DATA SCIENCE IN LIFE SCIENCE

Identify epigenetic alterations associated with Alzheimer's disease and classification of gene expressions between healthy and sick patients Project Week 05 - Group 02

Pushpa Koirala, Christina Kirschbaum and Melika Moradi

Full list of author information is available at the end of the article

1 Project

This week, we finally finished the ChIP-seq analysis for our project. Additionally, we worked on the Machine Learning, where we were only able to make slow progress because of problems that arose while preparing the ChIP-seq data.

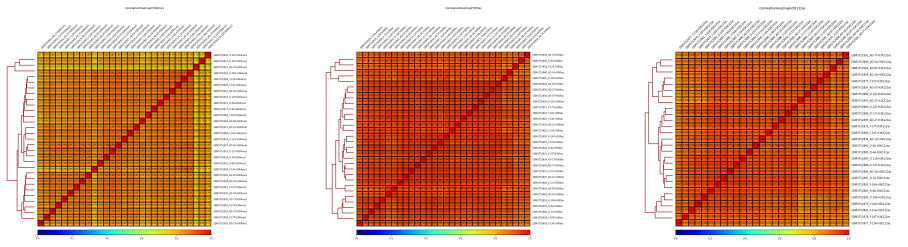
Because we did not finish the Machine Learning this week, we are still behind our timeline. We will now work intensely on completing this part as soon as possible, to move on to preparing our presentation, cleaning up our code and writing the final report.

2 ChIP-seq Data Analysis

2.1 Analysis of H3K4me1, H3K9ac, H3K122ac, H3K27ac

Due to the ongoing issues with Drompaplus during the analysis, we decided to start the analysis on the other datasets with deepTools.

First, the bigWig files were processed with multiBigwigSummary, which computes the average scores for each of the files in every genomic region. The generated coverage files were used for the tools plotCorrelation with Pearson as well as plotPCA.



(a) Correlation Heatmap for H3K4me1

(b) Correlation Heatmap for H3K9ac

(c) Correlation Heatmap for H3K122ac

Figure 1: Correlation Heatmaps for ChIP-seq data

In the correlation heatmaps it is visible, that the samples are not clusered by their respective groups young, old and diseased by the program. Moreover, especially for H3K9ac and H3K122ac the correlations between all samples are very high, so no

conclusion can draw from the results. For the PCA plots, the samples always were in one big cluster, too.

Then, for every sample, the scores per genome regions were calculated with computeMatrix from the bigWig files. Afterwards, the matrixfiles were used for the tool plotHeatmap.

Here, we've chosen a few samples to focus our attention solely on only histone modification and, i.e, H3K27ac. First .npz file was created using multiBamSummary. The coverage calculation is done for consecutive bins of equal size. This was important to assess the genome-wide similarity of BAM files. As previously stated, running the program in bins mode is used to calculate the rad coverages for genomic regions. When plotCorrelation is used to display pairwise correlation values between the read coverages, the result of multiBamsummary is a compressed numpy array (.npz).Color intensity and hierarchical clustering are used to represent the correlation coefficients between pairs of data points. We calculated the Spearman correlation coefficients of read counts for this modification H3K27ac. The dendrogram depicted in the figure shows which samples have the most similar read counts.

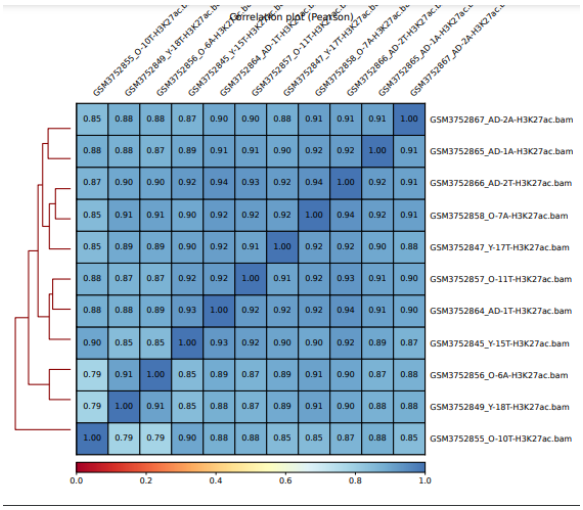


Figure 2: Correlation Analysis computed by multiBamSummary

2.2 Analysis of peaks

The data provided also included bed files for the peaks. For the peaks, the coverage per chromosome was plotted as well as the peak count frequency, the feature distribution and the distribution of transcription factor binding sites with ChIPSeeker.

A differential peak analysis and functional enrichment analysis was started. However, this is not working, or the results are not usable up to now. It will be abandoned now in favor of the Machine Learning and will maybe be finished if there is time left.

As observed in the diagram 2, the result described the per-chromosome coverage. So as to conduct further analysis, we have once more chosen to examine a comprehensive overview of all sample peaks.

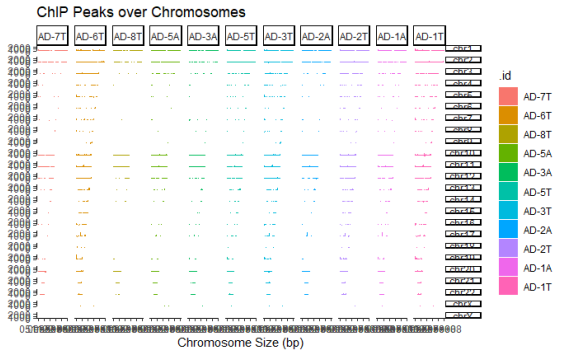


Figure 3: Coverage per Chromosome of the peaks for all AD samples

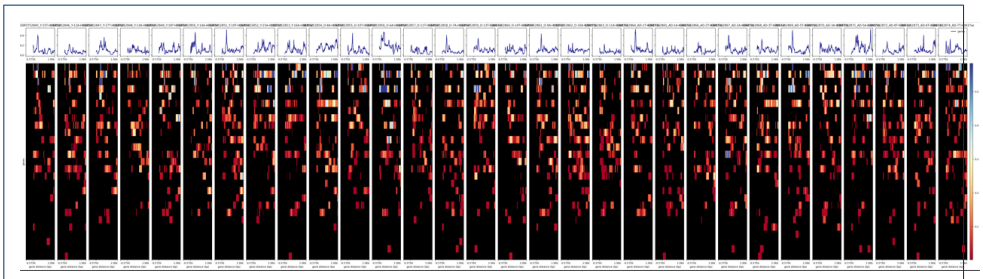


Figure 4: Coverage per Chromosome of the peaks for all samples

As the main focus was to distinguish the coverage between old, young and AD samples, we have decided to have a little closer look into the old, young and AD samples, as seen in figure below 5.

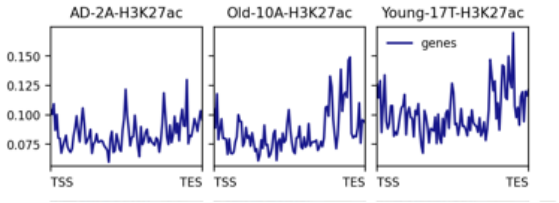


Figure 5: Heatmap scale region of old, young and AD samples

3 Machine Learning

To cluster the data hierarchically, we created the count matrix from RNA-seq and 4 ChIP-seq modifications.

For the hierarchical clustering, we have to build a distance matrix. For RNA-seq, we did the clustering by normalizing the count matrix and creating a distance matrix. We created the hierarchical clustering in R with function hcluster().

For ChIP-seq there were some problems with bam files. Nevertheless, we were able to fix the indexes in bam files with samtools and create the count matrix with bedtools multicov from of each modification. Now we have to create the distance

matrix of each ChIP-seq and cluster them together with RNA.
RNA-seq Clustering with DESeq2 package in R:
First we created DESeq matrix from RNA count matrix and metadata file (GSE159699-summary-count.star.txt, star-col.txt). Then we created DESeq object from the matrix for fitting the model. After we had the DESeq object,we should check how replicates cluster based on the expression levels of the genes. Here we used various clustering analyses.
We used functions in R to calculate the degree of dissimilarity between replicates and conditions based on the expression levels of the genes and then we plotted hierarchical clustering between replicates. We used vst()(varianceStabilizingTransformation), which is part of the DESeq2 package for normalization the data.
After we had the normalized the data we did export and save the normalized counts in a text file. To do this we first used he function assay() to extract normalized counts in readable format from the DEseq object and then we saved the normalized data in a table.
To perform hierarchical clustering we did calculate a distance matrix between all replicates based on their normalized gene counts with function dist(). At the end we did the hierarchical clustering with distance matrix and using of hclust function in R. R script will be on GitHub.

4 Contributions

Member	Last week's work
Christina	ChIP-seq analysis with deeptools, working with peak files
Pushpa	Chip-seq analysis DeepTools - intense gene expression analysis on old, young, samples using multiBamSummary,peaks analysis
Melika	Machine Learning

5 Plans for the upcoming week

- Machine Learning
- STRING analysis
- Prepare presentation for Monday