# How to process data for other cell lines

## Step 1: Data Curation

Create a folder for all cell-specific Hi-C, ChIPseq, histone modification, and methylation data. You may our curated data for non-cell-line specific genomic data, such as distance from centromere (relative chromosomal location), gene density (defined by number of TSS), and number of transcription factor binding sites of 161 transcription factors.

Download, obtain, or generate **narrowPeaks bed files** for 8 histone modification proteins and 12 transcription factors as listed in List-of-Bed\_narrowPeaks files. Create a .txt file with at minimum two columns: file\_name and target name.

Download, obtain, or generate **Hi-C data** for your cell line. Processed Hi-C data with identified or calculated TAD or insulated regions can also be used. You may use the 3D: Genome Browser to attain TADs for your cell line.

We have included some data in GitHub for your usage, however due to file size limitations, some files were excluded. Please contact [Jacqueline.chyr@uth.tmc.edu](mailto:Jacqueline.chyr@uth.tmc.edu) for more data, including raw and processed files.

## Step 2: Data reading and preprocessing

Run **1\_gen\_pre\_info.R** to read genomic and epigenomic files and to preprocess the files. Be sure to update the directories.

**TAD boundaries** are defined as regions between two adjacent TADs with insulation scores greater than 0.15 (using insulation square analysis at a 40 kb resolution). If TAD regions were used, TAD boundaries are defined as the region half-way between two TADs. Due to variations in replicates, the width of boundaries were extended to 200 kb.

TF ChIPseq and Histone Modification ChIPseq data are mapped to hg19 with BWA-MEM and **narrow peaks files** were called with MACS2. They should each contain ten columns: chrom, start, end, name, score, strand, sig, pv, qv, and peak. Average signal values for narrow peaks were used. Values per bin were normalized to 0-1.

**DNA methylation** from Illumina Methylation 450K BeadChip array was used in this study. Annotation information for the 450K BeadChip array is provided. Average beta values were used.

## Step 3: Sample preparation

Run **2\_sample\_preparation.R** to **generate samples** with the corresponding features. The entire genome is binned into 10 kb regions. Sample region in the telomere and centromeres were excluded. Average signal values for each bin are normalized from 0 to 1. Ten upstream and ten downstream neighbors’ genomic and epigenomic information were included as feature. Save the large data table and load it into the next step:

## Step 4: PredTAD model

Run **3\_ PredTAD\_model.R**. PredTAD requires H2O.ai in R. Documentation for this can be found at <https://www.h2o.ai/>. Gradient Boosting Machine (GBM) is used as the classifier for **PredTAD**. The samples of this model are 10 kb genomic regions. The input features are genomic and epigenomic features, such as histone modification, transcription factors, number of TSS, etc. The output is whether the region is a TAD boundary or non-TAD boundary.

# How to test new samples using our trained model

You can load GBM\_model\_R\_1597829376110\_1 for **trained model**, then test with new data. New data for other cell lines and samples need to be processed in the same manner as the trained data (see above).