Chip-seq Data Processing Steps

Step 1: Extract Transcription Start Sites (TSS)

Script: `step1_extract_tss_test.py`

This Python script loads a GTF file and extracts the Transcription Start Sites (TSS) for genes or transcripts. It creates a BED file with TSS regions extended by 2000 bp upstream and downstream.

Output: A BED file named `tss_regions.bed` which will be used for read count calculations.

Step 2: Sort and Index BAM Files

Script: `step2_run_bedtools.sh`

A bash script that sorts and indexes BAM files according to the reference genome. It uses `samtools` for sorting and indexing the BAM files.

Output: Sorted BAM files named `sorted_<original_name>.bam` and their index files.

Step 3: Calculate Read Coverage

Script: `generate_gene_read_counts.sh`

Another bash script that uses `bedtools coverage` to calculate the read coverage over the TSS regions for each BAM file.

Output: Coverage files named `<original_name>_tss_counts.txt` for each BAM file.

Step 4: Sum Read Counts for Each Gene Within Individual Samples

Script: `sum.py` (Repeated for each file)

This Python script reads the coverage file for a sample, assumes it is space-separated, and contains gene names and read counts. It sums up the read counts for genes that appear multiple times.

Output: Summed count files named `summed_gene_read_counts_<sample_name>.txt` for each sample.

Step 5: Merge Summed Read Counts Across All Samples

Script: `merge_counts.py`

The final Python script reads all the summed read count files and merges them into a single table. The merge is done on the gene names to create a table where each row represents a gene, and each column represents the summed read counts for that gene across all samples. Missing values are filled with zeros to account for genes that are not present in all samples.

Output: A single merged file named `merged_gene_read_counts.txt` which is a comprehensive table with the structure described above.

Summary of Outputs

- `tss_regions.bed`: BED file with TSS regions for each gene.
- `sorted_<original_name>.bam`: Sorted BAM files for each sample.
- `<original_name>_tss_counts.txt`: Coverage count files for TSS regions per sample.
- `summed_gene_read_counts_<sample_name>.txt`: Files with summed read counts for each gene per sample.
- `merged_gene_read_counts.txt`: Final merged table with genes as rows and each sample's read counts as columns.