PreSTIGE User's Guide:

The PreSTIGE algorithm is accessible online through Galaxy[1], at prestige.case.edu. In this way the user does not have to install or download new software but can use Galaxy's built-in tools for analysis. To run the PreSTIGE algorithm on your sample of interest you will need 3 files:

-H3K4me1 Peaks File
-H3K4me1 wig tracks
-Gene expression data

Suggested Use:
We suggest the user inputs all H3K4me1 peaks for the sample of interest and then extracts the peaks of interest such a transcription factor binding sites or H3K27ac/DHS sites from the generated prediction file. The PreSTIGE algorithm will take the user's peak file and combine these peaks with the peaks generated in the 12 cell line comparator set. Therefore the output peaks file may have slightly different coordinates than the input file. Predictions may be generated for the candidate sites the PreSTIGE algorithm uses by default in addition to the sites the user supplies.

If you are only interested in predictions for the peaks you supply or the a subset of these peaks you can upload the desired peaks list or use intersectBed to extract your sites of interest.

If you are interested in super, stretch or stitched enhancers, we suggest starting with the individual H3K4me1 peaks and then comparing the results with coordinates of broader enhancer peaks.

PreSTIGE currently makes predictions for genes found in RefSeq gene list. Options to make predictions for genes found in the Gencode gene list will be made available soon.

1) H3K4me1 Peaks Files
    a. This can be generated through any peaks calling program. (We suggest MACS[2]*, see H3K4me1 wig tracks)
    b. The file should be of the format, chr[TAB]start coordinate[TAB] stop coordinate. There should be no headers in the file.
    c. Some peak calling programs allow for negative coordinates to be output, (ie. chr1 -170 1400). The bedtools commands utilized by PreSTIGE will not run if negative coordinates are contained within the peaks file.
        i. This can be corrected with the following command: For a peaks file that is: chr start stop

```
cat peaks_file.bed | awk '{OFS="\t"}{if($2<0){$2=0};print}' > corrected_peaks.bed
```

ii. Or these peaks can be removed:

```
cat peaks_file | awk '($2>0)' | filtered_peaks_.bed
```

2) H3K4me1 wig tracks
    a. EVERY chromosome in the wig file must be correctly sorted in ascending order by coordinate for variableStep wig tracks. FixedStep wig tracks may only contain one header per chromosome.
    b. We suggest using MACS* with the generate wig tracks option as the wig tracks will be sorted formatted compatible with PreSTIGE. Combine all chromosome wig tracks into one file and then upload this file to Galaxy. NOTE: this will be a large file and therefore make take some time to upload.

*There is a Peak Calling tool available on the lefthand toolbar of prestige.case.edu. Uploading bam files (generated by ChIP-seq alignment) to Galaxy and then executing Peak Calling here will generate both the peaks files and wigs file needed for PreSTIGE. NOTE: this runs the peak calling off Case Western's servers so may be slow and should only be used for PreSTIGE analysis.

    c. Alternatively, wig tracks can be sorted using Cistrome's[3] wig formatting script (available on the left toolbar of prestige.case.edu) Operations on Wig Files.

3) Gene Expression Data
    a. Two types of gene expression can be utilized by PreSTIGE, RNA-seq and microarray
    b. Files should be of the format: gene-name[TAB]expression[optional TAB][optional gene locus id]. There should be no headers in the file.
    c. The FPKM (or RPKM) must be determined from RNA-seq data. We suggest using Tophat[4] and Cufflinks[5] with the RefSeq GTF file. The genes.fpkm_tracking file can then be used to obtain the gene name, fpkm and locus
        i. Because some genes are annotated with multiple start and stop coordinates if you use cufflinks with the RefSeq GTF file you can then use the option to merge by gene coordinate. This option merges by gene_name_coordinates so that FPKM of genes with multiple annotations are compared accurately.
    d. Microarray data can also be used, though RNA-seq is ideal. Microarray data is quantile normalized against the FPKM values of the RNA-seq data of the comparator set. Then shannon entropy is used to compare datasets. This enables microarray and RNA-seq data to be compared, but due to differences in the dynamic ranges microarray data is not ideal.

4) Obtaining the Output:
    a. Troubleshooting
        i. The PreSTIGE_log.txt will help in evaluating an error. View the log file and determine the source of the error. Most common errors include negative coordinates in peaks file, unsorted wig tracks, headers in the gene expression or peaks file, or the wrong file being selected in the drop-down menus.
        ii. Check file formats using the above guide
    b. Timing:
        i. Uploading large wig tracks will take time, wait to start PreSTIGE analysis until the file is uploaded.
        ii. PreSTIGE involves generating ChIP-seq output for all peaks identified in the comparator-set and in the cell line of interest and then does the comparative analysis to generate predictions. This process will take 3 or more hours depending on the number of peaks to be processed.
        iii. If you need to use PreSTIGE on a large number of datasets we may be able to assist you in running PreSTIGE locally.
    c. Ouput
        i. The output files include: Predicted Pairs for sample of interest, and the predicted pairs for all samples of the comparator set using the peaks uploaded by the user.
        ii. The output columns include the following
            1. Sample Name
            2. Gene
            3. Chromosome
            4. Enhancer start coordinate
            5. Enhancer stop coordinate
            6. H3K4me1: maximum signal in enhancer window
            7. H3K4me1_spec: Q scores from Shannon entropy (low score is highly specific)
            8. gene_expr: gene expression value
            9. gene_spec: Q scores from Shannon entropy (low score is highly specific)
            10. within_ctcf:
                a. CTCF = enhancer and gene are within the same CTCF domain
                b. NA = enhancer and gene are separated by CTCF site.
            11. within_100kb:
                a. <100kb = enhancer and gene are within 100kb
                b. NA = enhancer and gene are greater than 100kb apart

1. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* **11**, R86 (2010).
2. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137 (2008).
3. Liu, T. et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome biology* **12**, R83 (2011).
4. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
5. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515 (2010).