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

Supplementary Table Legends

Table of Contents

# Supplementary Table 1.

Data sources for Hi-C matrices used to call topologically associating domains with Arrowhead, as well as loop boundaries obtained by Peakachu.

# Supplementary Table 2.

A complete list of genomic annotations used to build the predictor space for all downstream models. The GRCh37/hg19 human genome assembly was used. “Genomic Class” - broad category of genomic features, “Element” - names of genomic features, “Cell line-Specific Source” - download URL specific to the cell line (not all annotations were provided by the same institutions).

# Supplementary Table 3.

Domain boundary data and class imbalance summaries across resolutions for Arrowhead and Peakachu on K562.

# Supplementary Table 4.

hg19/GRCh37 genomic coordinates of *preciseTAD*-predicted boundary regions (PTBR) and points (PTBP) for GM12878 and K562 cell lines, using models trained on Arrowhead TAD and Peakachu chromatin loop boundaries as ground truth. For PTBRs, the start and end coordinates define the clusters of spatially proximal bases with the probability of being a boundary equal to 1. For PTBPs, the start and end (start+1) coordinates define the most likely boundary point within each PTBR.

# Supplementary Table 5.

Additional performance metrics when implementing a random forest using Arrowhead/Peakachu ground truth TAD/loop boundaries. Performances are averaged across each holdout chromosome that was reserved for testing.