**tss enrichment analysis data processing**

**1. Extract coordinate values**

* clean\_summary.ipynb
* takes the raw data file summary.xlsx and extracts the chr, chromosomeStart, and chrEnd from peak data
* saved file name-> H3K9ac.csv

**A screenshot of a computer

Description automatically generated**

**2. Save the coordinates only to .bed**

* save chr, chromStart, chromEnd to H3K9ac\_coords.bed

**3. Assign gene to each peak**

* assign\_gene.ipynb
* takes H3K9\_coords.bed and tss\_regions\_mm10.bed files and assigns gene to each peak based on closest distance to tss bed coordinates
* result saved to H3K9\_nearest.csv

**4. Combine csv results to summary\_edited.xlsx**

**5. Filter distant genes >10kb**

* distance\_filt.ipynb
* filter out all genes with assigned distance greater than 10kb from the tss
* save to H3K9ac\_distfilt.csv

**6. Combine files to make summary\_distfilt.xlsx**

**7. Drop rows to assign only one entry per gene (one row per gene)**

* compress\_coords.ipynb
* input: summary\_distfilt.xlsx
* output: onerowperGene.xlsx
* only keeps the gene row with the closest distance value

**8. Use boxplot to visualize data**

* input: boxplot\_H3K9.ipynb and boxplot\_H3K27.ipynb
* input2: sig\_hdac3.txt ... etc
* output: onerowperGene.xlsx

---------------------------------------------------------------------------------------------------------------

October 28th, 2024

goal: to see what proportion of hdac3/nfil3/nr1d1 genes have acetylation signal. Previously, genes without ChIP signal were disregarded.

Differences: No filtering based on distance

1. Use compress\_coords.ipynb on summary\_edited.xlsx

* output: onerowperGene\_no\_distfilt.ipynb

2. boxplot\_H3K9\_28Oct24 to make pie charts of signal/no signal