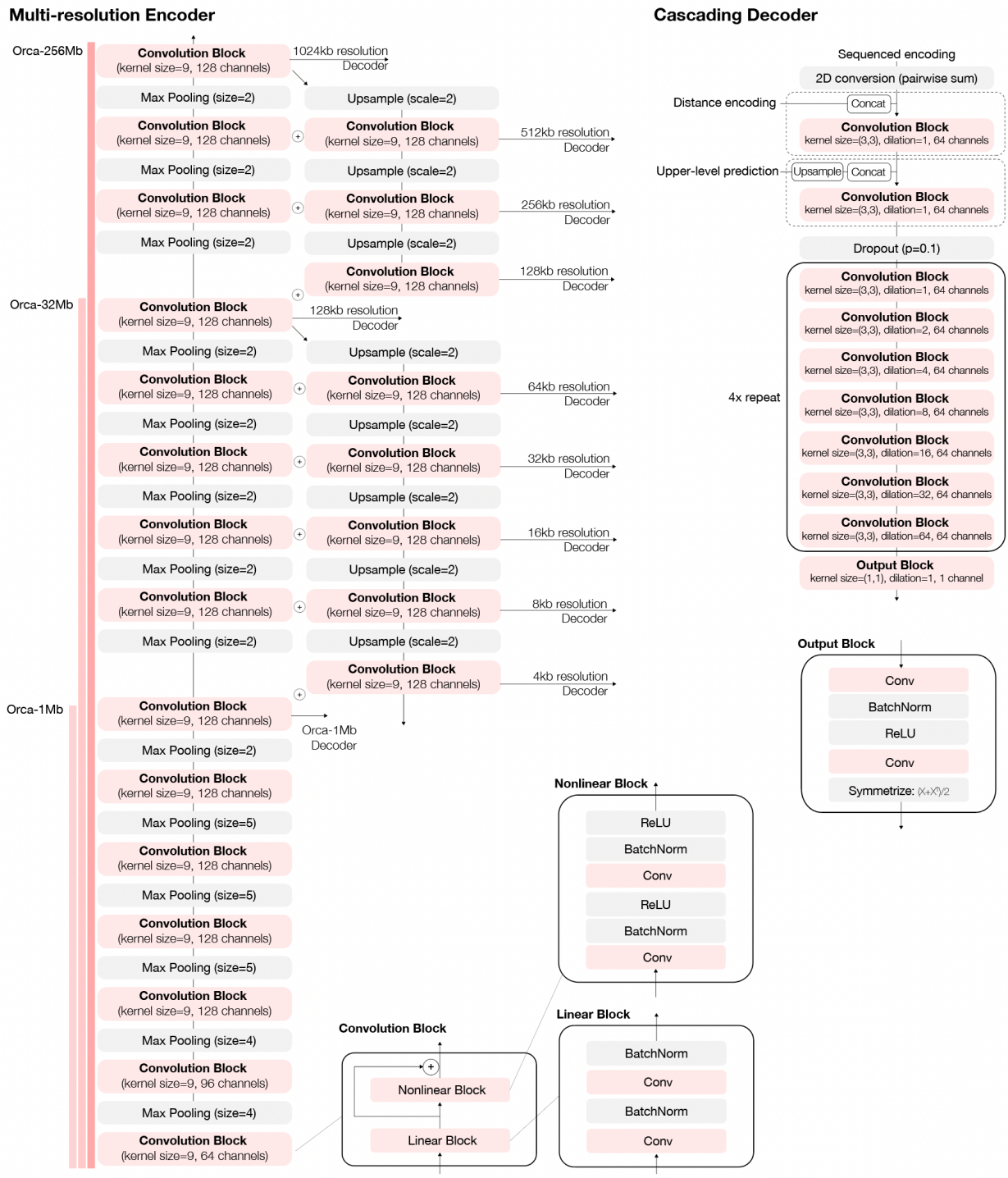
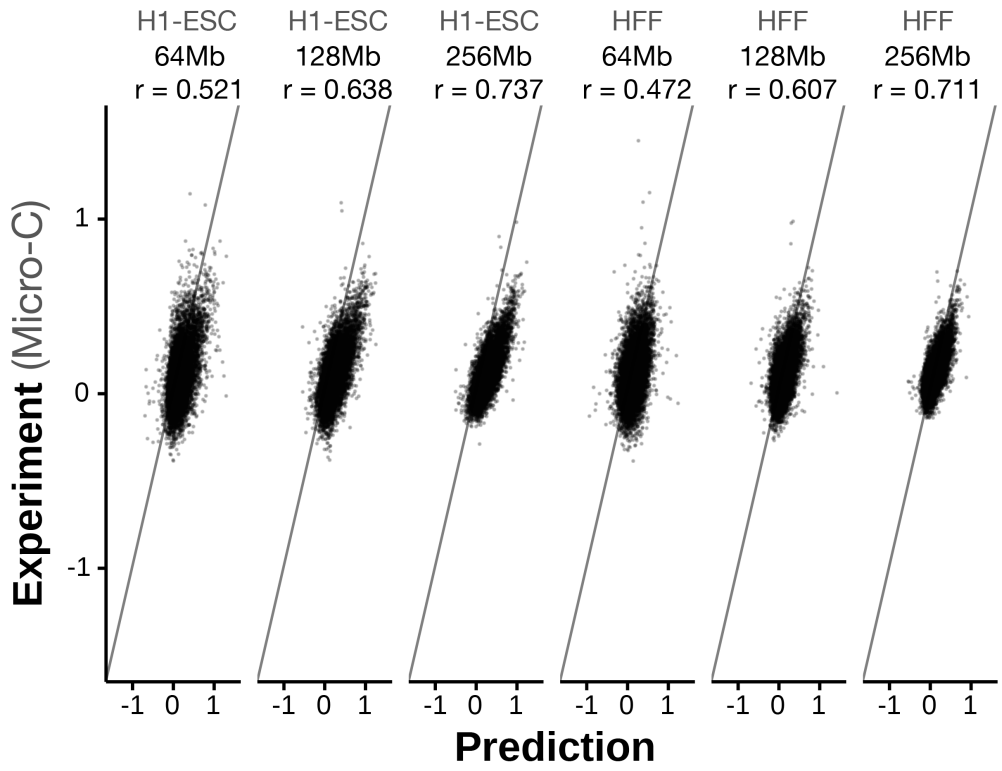
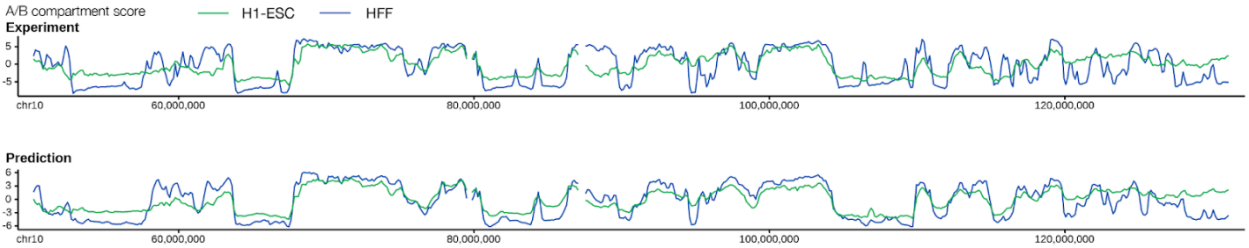


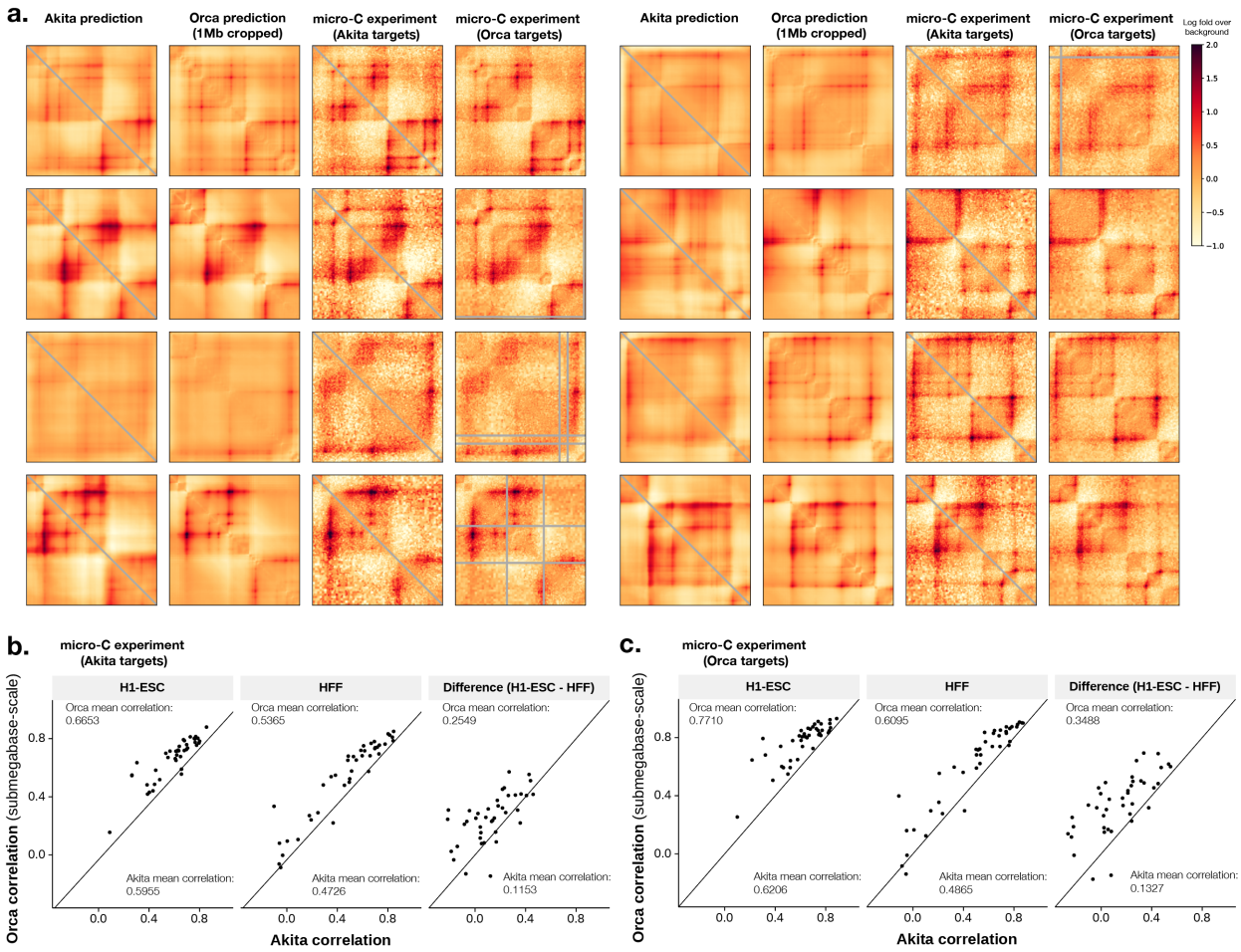
Supplementary Figure 1. The deep learning model architecture of Orca. The Orca architecture consists of a multi-resolution encoder and a series of cascading decoders. The multi-resolution encoder outputs sequence encodings at multiple resolutions and can be trained in three stages (Orca-1Mb, Orca-32Mb, and Orca-256Mb). Each decoder takes the encoder output at a specific resolution and the prediction from the upper-level decoder as input. The basic building block of both the encoder and the decoders is the Sei-style convolution blocks with dual linear and nonlinear path design.



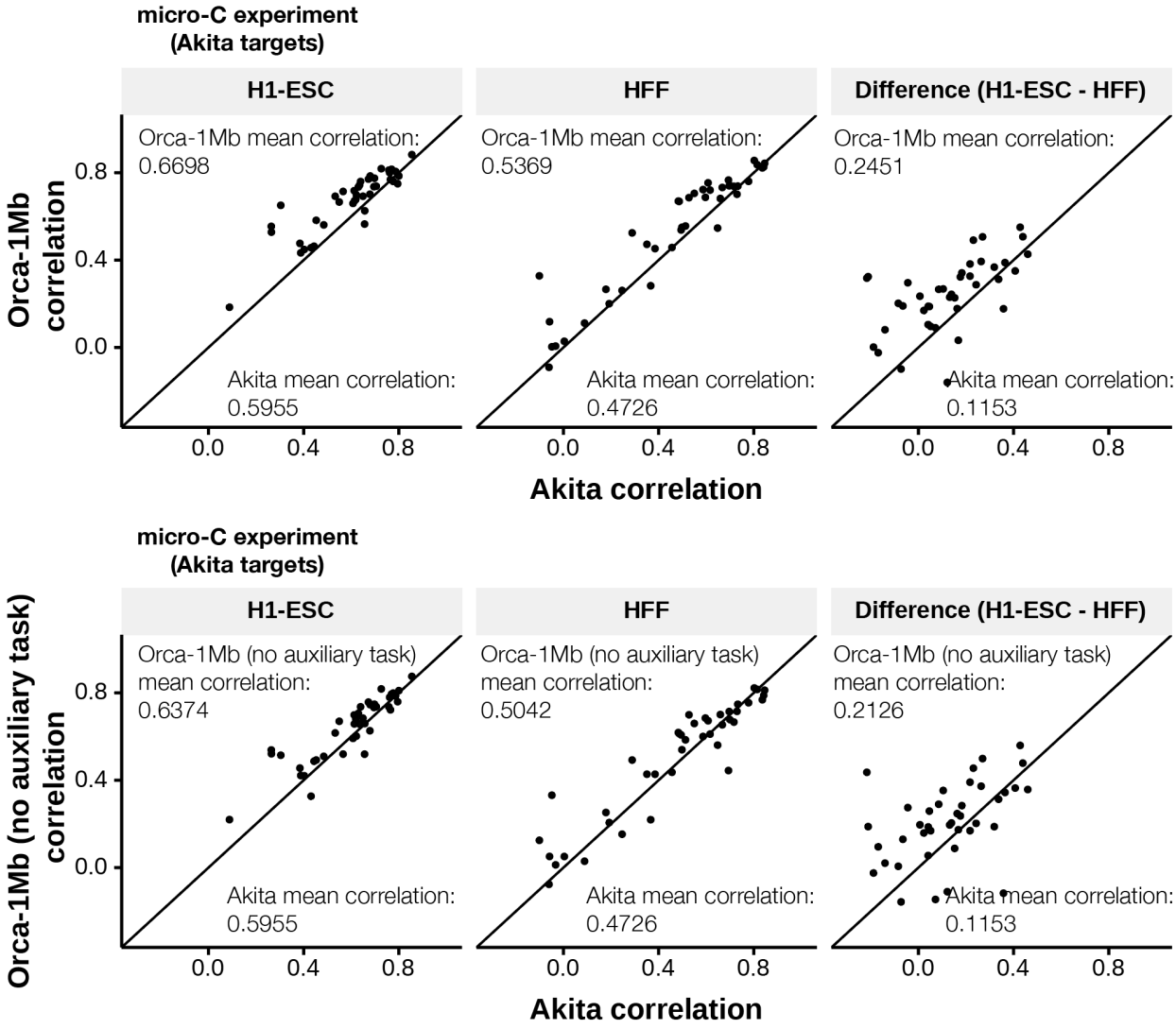
Supplementary Figure 2. Performance of interchromosomal predictions. Scatter plot comparison of the predicted interchromosomal interaction scores with the micro-C measured interaction scores (log fold over background) for the holdout test chromosomes. 10,000 randomly subsampled scores are shown in each panel. The overall Pearson correlations are annotated. The genome interactions are represented by the log fold over background scores for both prediction and experimental data.



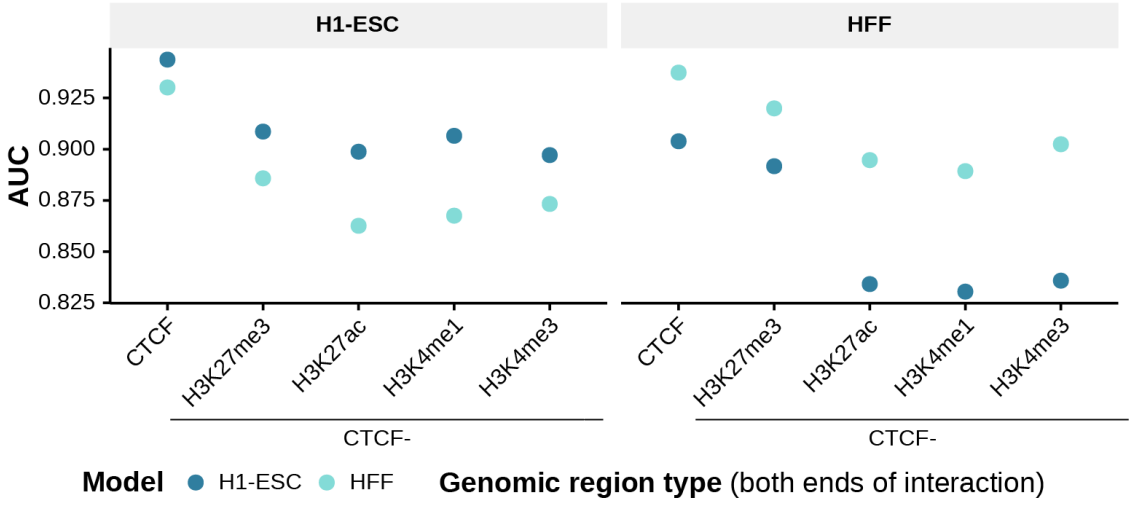
Supplementary Figure 3. Comparison of A/B compartment score for H1-ESC and HFF cell types with Orca prediction-derived A/B compartment scores in an example genomic region. Observed(top) and predicted (bottom) chromatin A/B compartment scores as computed by the first principal component (PC) of the interaction matrix (high score indicates A compartment). Predicted chromatin compartments are derived from 32Mb level predictions with 3.2Mb step size.

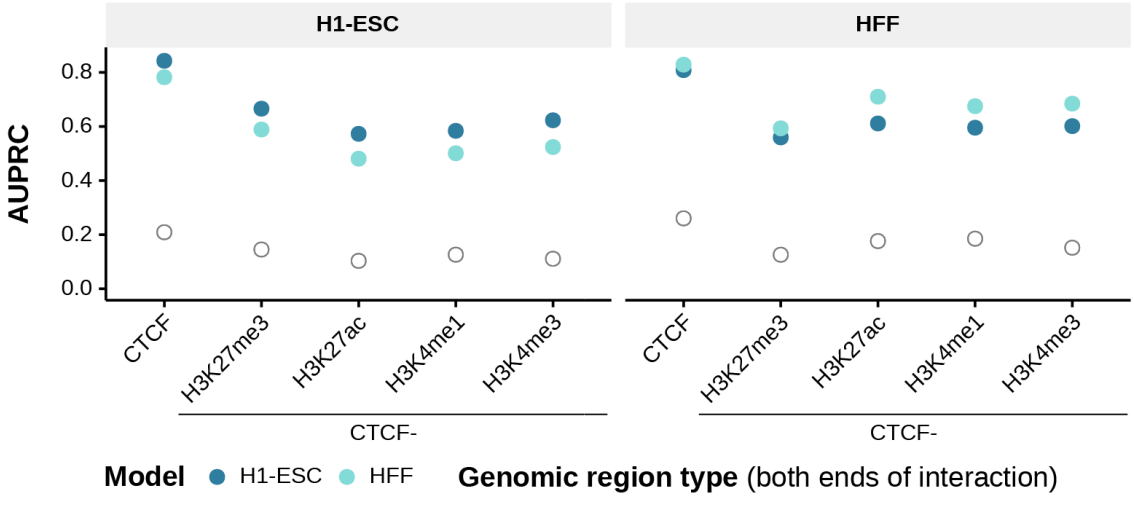


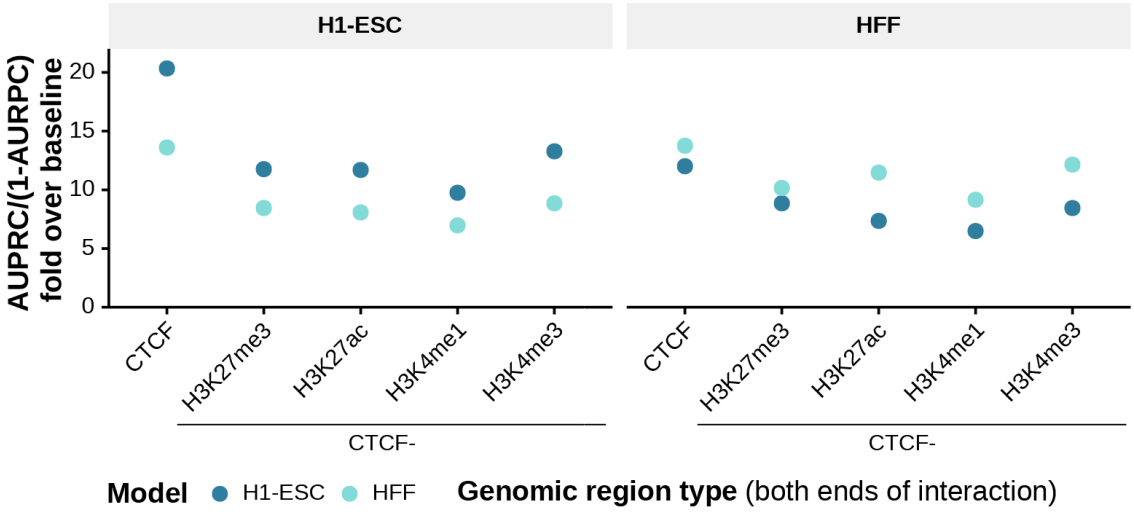
Supplementary Figure 4. Performance comparison with Akita on submegabase-level predictions. a). Randomly selected submegabase-scale prediction examples for H1-ESC from the test set of Akita that are also located in Orca test chromosomes. The Orca-1Mb prediction and experimental matrices are cropped to match the Akita output range. For any prediction or target matrix, each score was subtracted by the average scores at the same distance in the same matrix to facilitate comparison across different preprocessed datasets. b-c). Scatter plot comparisons of performance evaluated on test set samples from Akita that are also located in Orca test chromosomes 9 and 10. Comparison based on Akita targets (b) and Orca targets (c) separately are shown. For both comparisons, we also performed rescaling, cropping, and Gaussian smoothing for Orca predictions and targets to match Akita preprocessing. The scatter plots compared submegabase-scale predictions for H1-ESC, HFF, and the difference between the two cell types. Pearson correlations are computed after the background subtraction described in (a) for both prediction and target to be more robust to different preprocessing.



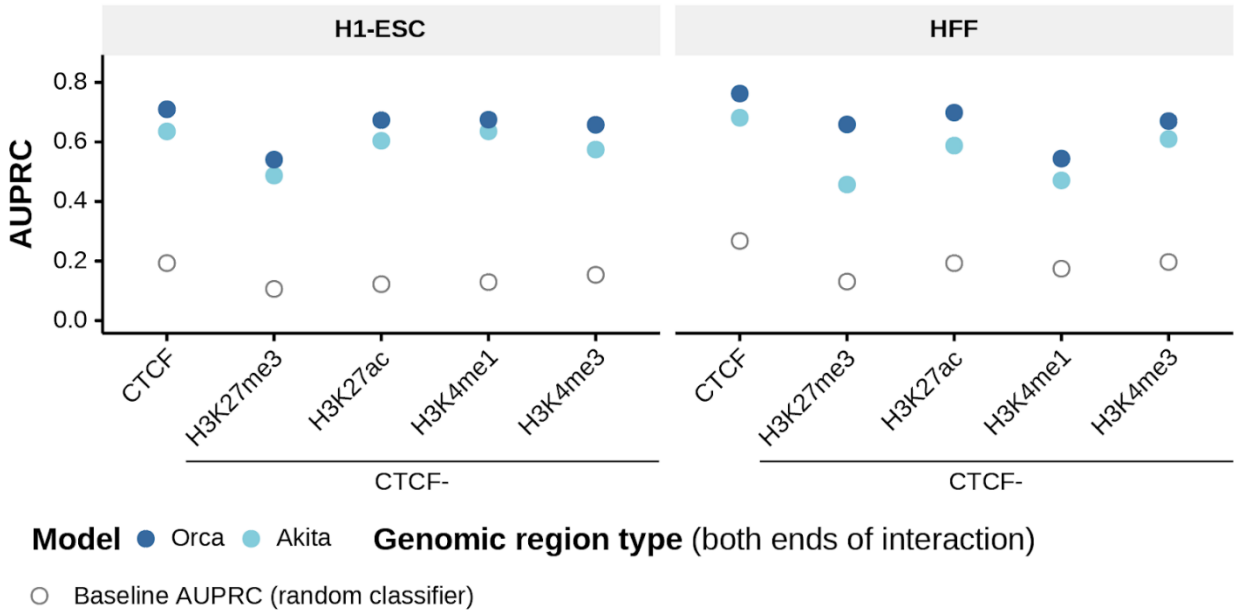
Supplementary Figure 5. Performance comparisons of ablated Orca models on common Akita and Orca test set at submegabase-scale. Scatter plot comparisons of performance evaluated on all test set samples from Akita that are also located in Orca test chromosomes. Evaluations are performed separately on the Akita target data and the Orca target data of this shared test set. The scatter plots compared submegabase-scale predictions for H1-ESC, HFF, and the difference between the two cell types. Pearson correlations are computed after background subtraction for both prediction and target to be more robust to different preprocessing. Specifically for any prediction or target matrix, each score was subtracted by the average scores at the same distance in the same matrix before computing correlation. For Orca dataset and predictions, we also performed rescaling, cropping, and Gaussian smoothing to match the Akita preprocessing.


H1-ESC
HFF
AUC
0.925
0.900
0.875
0.850
0.825
CTCF
H3K27me3
H3K27ac
H3K4me1
H3K4me3
CTCF-
Model
H1-ESC
HFF
Genomic region type (both ends of interaction)

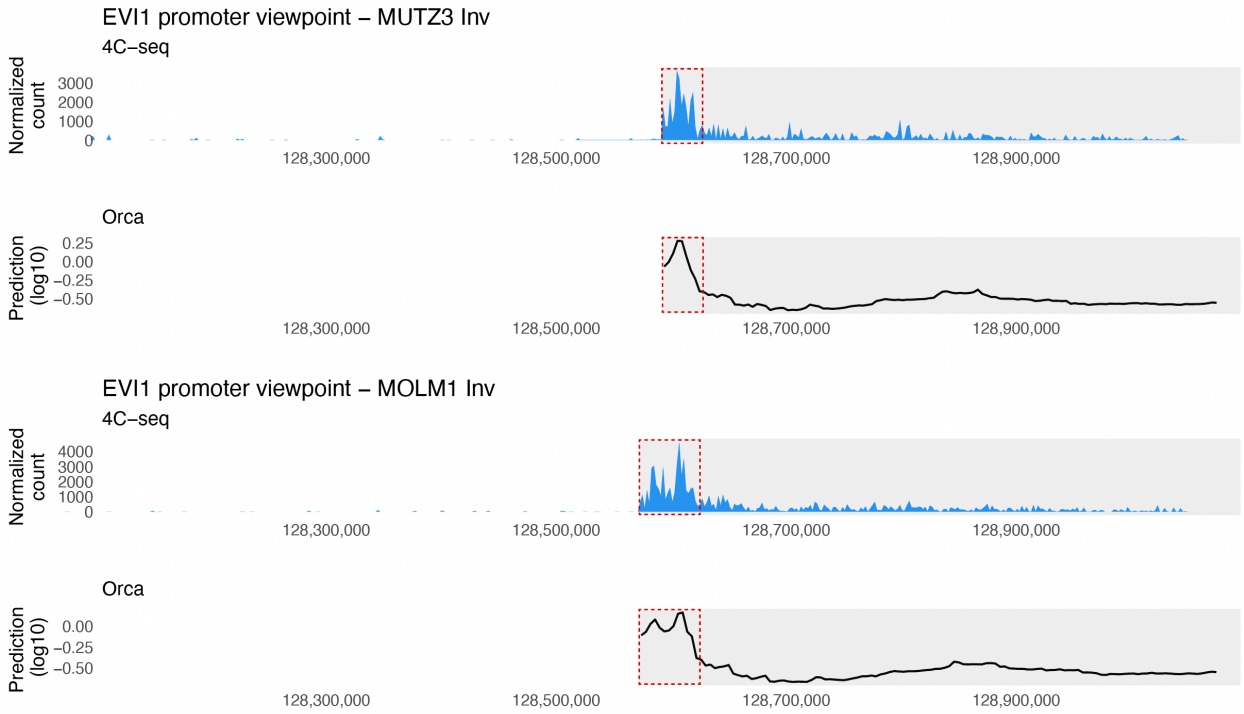

H1-ESC
HFF
AUPRC
0.8
0.6
0.4
0.2
0.0
CTCF
H3K27me3
H3K27ac
H3K4me1
H3K4me3
CTCF-
Model
H1-ESC
HFF
Genomic region type (both ends of interaction)


H1-ESC
HFF
AUPRC/(1-AURPC)
fold over baseline
20
15
10
5
0
CTCF
H3K27me3
H3K27ac
H3K4me1
H3K4me3
CTCF-
Model
H1-ESC
HFF
Genomic region type (both ends of interaction)

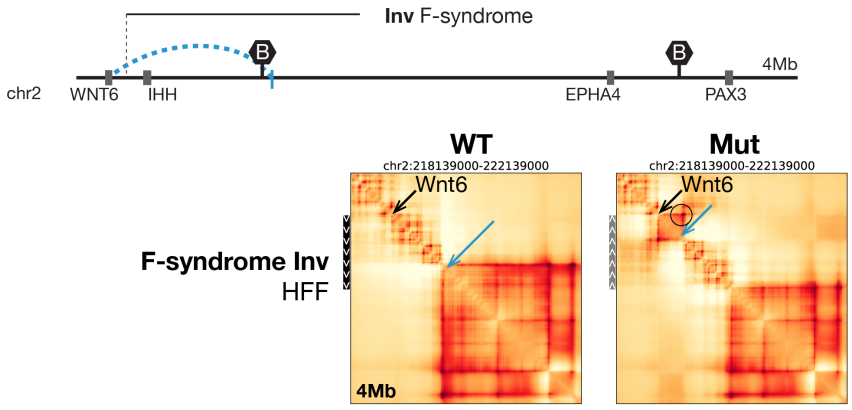
Supplementary Figure 6. Orca model performance in predicting genome interactions in different genomic region types. For each genomic region type (CTCF, H3K27me3, H3K27ac, H3K4me1, H3K4me3; the last 4 region types are also filtered to CTCF-negative regions, as indicated by 'CTCF-'), we tested the model performance in predicting genome interactions (labeled with a threshold of 90 percentile log fold over background genome interaction scores) at 1Mb level with both sides located in the genomic region type, against non-interaction pairs from any region type (top panel, with AUROC) or from the same region type (mid and bottom panel, with AUPRC), on the holdout test chromosomes. The genomic region type is annotated based on >95 percentile signal scores and non-CTCF types are additionally filtered with <90 percentile CTCF binding scores. For the middle panel, the proportion of positives is shown in gray circles as the baseline of AUPRC. The signal scores are obtained from ENCODE accession ids ENCFF473IZV, ENCFF623ZAW, ENCFF423TVA, ENCFF584AVI, ENCFF912ZUR, ENCFF761RHS, ENCFF442WNT, ENCFF078JZB, ENCFF449DEA, and ENCFF027GWJ.



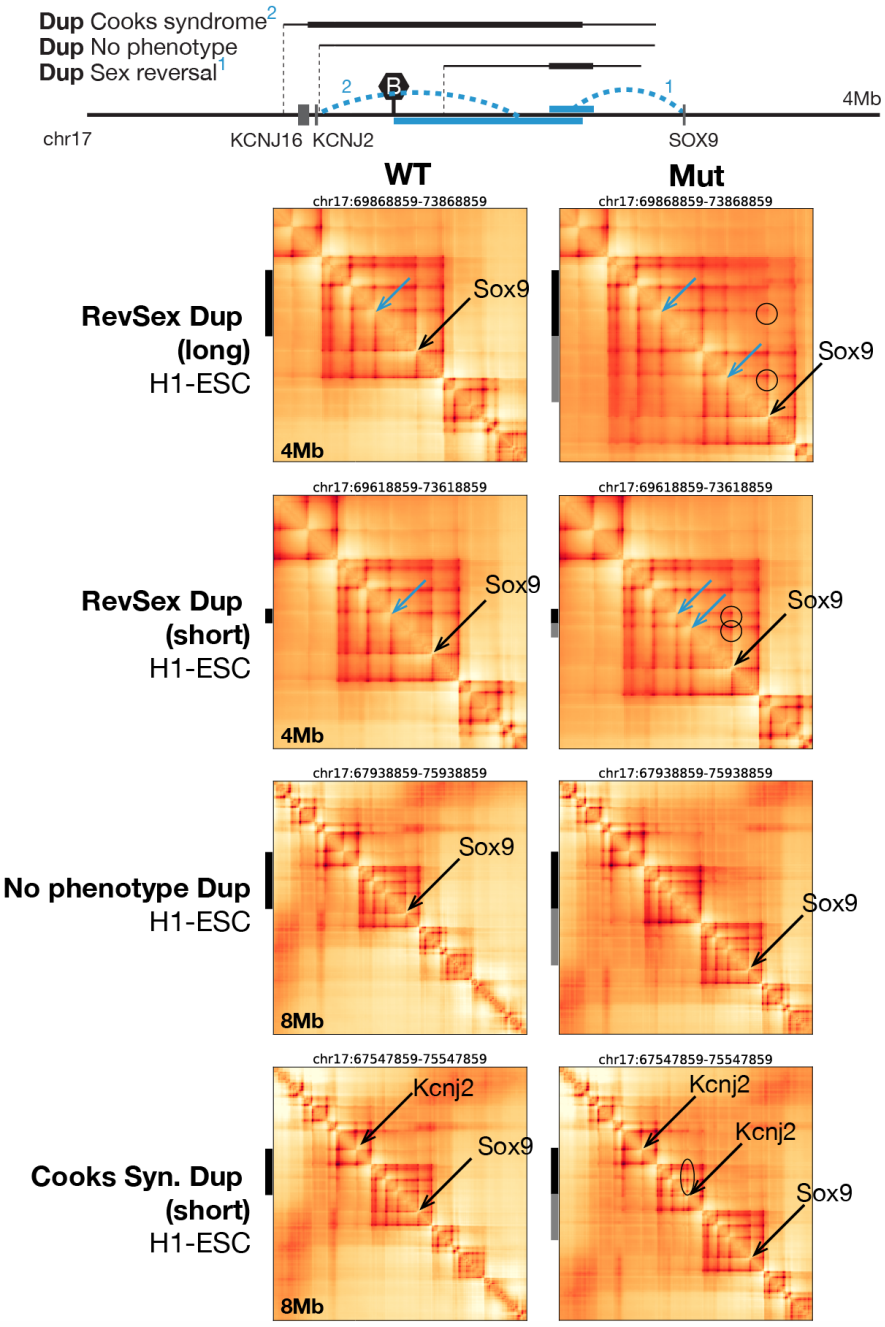
Supplementary Figure 7. Submegabase-scale area under precision recall curve (AUPRC) comparison with Akita for predicting genome interactions in different genomic region types. The baseline AUPRC, or the proportion of positives is shown in gray circles. For each genomic region type (CTCF, H3K27me3, H3K27ac, H3K4me1, H3K4me3; the last 4 region types are also filtered to CTCF-negative regions, as indicated by 'CTCF-'), we tested the model performance in predicting genome interactions (labeled with a threshold of 90 percentile log fold over background genome interaction scores) at 1Mb level with both sides located in the genomic region type, against non-interaction pairs from the same region type, on the holdout test chromosomes. The genomic region type is annotated based on >95 percentile signal scores and non-CTCF types are additionally filtered with <90 percentile CTCF binding scores.



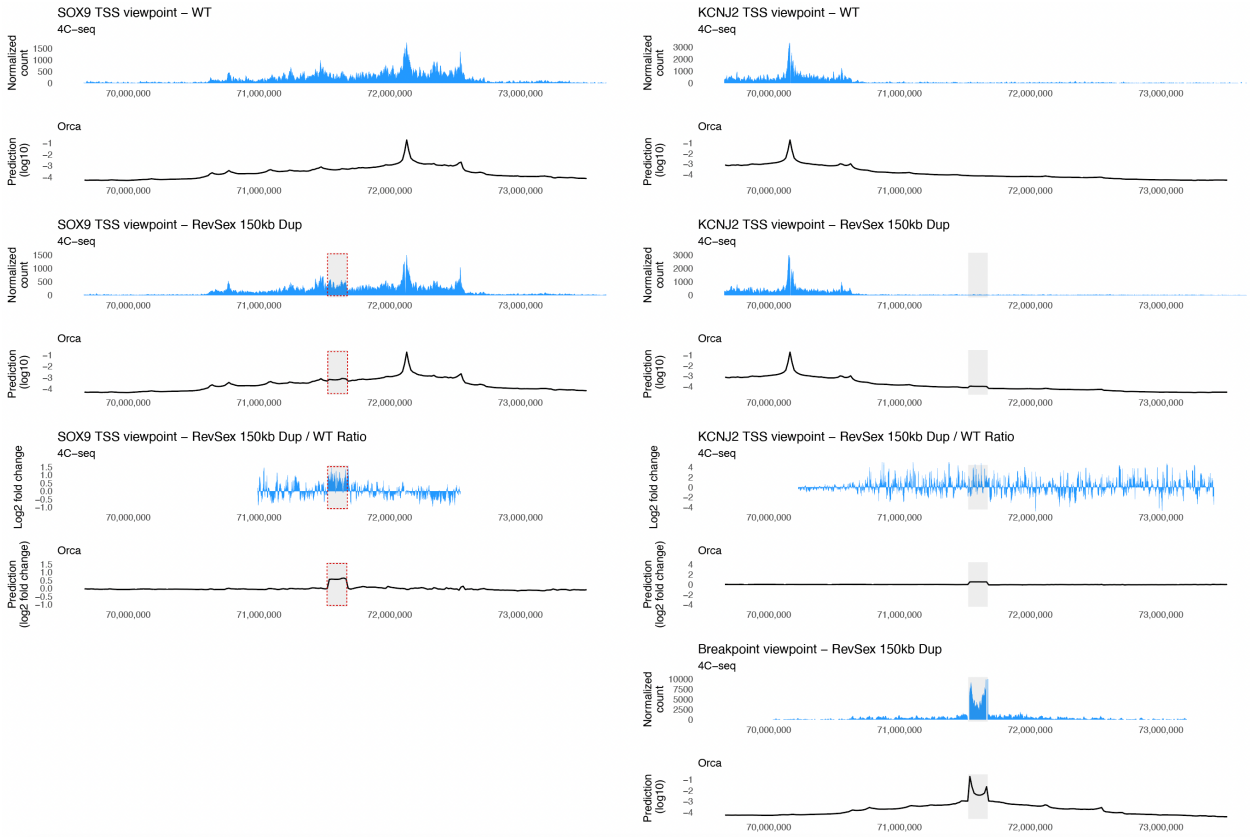
Supplementary Figure 8. Comparison of Orca prediction with 4C-seq experimental measurement for inversion variants near EVI1 in Leukemia from Gröschel et al. 2014. The normalized counts from 4C-seq and the log10 predicted interaction scores (log fold over background) at the 4C-seq point-of-view are shown. The observed and predicted gain of interaction sites relevant to the phenotype is highlighted with the red dashed line box.



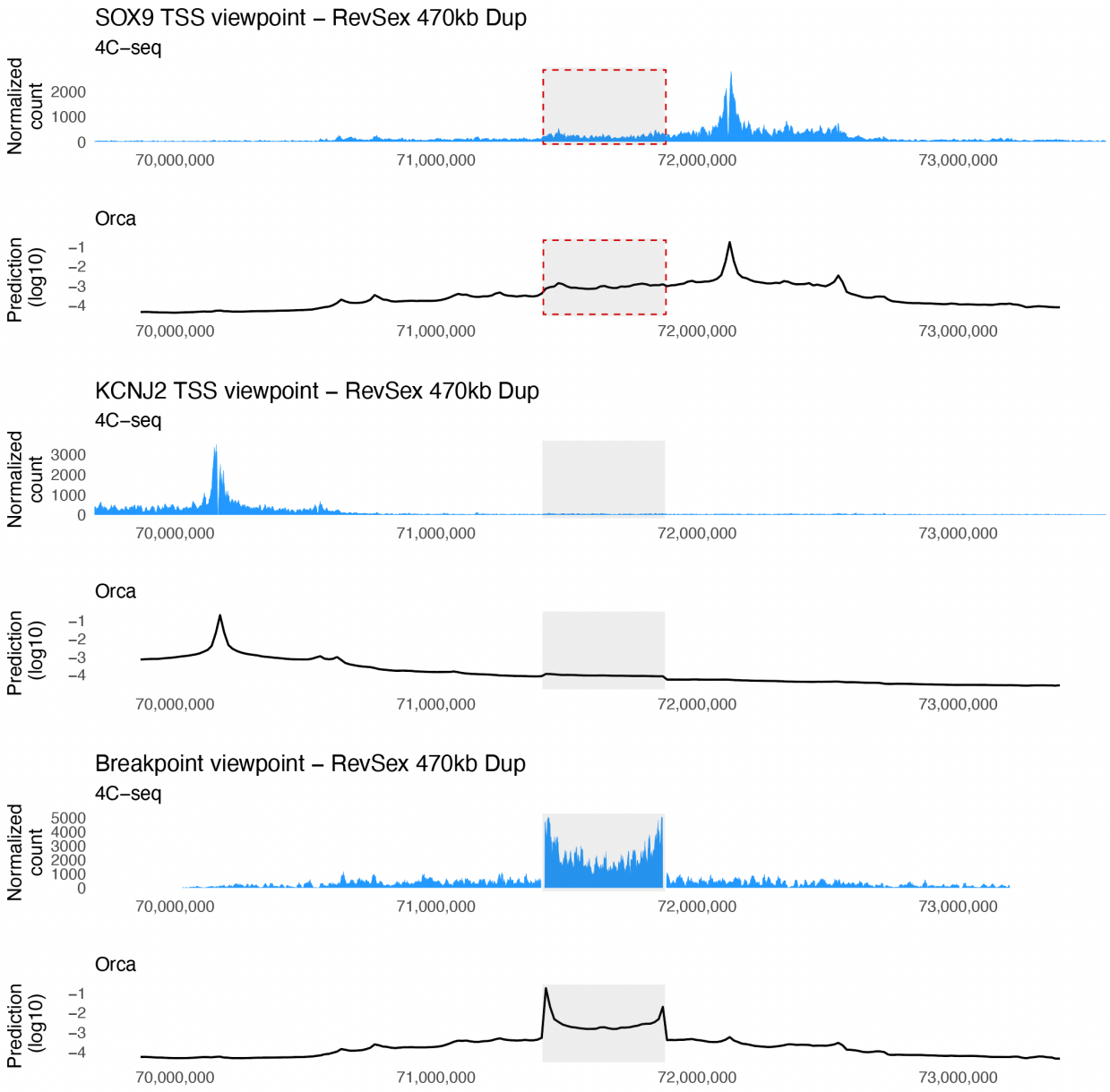
Supplementary Figure 9. Prediction of F-syndrome Inversion variant outcome in WNT6-PAX3 region. The Orca genome interaction predictions are represented by the log fold over genomic-distance-based background scores. Positions of the major genes affected by the structural variants are indicated by black arrows and known enhancer regions involved are indicated by blue arrows. Black and gray bars on the left side indicate genomic intervals involved in the structural variants pre and post mutation. Full multiscale prediction results for both H1-ESC and HFF cell types as well as micro-C observations in the cell types are included in Supplementary Data 3, and validations results for all structural variants are summarized in Supplementary Table 2.



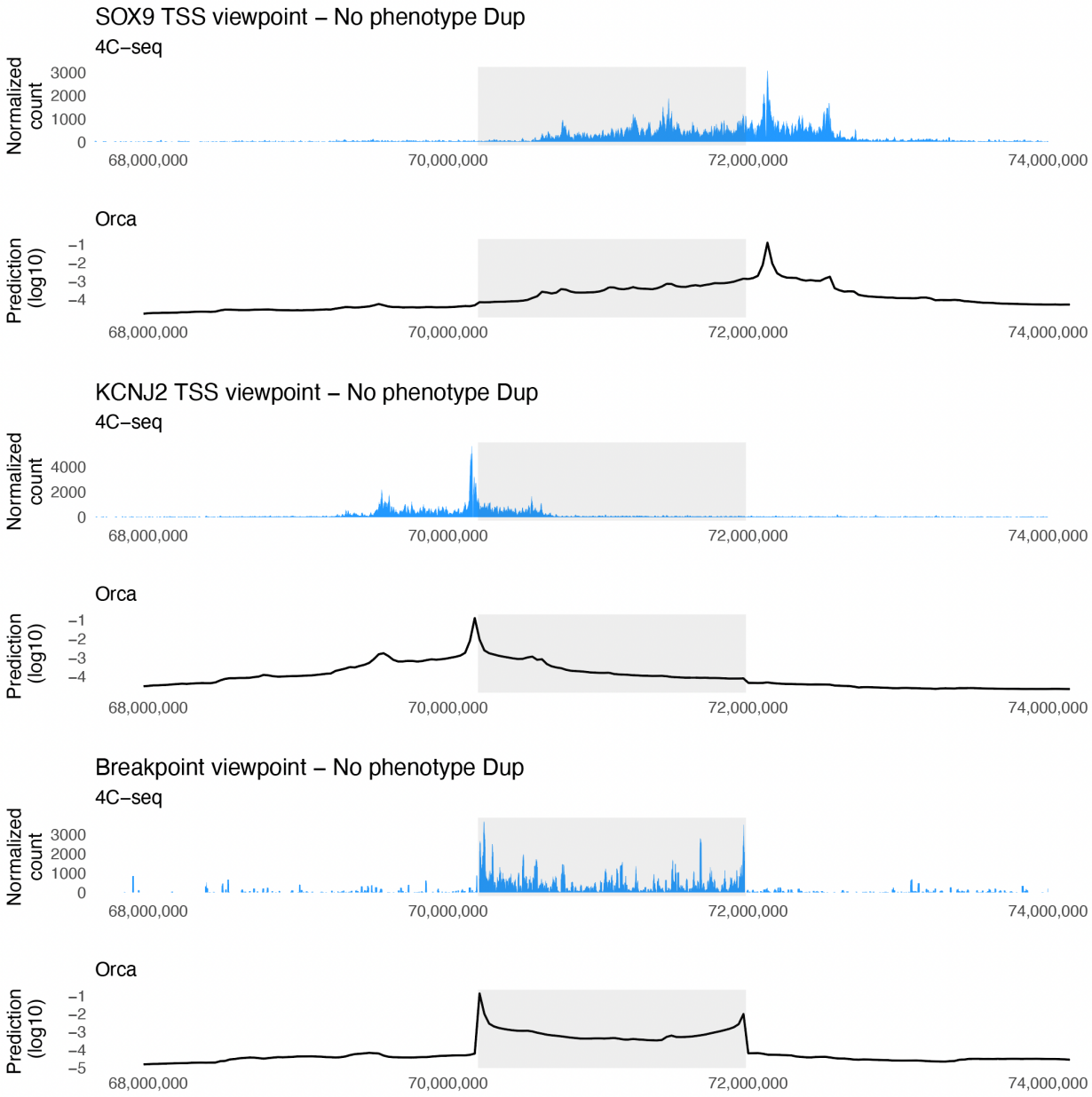
Supplementary Figure 10. Orca predictions of multiple structural variants with complex phenotypic outcomes in KCNJ2-SOX9 region. The Orca genome interaction predictions are represented by the log fold over genomic-distance-based background scores. Positions of the major genes affected by the structural variants are indicated by black arrows and known enhancer regions involved are indicated by blue arrows. Ectopic interactions caused by the variants are indicated by circles. Black and gray bars on the left side indicate genomic intervals involved in the structural variants pre and post mutation. Full multiscale prediction results for both H1-ESC and HFF cell types as well as micro-C observations in the cell types are included in Supplementary Data 3, and validations results for all structural variants are summarized in Supplementary Table 2.



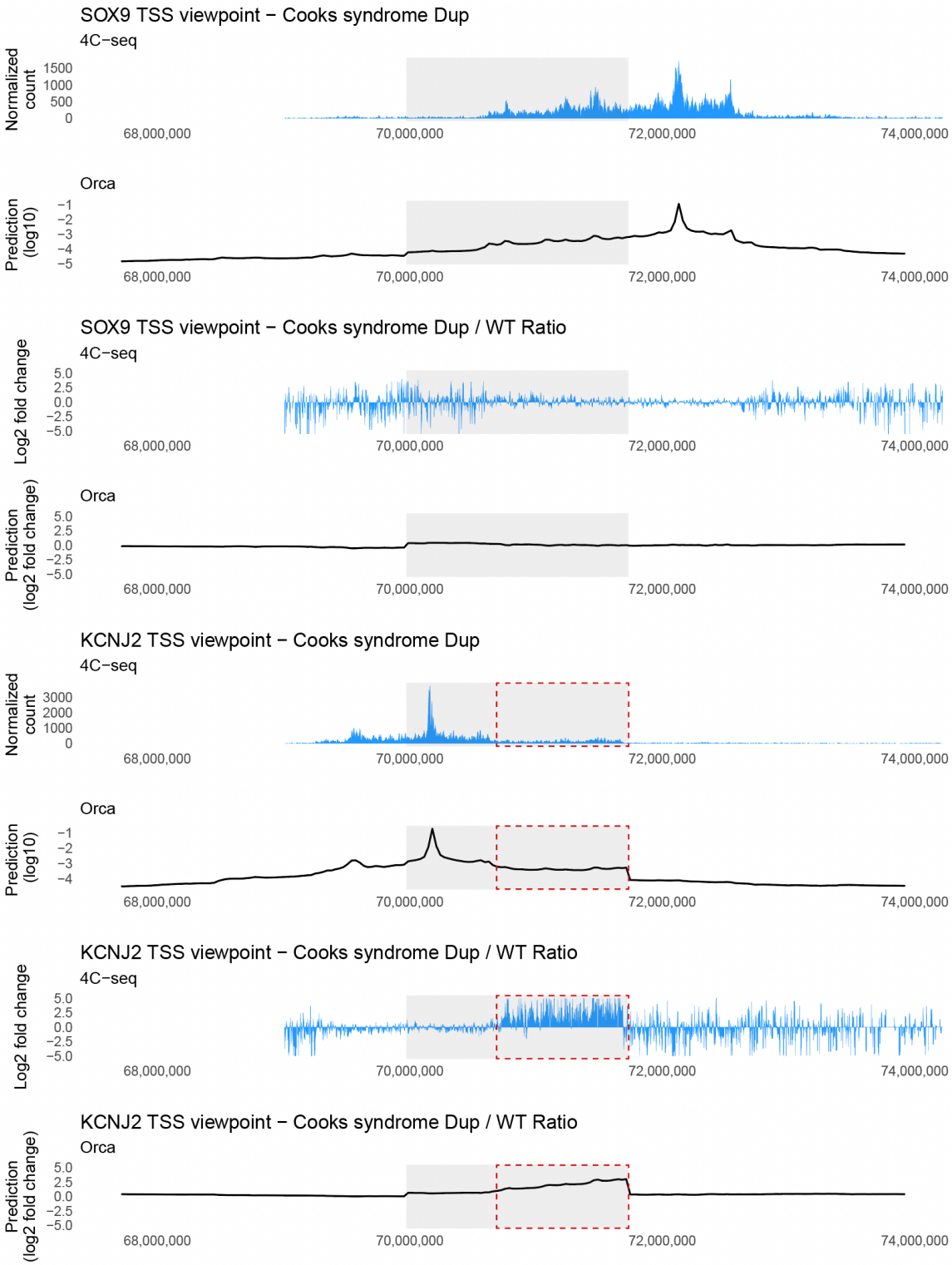
Supplementary Figure 11. Comparison of Orca prediction with 4C-seq experimental measurement for a RevSex duplication variant from Franke et al. 2016. The normalized counts from 4C-seq and the log10 predicted interaction scores (log fold over background) at the 4C-seq point-of-view are shown. The observed and predicted gain of interaction sites relevant to the phenotype is highlighted with the red dashed line box.



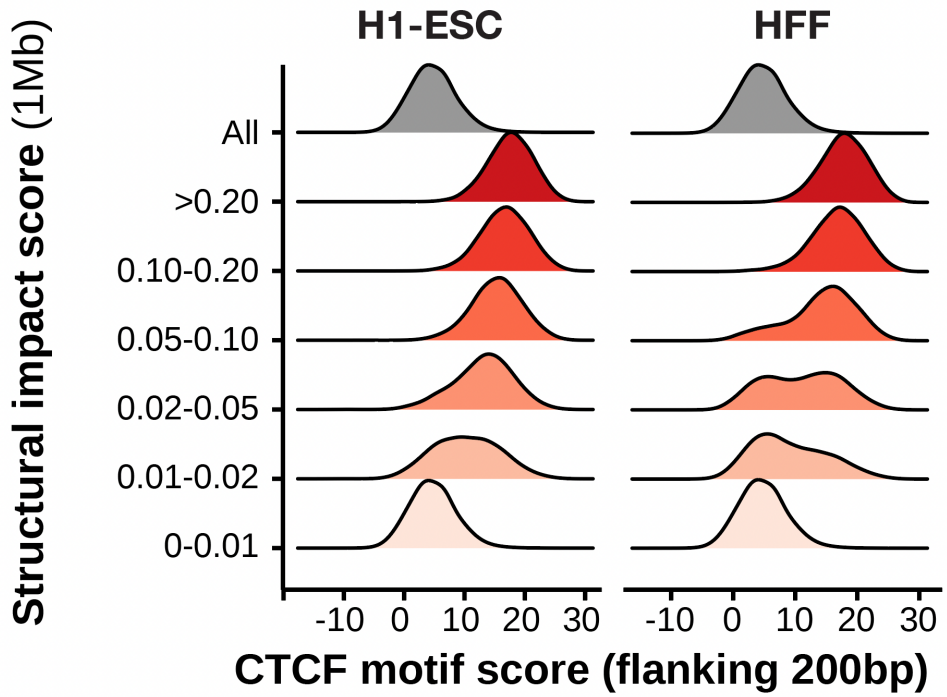
Supplementary Figure 12. Comparison of Orca prediction with 4C-seq experimental measurement for a long RevSex duplication (470kb) variant from Franke et al. 2016. The normalized counts from 4C-seq and the log10 predicted interaction scores (log fold over background) at the 4C-seq point-of-view are shown. The observed and predicted gain of interaction site relevant to the phenotype is highlighted with the red dashed line box.



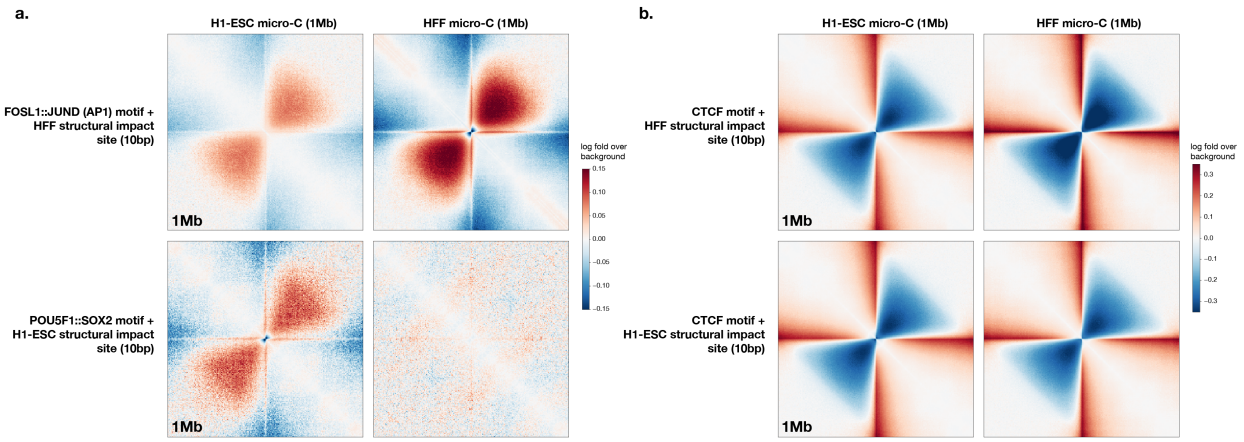
Supplementary Figure 13. Comparison of Orca prediction with 4C-seq experimental measurement for a "No phenotype" variant from Franke et al. 2016. The normalized counts from 4C-seq and the log10 predicted interaction scores (log fold over background) at the 4C-seq point-of-view are shown. The observed and predicted gain of interaction site relevant to the phenotype is highlighted with the red dashed line box.



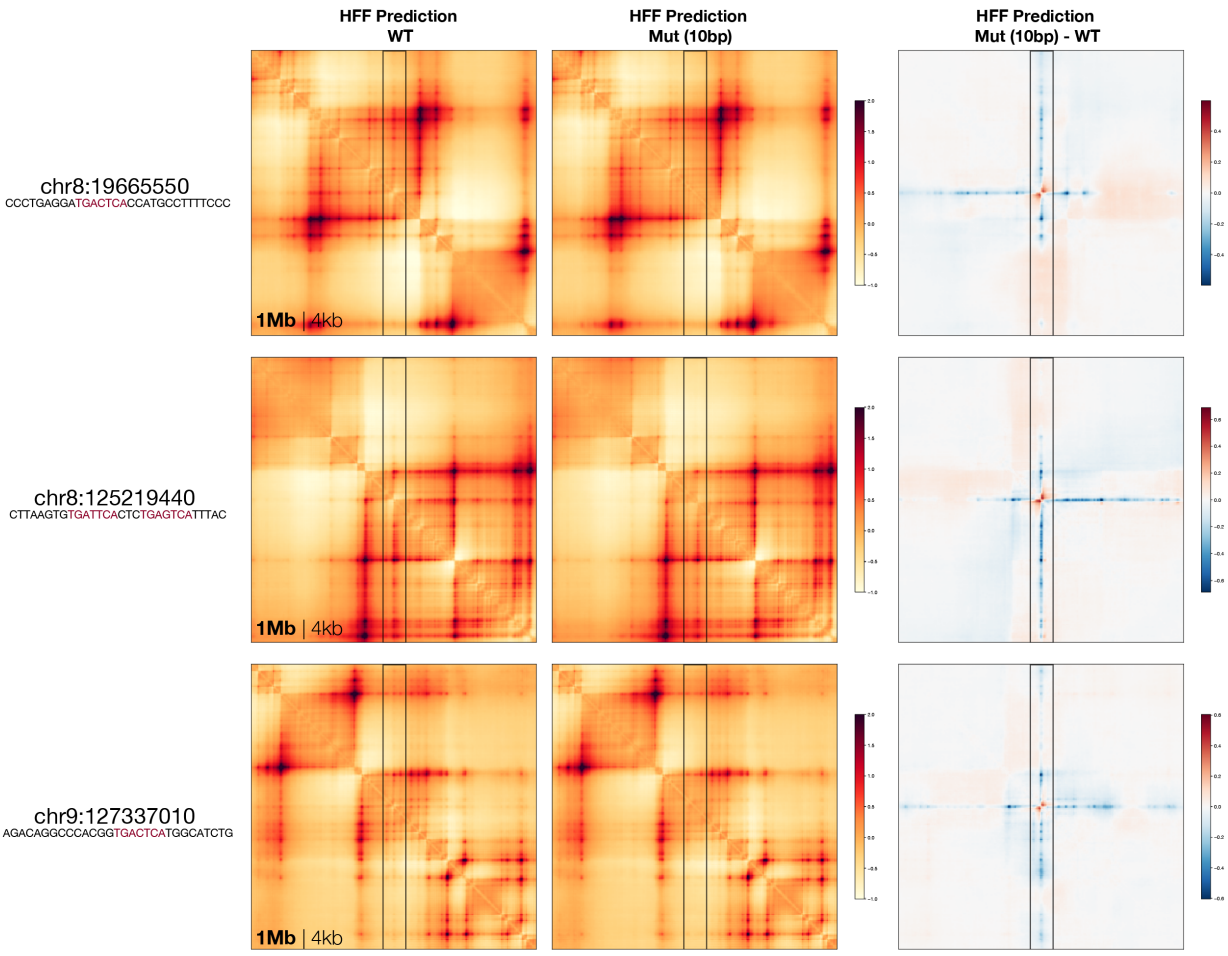
Supplementary Figure 14. Comparison of Orca prediction with 4C-seq experimental measurement for a Cooks syndrome variant from Franke et al. 2016. The normalized counts from 4C-seq and the log10 predicted interaction scores (log fold over background) at the 4C-seq point-of-view are shown. The observed and predicted gain of interaction site relevant to the phenotype is highlighted with the red dashed line box.



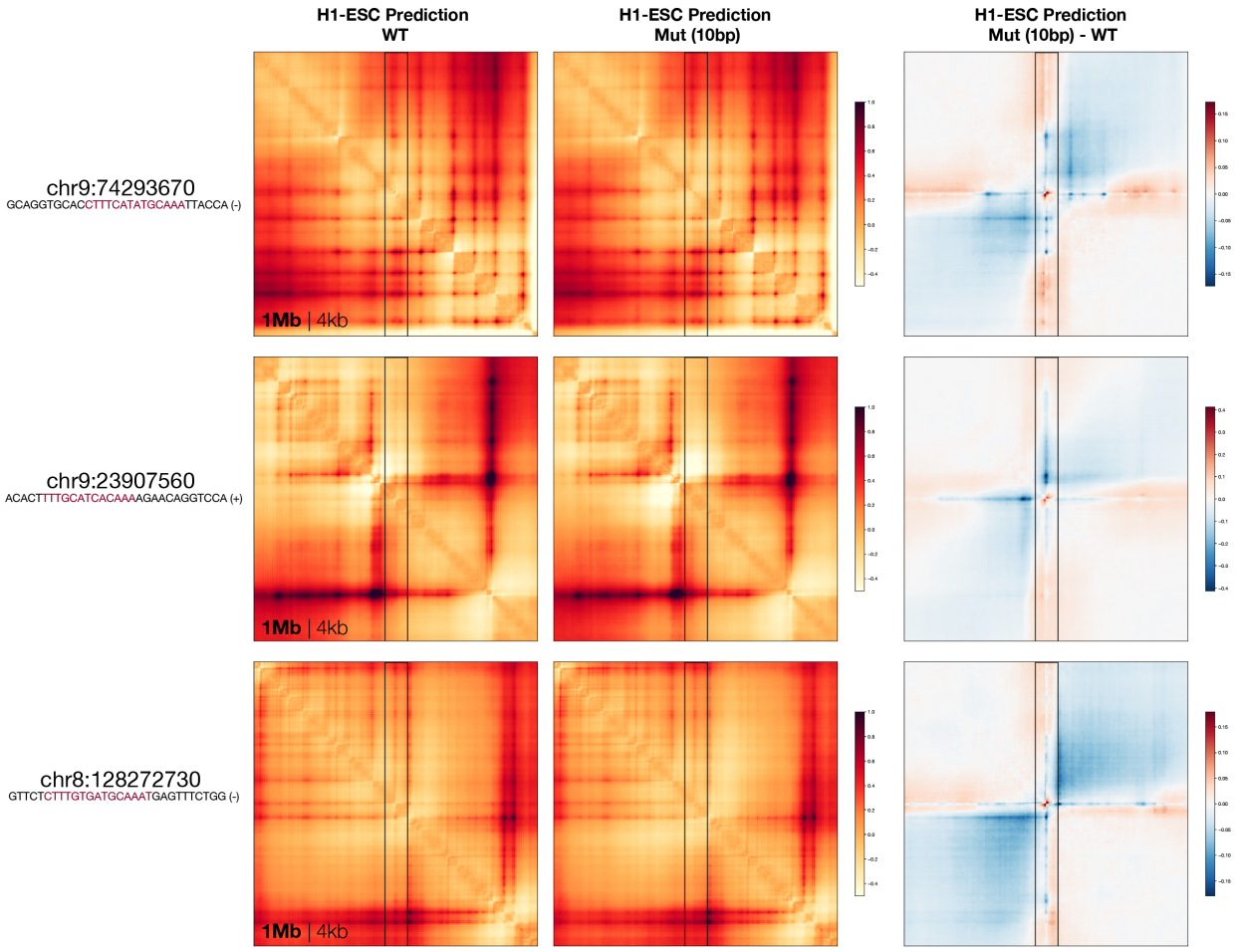
Supplementary Figure 15. CTCF motifs are observed within 200bp distance to most 10bp sequences with the strongest tier (>0.10) of structural impact score (1Mb). 10bp sequences are stratified by structural impact score (1Mb) for H1-ESC (left) and HFF (right). CTCF motif scores represent log odds scores.



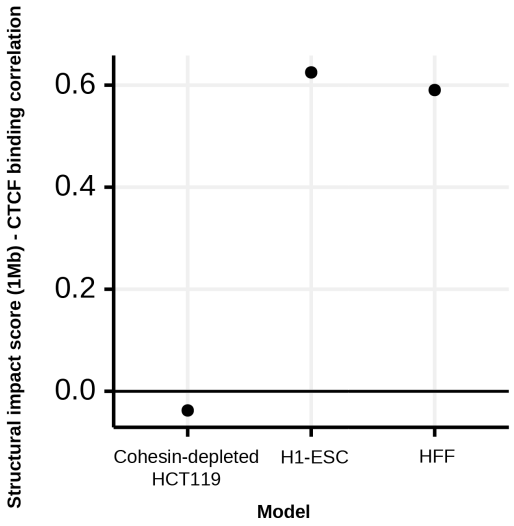
Supplementary Figure 16. Comparison of experimental H1-ESC and HFF micro-C genome interaction patterns at motif sites with predicted local structural impact. FOSL1::JUND and POU5F1::SOX2 structural impact sites showed strongly cell-type-specific interactions while the pile-up patterns of CTCF sites are largely cell-type-invariant. Average log fold over background micro-C contact matrices (1Mb) centered at the motif positions (log odds > 10) with predicted structural impact. The structural impact score (1Mb) threshold is > 0.02 for FOSL1::JUND or OCT4::SOX2 motifs (a) and >0.1 for CTCF motifs (b).



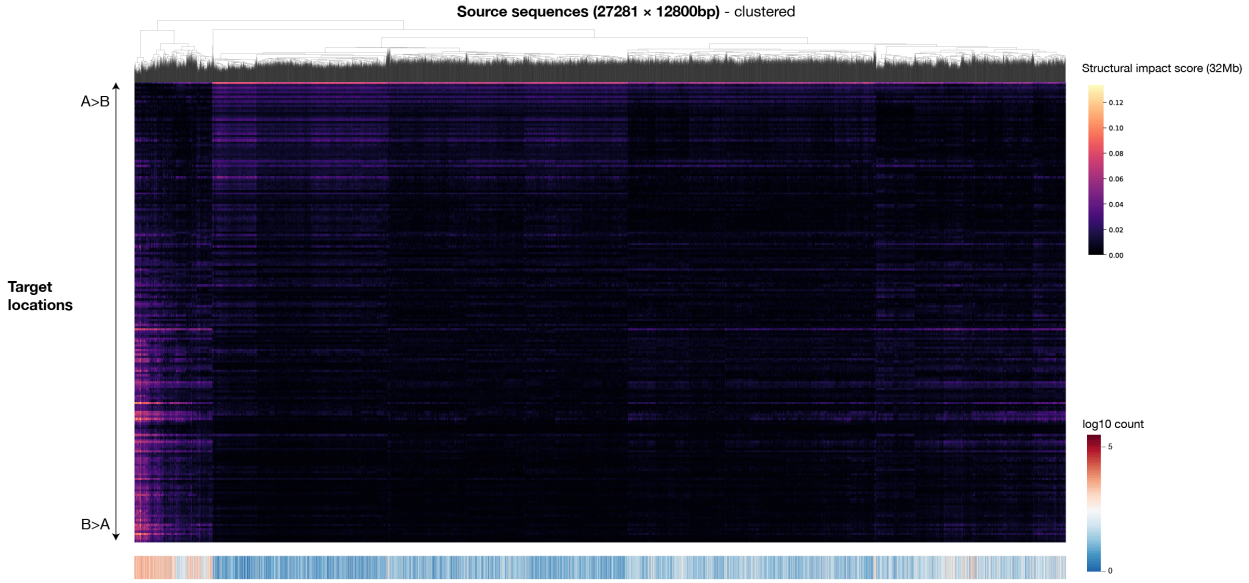
Supplementary Figure 17. Examples of predicted genomic interaction disruption caused by FOS::JUN motif mutation. The start position of each 10bp mutation and the mutated sequence including the 10bp flanking sequence are shown on the left. The FOS::JUN motif sequence is indicated in red. The predictions shown are for 1Mb region with 4kb resolution.



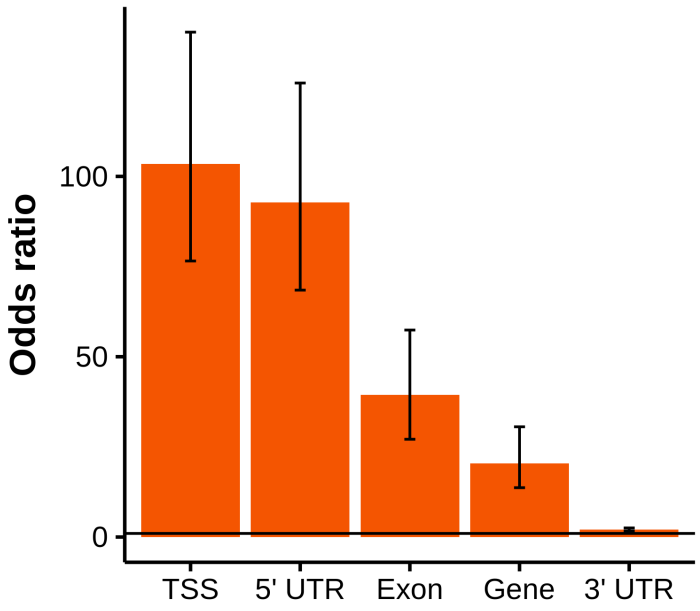
Supplementary Figure 18. Examples of predicted genomic interaction disruption caused by POU5F1::SOX2 motif mutation. The start position of each 10bp mutation and the original sequence including 10bp flanking sequence are shown on the left. The POU5F1::SOX2 motif sequence is indicated in red and the motif directionality is indicated by '+' and '-' signs. The predictions shown are for 1Mb region with 4kb resolution.



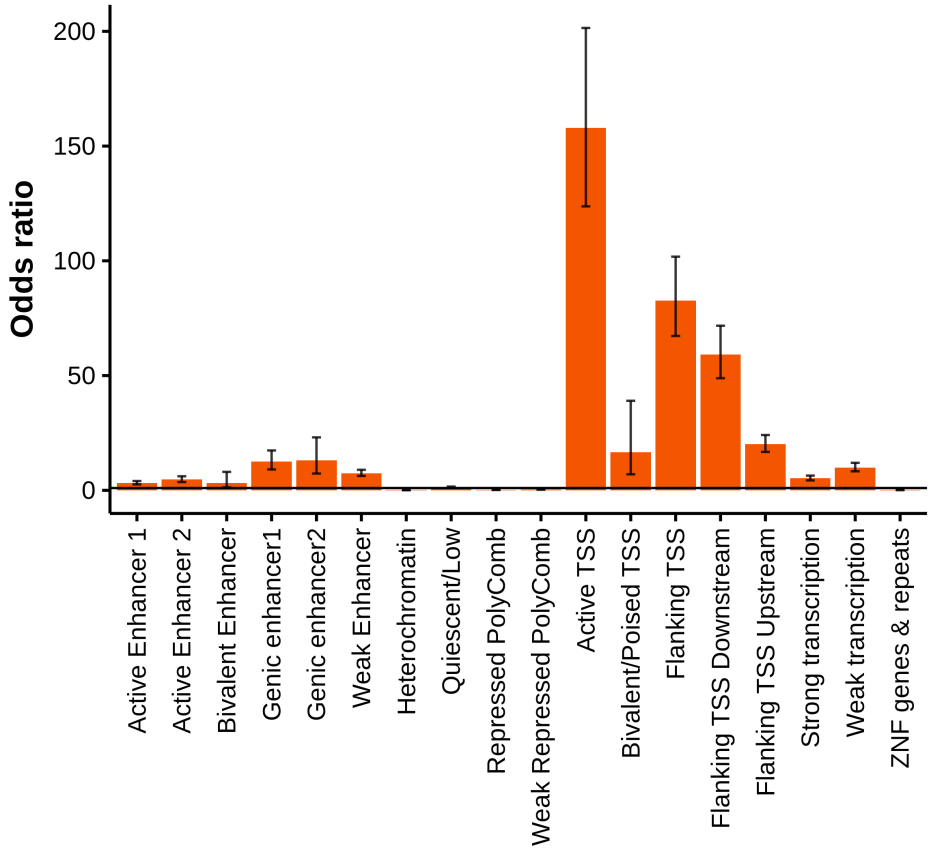
Supplementary Figure 19. Cohesin-depleted HCT116 model shows no strong CTCF motif dependency. Pearson correlation between structural impact score (1Mb) and CTCF binding signal (ENCFF115GQW, ENCFF147GRN, ENCFF761RHS) across holdout chromosomes chr8, 9, and 10 are computed (binned to 1000bp bins by taking the max).



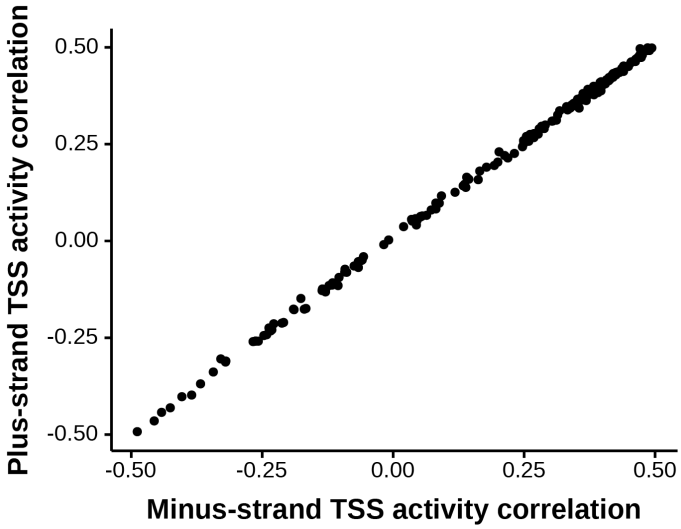
Supplementary Figure 20. Visualization of virtual screen sequence activity on chromatin compartment alteration (continued). All 27981 12800bp source sequences (covering chr8, 9, and 10) are clustered by hierarchical clustering. Target locations are ordered by the main mode of compartment change detected at the target site (from top: A>B to bottom: B>A), which is quantified by the loading of the first principal component of the whole sequence structural impact score (32Mb) matrix. TSS activity measured by FANTOM CAGE signal (max count across samples) summed over the genomic interval of each sequence is shown at the bottom for comparison.



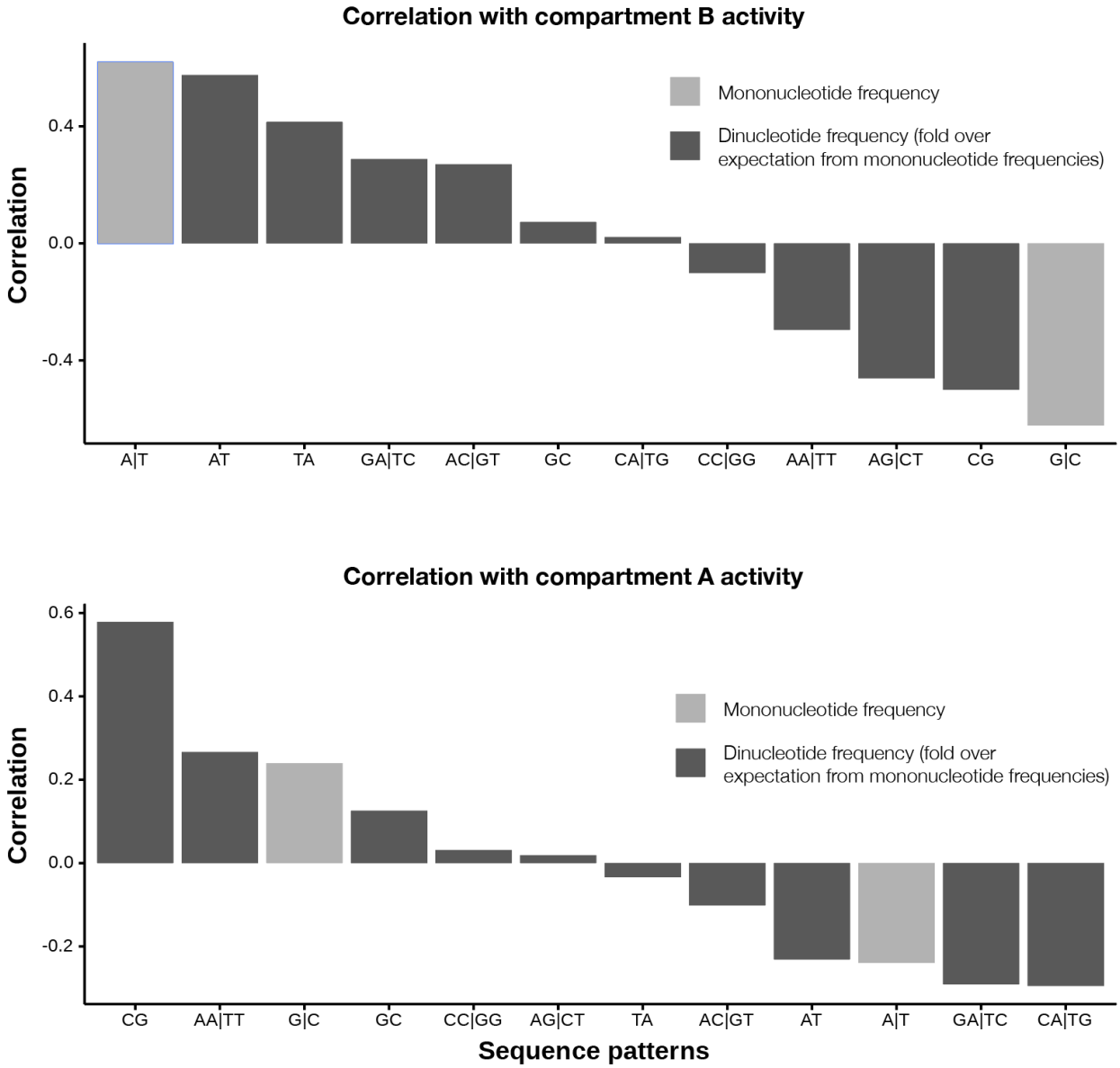
Supplementary Figure 21. Compartment A activity is selectively enriched in TSS. For all 12800bp sequences in the large-scale compartment activity screen, enrichment odds ratios are computed for sequences overlapping with each gene annotation and sequences with top 2% compartment A activity as quantified by the first principal component of the sources-by-targets structural impact score (32Mb) matrix (Methods). Total number of sequences n = 27281. The center of the error bars represents the odds ratios and the error bars show 95% confidence interval of the odds ratios.



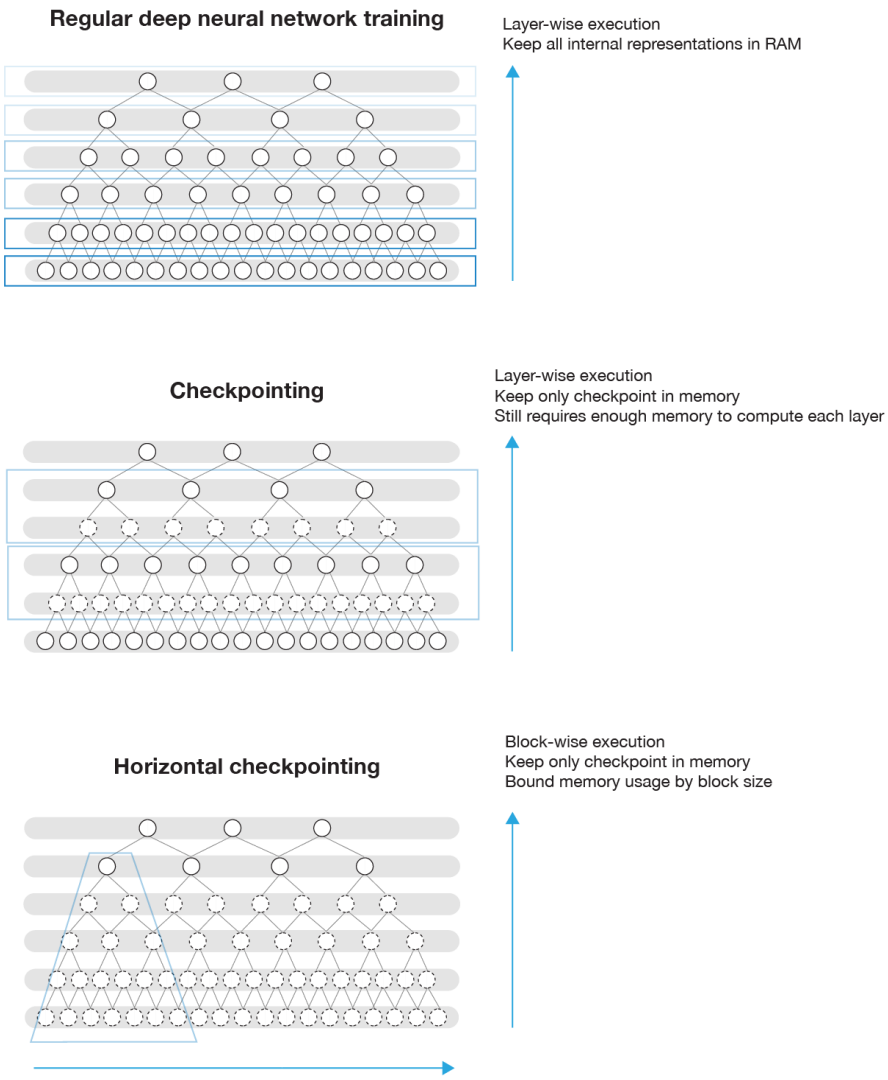
Supplementary Figure 22. Compartment A activity is selectively enriched in active TSS chromatin states. For all 12800bp sequences in the large-scale compartment activity screen, enrichment odds ratios are computed for sequences overlapping with each of the 18 states from an HCT116 chromatin state annotation[39] and sequences with top 2% compartment A activity as quantified by the first principal component of the sources-by-targets structural impact score (32Mb) matrix (Methods). Total number of sequences n = 27281. The center of the error bars represents the odds ratios and the error bars show 95% confidence interval of the odds ratios.



Supplementary Figure 23. No TSS strand specificity detected for chromatin compartmental alteration activity of sequences across target locations. Each dot represents a target location. For each target location, the correlation between the compartmental alteration activity and the plus- (y-axis) or minus- (x-axis) strand-specific FANTOM CAGE TSS activity (log max signal across samples, with pseudocount 1) across source sequences are shown.



Supplementary Figure 24. Mono- and di- nucleotide content correlates with compartment A and B activities. The correlations between mono- and di-nucleotide pattern frequencies (Dinucleotide frequency is divided by the expectation computed from mononucleotide frequencies) with compartment A>B (top panel) and B>A (bottom panel) activities of 27981 x 12800bp source sequences covering chr8, 9, and 10 are shown. The mono- or di-nucleotide sequence patterns are shown in the x-axis (reverse complement pattern pairs are grouped together).



Supplementary Figure 25. Illustration of horizontal checkpointing mechanism for memory-efficient training of hierarchical deep learning models. Regular deep neural network training (top panel) executes the model in a layer-wise order and stores all internal representations, resulting in high memory cost. Memory checkpointing technique (center panel) saves memory by only storing internal representations at the checkpoints, and backpropagation is allowed by recomputing the internal representations in between checkpoints during the backward pass. This is still infeasible for very large models where computing even a single layer requires too much memory. For hierarchically structured models, horizontal checkpointing (bottom), which is based on the memory checkpointing technique, can dramatically decrease the memory consumption by executing the model in horizontal blocks of custom size, allowing training ultra-large sequence models.