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

	peak	chr	chromStart	chromEnd
0	chr1_4536689_4539694	chr1	4536689	4539694
1	chr1_4571028_4572577	chr1	4571028	4572577
2	chr1_4774286_4777435	chr1	4774286	4777435
3	chr1_4783493_4786944	chr1	4783493	4786944
4	chr1_4800555_4802981	chr1	4800555	4802981
...	...	...	...	...
17808	chrY_90809994_90814071	chrY	90809994	90814071
17809	chrY_90819693_90819950	chrY	90819693	90819950
17811	chrX_GL456233_random_38117_39776	chrX	38117	39776
17812	chrX_GL456233_random_110914_112210	chrX	110914	112210
17813	chrX_GL456233_random_159163_161681	chrX	159163	161681

17747 rows × 16 columns

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