Janet Pan
PABIO 516 Final Report
6/8/2023

RNA-seq Analysis for Genetic and Genomic Analyses of *Drosophila melanogaster* Models of Chromatin Modification Disorders

Introduction

SSRIDDS and CdLS are syndromic neurodevelopmental Mendelian disorders of chromatin modification with overlapping clinical phenotypes. SSRIDD patients show a spectrum of clinical phenotype including neurodevelopmental delay, intellectual disability, hypotonia, seizures, as well as a cardiac, digit, and craniofacial anomalies. CdLS patients also exhibit a clinical spectrum including intellectual disability, hirsutism, synophrys, and digit, craniofacial, and CNS anomalies. For both of these diseases, the phenotypes are strongly linked to a particular gene, but the severity of these phenotypes can greatly differ among different variations within the same gene. The shared biological mechanisms behind these diseases are very little known. In this analysis, I am interested in the differential gene expression comparing one of the disease, CdLS, and how the most significant genes found in this disease-associated sample when compared to control samples are expressed in SSRIDD-associated samples.

Method

Data Filtering

The data was first filtered genes to exclude low variance and low counts. A scatter plot plotted for mean of log2 counts and variance was created with polynomial models for up to 6 degrees. The data was filtered for mean greater than 2, where the curves begin to drop.

Deseq analysis

I then performed 2 differential gene expression analyses using deseq: CdLS vs control and SSRIDD vs control with pydeseq2. The genes that are significant in CdLS were selected (adjusted p-value <0.05). Within these significant genes, they were ordered by log2 fold change from largest to smallest and assigned a rank value. The highest log2 fold change was ranked first, and the lowest log2 fold change was ranked last. An elbow plot was created to determine the cutoff threshold for log2 fold change. The data was then filtered for log2 fold change that is equal or greater than 2, which is where the curve begins to flatten out. In total, there are 51 genes that passed the filters.

Data visualization

A cluster map was created to visualize the log2 fold gene expression change in CdLS samples relative to mean expression in control samples for the top 25 most significantly expressed genes. I also made a PCA plot to see the expression pattern in CdLS, SSRIDD, and control samples for all the significant genes found in CdLS vs control analysis. Last, I plotted a volcano plot to compare the log2 fold gene expression change and their associated p-value in SSRIDD vs control for all the significant genes found in CdLS samples.

Discussion

When comparing CdLS associated knockdown samples with control samples, SMC3 lines show substantial reduction in gene expression compared to control samples in most of the top 25 significant genes while RNAi line SMC1 samples show increased gene expression compared to control samples in some of the genes found.

To assess the driving force between the CdLS, SSRIDD and control samples, a PCA plot was conducted. According to the PCA plot, it shows about half of the samples differentiate between CdLS and SSRIDD, and the rest of the samples are clustered together with control samples. This implies that even within the genes that exhibit the

greatest differential expressions in the CdLS samples, they are still quite similar to control and SSRIDD samples. This could explain the overlapping clinical phenotypes between CdLS patients and SSRIDD patients.

I then selected all the genes are found significant in log2 fold change in both CdLS and SSRIDD samples. These genes all show positive gene expression fold change in SSRIDD, but all display negative gene expression fold change in CdLS samples. I put these genes into flybase to find equivalent human orthologs, but a lot of them are undocumented. This implies that the leading mechanism/gene that is driving the difference between CdLS and SSRIDD-associated diseases are still yet discovered, or these genes do not have equivalent human orthologs. There is also no significant gene enrichment pathway found from these genes, so it is difficult to make any conclusion about the biological process involved. An explanation could be that the genes that are oppositely expressed in CdLS and SSRIDD samples are in the non-coding regions and therefore the function of these genes are unknown. And since the spectrum of the clinical phenotypes of these diseases varied, it is also challenging to conclude how these genes affect the diseases traits. However, this can also be due to a limited number of genes after filtering and the normalization method which did not take gene length into account. Further research is needed to investigate the biological mechanisms differentiating CdLS and SSRIDD associated diseases.

Result

Fig. 1 Cluster map (top left) shows lo2 fold gene expression change in CdLS samples relative to mean expression in control samples for top 25 most significantly expressed genes from the CdLS vs control analysis. The orange bar indicates CdLS samples and the blue bar indicates control samples.

Fig2. PCA plot (top right) shows quality control to see the gene expression pattern between CdLS, SSRIDD and control samples for all the 51 significant genes found in CdLS when compared to control samples.

Fig3. Volcano plot (bottom) shows genes with log2 fold gene expression change and their associated adjusted p-value (transformed to negative log10). The orange color indicates genes that are significant (adjusted p-value < 0.05) and the blue color indicates genes that are not significant in SSRIDD vs control. The highlighted section shows the genes that are found significant in both CdLS- and SSRIDD- associated samples.