properties. (a) Segment end-to-end distances for the base model with a range of interaction energies. The theta condition is achieved with an interaction energy between 0.8 kT and 1.0 kT. (b) Equivalent for the optimized *Nanog* system. The distances scale faster with higher interaction energies, but this trend begins to plateau with an interaction energy of 1.0 kT or higher. In the “rc15” simulations, we directly optimize the short range nucleosome interactions instead of the interaction potentials based on the contact indicator function. It proved challenging to completely reproduce the reference data with this strategy, but nonetheless it provides comparable values for our observables of interest as our standard simulations. (c, d) Simulated $P(s)$ curves before and after optimization. (e, f) The average local nucleosome concentration, defined as the average concentration within the indicated radius of each nucleosome. Strong nucleosome interaction energies favor short range interactions, whereas weak interactions favor long range interactions. (g) Representative simulation frames colored by packing domain assignments for different nucleosome interaction energies. Protrusions are shown in black. To our eyes, conformations are qualitatively similar in all but the 0.0 kT condition. (h) Size of domains and protrusions for different interaction energies. (i) Nucleosome interactions have minor impacts on protrusion probabilities.

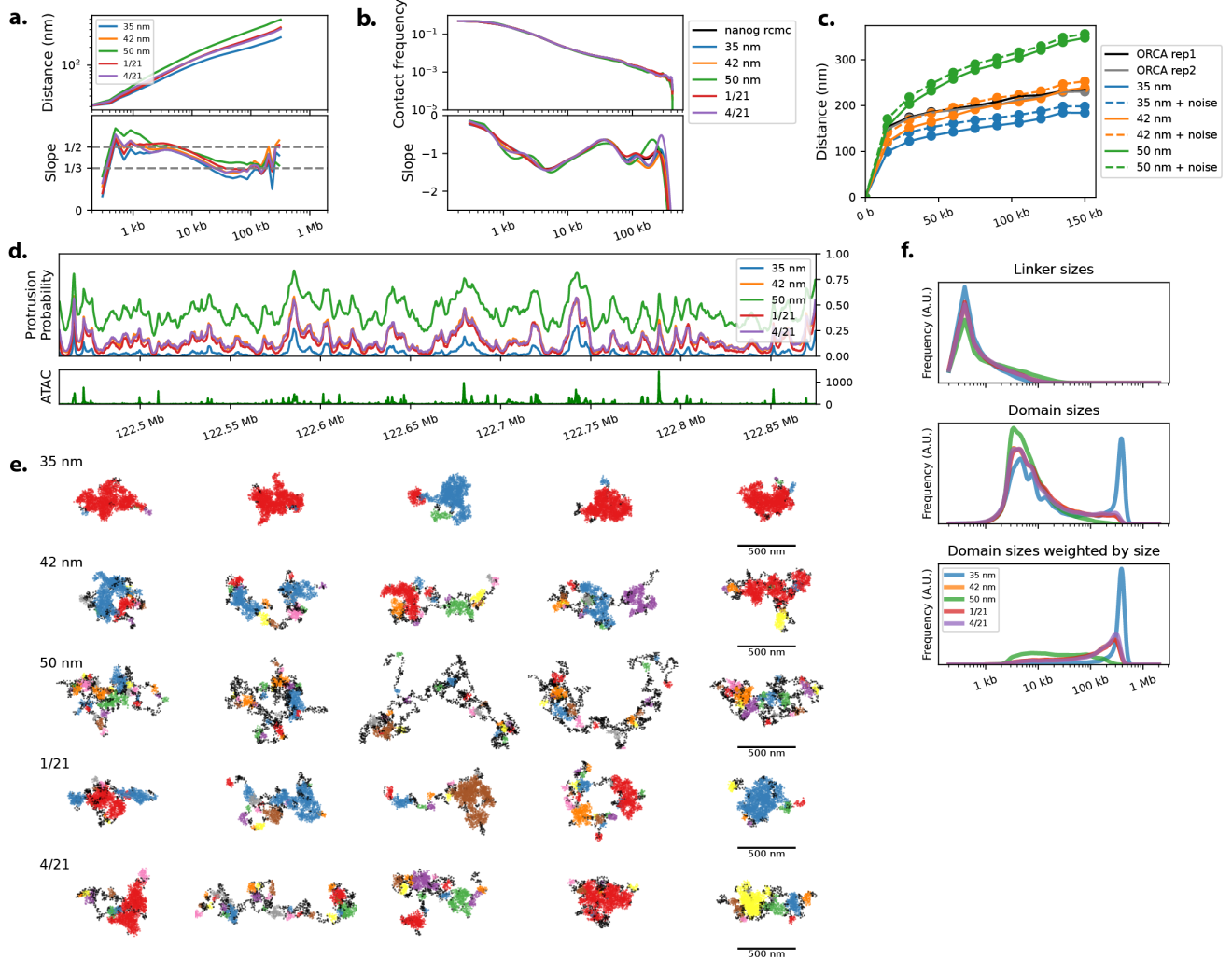


Fig. S13: Condensed domains and protrusions are present well beyond the range of contact indicators supported by chromatin tracing data. (a) Segment end-to-end distances for the optimized *Nanog* system with a range of capture radii and slope values ($\sigma = 1/21$ and $\sigma = 4/21$). The results in the main text all use a capture radius of 42 nm. (b) Simulated $P(s)$ curves after optimization, each computed using the indicated contact indicator. This is included to show that the optimizations are converged. (c) Comparison to chromatin tracing data (ORCA) shows that a capture radius of 50 nm (35 nm) results in a system that is on average too large (small). (d) Increasing the capture radius generally increases the probability of a position occurring in a protrusion. (e) Representative simulation frames colored by packing domain assignments for different nucleosome interaction energies. Protrusions are shown in black. (f) Size of domains and protrusions for different capture radii.

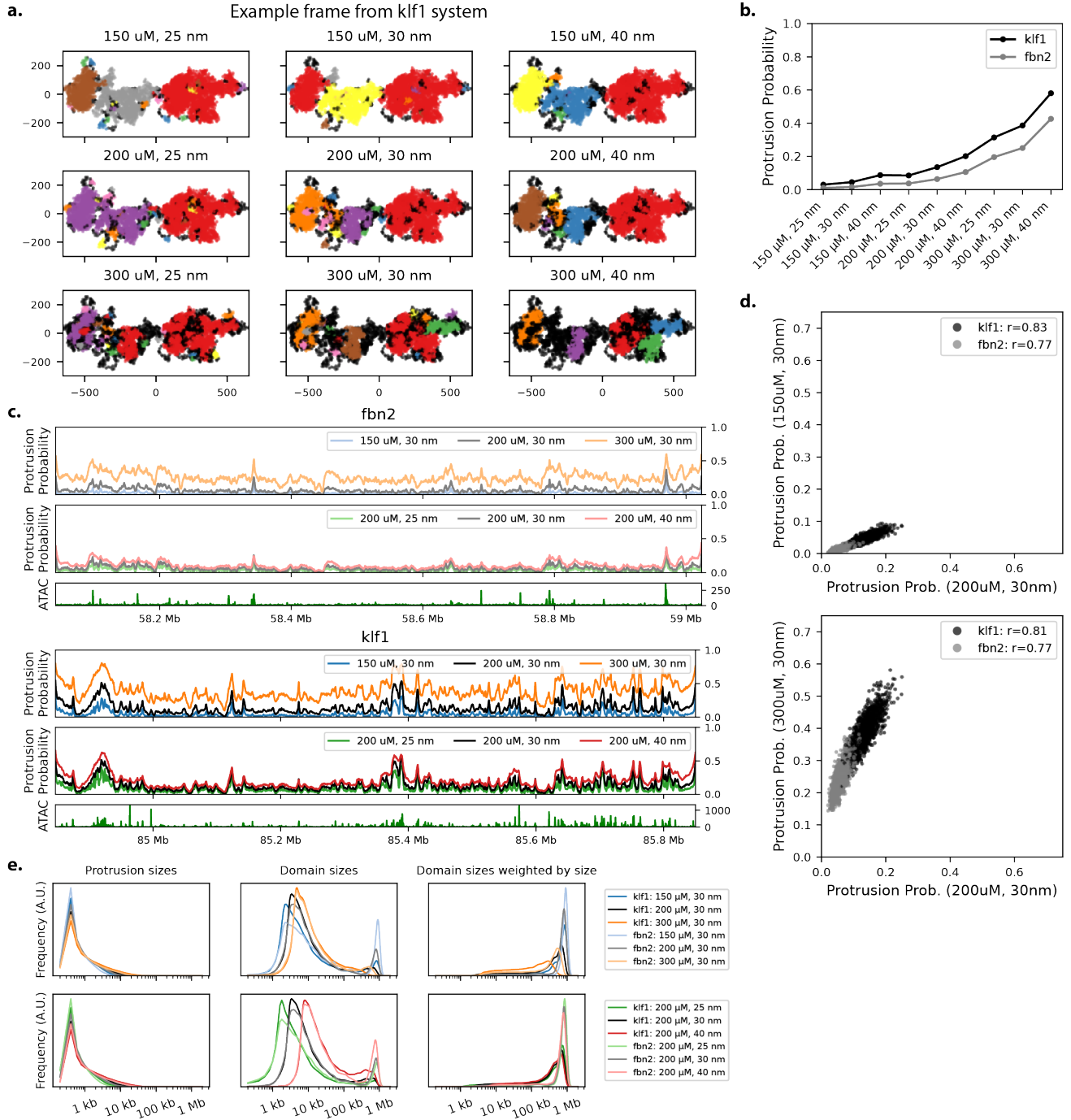


Fig. S14: Sensitivity of conclusions to DBSCAN parameters. (a) An example simulation frame showing DBSCAN clustering results with a range of parameters. The center panel (200 μ M, 30 nm) shows the clustering used in the main text figures. (b) Average protrusion probabilities for the *klf1* (euchromatin) and *fbn2* (heterochromatin) for a range of parameters. Increasing the concentration or radius increases the protrusion probability. Importantly, these changes are proportional between the two regions. (c) Protrusion probabilities as a function of genomic position for a range of concentrations (top panels) and range of radii (middle panels). (d) Direct comparison of protrusion probabilities with different parameters. The legend shows the Spearman correlation coefficient. (e) Protrusion and domain sizes for a range of concentrations (top) and range of radii (bottom).

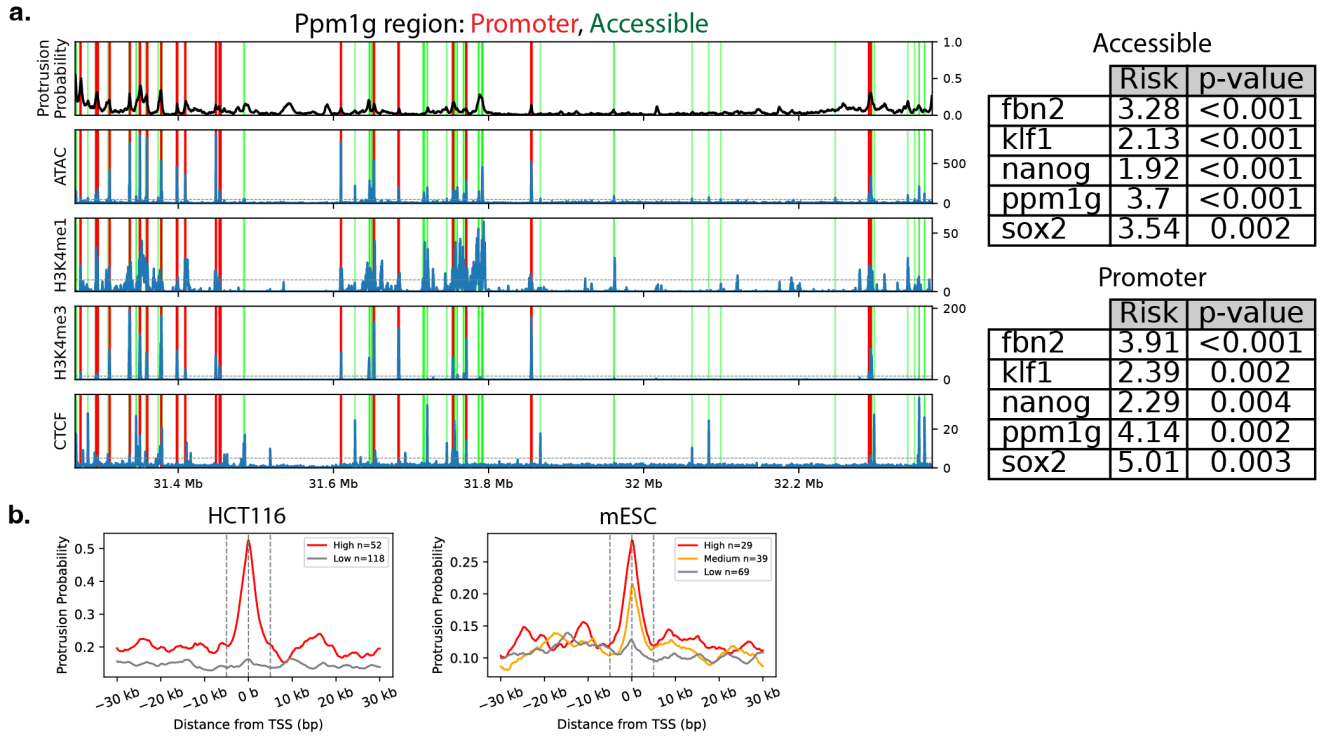


Fig. S15: Enrichment of regulatory elements in protrusions is statistically significant. (a) The left panel shows an example region showing protrusion probabilities and several epigenetic tracks. Vertical red lines indicate active promoters and green lines show accessible regions. Some green lines are obscured by red lines. The horizontal lines in the ATAC and H3K4me3 indicate the thresholds used to define accessible sites and active promoters, respectively. The right panel shows risk ratios and p-values for the two sets of features being enriched for protrusions. (b) The mean protrusion probability centered on transcription start sites (TSS) with different levels of gene expression aggregated over all genes in the regions studied. For HCT116 cells, genes were defined using GRCh38 GENCODE V29 and the first listed transcript was selected. Gene expression values were obtained from ENCFF435PHM and considered high for transcript per million (TPM) > 2 and otherwise low. For mESCs, genes were defined using ENSEMBL GRCm39 release 115 and the first listed transcript was selected. Gene expression values were obtained from GSE123636, and considered high for counts per million (CPM) > 50, medium for $1 < \text{CPM} < 50$, and low for $\text{CPM} < 1$.