preciseTAD: A machine learning framework for precise 3D domain boundary prediction at base-level resolution

Supplementary Figure Legends

Table of Contents

[Supplementary Figure 1. The transformation of genomic distances normalizes their distributions. 1](#_Toc51956586)

[Supplementary Figure 2. Psuedocode. 2](#_Toc51956587)

[Supplementary Figure 3. Determining optimal data level characteristics for building TAD boundary region prediction models on K562. 2](#_Toc51956588)

[Supplementary Figure 4. SMC3, RAD21, CTCF, and ZNF143 transcription factors accurately predict TAD and loop boundaries in K562. 2](#_Toc51956589)

[Supplementary Figure 5. preciseTAD boundaries are enriched for known molecular drivers of 3D chromatin. 3](#_Toc51956590)

[Supplementary Figure 6. *preciseTAD* boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Arrowhead boundaries. 3](#_Toc51956591)

[Supplementary Figure 7. *preciseTAD* boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Peakachu boundaries. 3](#_Toc51956592)

[Supplementary Figure 8. *preciseTAD* boundaries are spatially closer to known molecular drivers of 3D chromatin. 3](#_Toc51956593)

[Supplementary Figure 9. The agreement between *preciseTAD*-predicted boundaries using Arrowhead- and Peakachu-trained models. 4](#_Toc51956594)

[Supplementary Figure 10. Pre-trained *preciseTAD* models accurately predict boundaries on cell lines using annotation data only. 4](#_Toc51956595)

[Supplementary Figure 11. Normalized Enrichment levels suggest t=1.0 and as the most optimal parameters for biologically relevant *preciseTAD*-predicted boundaries. 4](#_Toc51956596)

# Supplementary Figure 1. The transformation of genomic distances normalizes their distributions.

Distances are measured as the number of bases from the center of a genomic bin to the nearest genomic annotation center. Density curves of distances before (red) and after (blue) performing a transformation across 10 kb, 25 kb, 50 kb, and 100 kb data resolutions for both the (A) GM12878 and (B) K562 cell lines. Each density curve represents an individual genomic annotation (77 total).

# Supplementary Figure 2. Psuedocode.

Pseudocode of the *preciseTAD* algorithm.

# Supplementary Figure 3. Determining optimal data level characteristics for building TAD boundary region prediction models on K562.

(A) Averaged balanced accuracies are compared across resolution, within each predictor-type: Signal, OC, OP, and Distance, and across resampling techniques: no resampling (None; red), random over-sampling (ROS; green), random under-sampling (RUS; blue), and synthetic minority over-sampling (SMOTE; purple) when using Arrowhead ground truth boundaries for K562. (B) Averaged balanced accuracies are compared for GM12878 and (C) K562 within each predictor-type: Signal, OC, OP, and Distance, and across resampling technique: no resampling (None; red), random over-sampling (ROS; green), random under-sampling (RUS; blue), and synthetic minority over-sampling (SMOTE; purple). Error bars indicate standard deviation away from the mean performance across each holdout chromosome used for testing.

# Supplementary Figure 4. SMC3, RAD21, CTCF, and ZNF143 transcription factors accurately predict TAD and loop boundaries in K562.

(A) Barplots comparing performances of TAD (Arrowhead) and loop (Peakachu) boundary prediction models using histone modifications (HM), chromatin states (BroadHMM), transcription factor binding sites (TFBS), in addition to a model containing all three classes (ALL). (B) Recursive feature elimination (RFE) analysis used to select the optimal number of predictors. Error bars represent 1 standard deviation from the mean cross-validated accuracy across each holdout chromosome. (C) Clustered heatmap of the predictive importance for the union of the top 8 most predictive chromosome-specific TFBSs. The columns represent the holdout chromosome excluded from the training data. (D) Receiver operating characteristic (ROC) curves and the corresponding average area under the curves (AUCs) when training and testing on K562 data (blue, Arrowhead ground truth; red, Peakachu ground truth) versus training on GM12878 and testing on K5628 data (black, dashed). The curves represent the average sensitivities and specificities across each holdout chromosome. The shaded areas around each curve represent 1 standard deviation from the average.

# Supplementary Figure 5. preciseTAD boundaries are enriched for known molecular drivers of 3D chromatin.

Signal profile plots comparing the binding strength of top TFBSs around flanked (A) Arrowhead called TAD boundaries (blue) and *preciseTAD*-predicted TAD boundaries (green) on K562, (B) Peakachu chromatin loop boundaries (red) and preciseTAD predicted loop boundaries (green) on K562, (C) Arrowhead called TAD boundaries (blue), Peakachu chromatin loop boundaries (red), and SpectralTAD called TAD boundaries (green) on GM12878 and (D) on K562.

# Supplementary Figure 6. *preciseTAD* boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Arrowhead boundaries.

Enrichment heatmaps comparing the signal distribution ofCTCF, RAD21, SMC3, and ZNF143 around Arrowhead-called TAD boundaries vs. *preciseTAD*-predicted TAD boundaries for (A) GM12878 and (B) K562 cell lines.

# Supplementary Figure 7. *preciseTAD* boundaries are more enriched for known molecular drivers of 3D chromatin, as compared with Peakachu boundaries.

Enrichment heatmaps comparing the signal distribution ofCTCF, RAD21, SMC3, and ZNF143 around Peakachu chromatin loop boundaries vs. *preciseTAD*-predicted TAD boundaries for (A) GM12878 and (B) K562 cell lines.

# Supplementary Figure 8. *preciseTAD* boundaries are spatially closer to known molecular drivers of 3D chromatin.

Boxplots comparing the log2 genomic distance distributions from predicted and called boundaries to the nearest (A) GM12878-specific and (B) K562-specific CTCF, RAD21, SMC3, and ZNF143 transcription factor binding sites. p-values are derived from the Wilcoxon Rank Sum test.

# Supplementary Figure 9. The agreement between *preciseTAD*-predicted boundaries using Arrowhead- and Peakachu-trained models.

Venn diagrams of boundary overlaps using (A) GM12878 and (B) K562 data. Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.

# Supplementary Figure 10. Pre-trained *preciseTAD* models accurately predict boundaries on cell lines using annotation data only.

Venn diagrams and signal profile plots comparing flanked predicted boundaries using Arrowhead (A, B) and Peakachu (C, D) trained models. (A, C) train on GM12878 and predict on GM12878 (red, GM on GM) vs. train on K562 and predict on GM12878 (blue, K on GM), (B, D) train on K562 predict on K562 (red, K on K) vs. train on GM12878 predict on K562 (blue, GM on K). Boundaries involving Arrowhead/Peakachu were flanked by 5 kb/10 kb, respectively.

# Supplementary Figure 11. Normalized Enrichment levels suggest t=1.0 and as the most optimal parameters for biologically relevant *preciseTAD*-predicted boundaries.

Linecharts illustrating the normalized enrichment (NE) between CTCF, RAD21, SMC3, ZNF143 and resolution-flanked *preciseTAD*-predicted boundaries for different combinations of thresholds (t) and epsilon parameter values (eps). NE was calculated as the total number of ChIP-seq peaks that overlapped within a given flanked boundary, divided by the number of boundaries that were predicted, and averaged over the number of annotations included in the model. Data from GM12878 (A) and K562 (B) cell lines, chromosome 22, at 5kb resolution was used. Error bars indicate standard deviation from the mean.