Universiteit Leiden

Master Computer Science

[Integrating transcriptomic sequencing data at single-cell/nucleus level reveals cellular heterogeneity in FSHD]

Name: [Mengyu Zhao]
Student ID: [2938413]
Date: [04/08/2022]

Specialisation: [Bioinformatics]

1st supervisor: [Katy Wolstencroft]
2nd supervisor: [Anita van den Heuvel]
Daily supervisor: [Dongxu]

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Background: Multiple publications have suggested that DUX4 is a key gene in the cure of FSHD. However, the expression of DUX4 is sporadic and heterogeneous. Conventional bulk RNA-seq is challenging to identify essential cells as they are masked by most unaffected cells in the sample. However, with the development of single-cell sequencing technology, some FSHD samples have been applied for this purpose. There is very little literature to explore the heterogeneous expression of DUX4 through integrating data on FSHD samples. In this study, we aimed to select the superior integration method to detect DUX4 expression by combining multiple data sets to explore the physiological significance of FSHD.

Methods: This study provides multiple integrations of five datasets related to FSHD through a macro perspective, an induction dataset focused on the sc level, and a patient dataset focused on the sn level in three parts. The integration is carried out using the harmony approach and creating a complete pipeline. Finally, a simple interactive shiny app is created using the R language.

Results: The analysis showed that the FSHD atlas was first successfully established by integrating all the datasets. Then, to gain more insight into the expression of DUX4 in FSHD patients and healthy controls, the conjecture that the introduction of DUX4i single-cell RNA-seq might capture cells affected by FSHD without detecting the DUX4 target gene was established. The results showed that although there is a link between the two associations, it is unclear whether this partial overlap is due to this conjecture or to myoblast growth, which requires us to design further experiments. Next, pseudotemporal trajectory analysis allowed us to dissect the transcriptome dynamics of FSHD1 and FSHD2 and define the transcriptome characteristics of the nuclei in different states. Finally create a simple interactive shiny app web page.

Contents

Chapter 1

Introduction

1.1 Background

1.1.1 Facioscapulohumeral Muscular Dystrophy

Facioscapulohumeral Muscular Dystrophy (FSHD) is a relatively common form of inherited muscular dystrophy, with most cases of FSHD being inherited in an autosomal dominant man- ner. The first muscles clinically affected by the disease are the face and shoulder muscles, and the symptoms (muscle weakness and wasting) usually progress to the lower body, affecting the humerus, trunk and leg muscles [1]. It also presents asymmetrically from side to side, but the symptoms that patients may experience and the time of onset varies from person to person. FSHD is the result of inappropriate expression of DUX4 on chromosome 4q35 in skeletal muscle [2]. D4Z4 repeat has multiple tandem copies with 3.3 kb. Each unit contains a copy of the intronless DUX4 gene [3]. DUX4 is a toxic transcription factor that activates hundreds of downstream genes, impairs muscle development, activates immune responses, increases susceptibility to oxidative stress and ultimately leads to activation of cell death pathways [4].Although the symptoms are similar in clinical presentation, at the genetic level FSHD can be divided into two categories, FSHD1 and FSHD2.FSHD type 1 (FSHD1)accounts for about 95 percent of the prevalence. FSHD1 is caused by loss of the D4Z4 unit at 4q35 on at least one allele to between 1 and 10 resulting in epigenetic derepression, allowing DUX4 to be transcribed from the most distal D4Z4 unit [5].A relatively small proportion of cases are FSHD type2(FSHD2). FSHD2 is caused by the mutation of epigenetic modifiers. Approximately 80 percent of

FSHD2 cases are resulted from mutation in SMCHD1, a regulator of chromatin methylation status including the D4Z4 repeat array[6].

1.1.2 Applications in Single Cell Sequencing Analysis

The rapid development of next generation sequencing (NGS) and third generation sequencing (TGS) technologies has brought about a dramatic change in the field of biological research. Previously, a large number of cells with sufficient DNA had to be sequenced so that the sequencing results would be a "complete" representation of these cells. Thus, due to cellular heterogeneity, cells with similar genotype (both having the FSHD1/2 genotype) have different DUX4 expression (phenotype) and much of the low abundance of information may be lost in the overall characterization. As we mentioned above, DUX4 and DUX4-induced responses in FSHD show very fragmented and highly heterogeneous patterns that would not be detected by bulk RNA-seq because they would be masked by the majority of normal cells in the sample. To compensate for the limitations of traditional high-throughput sequencing, single-cell sequencing technologies have emerged. Single-cell sequencing is a new technology for high-performance sequencing of the genome, transcriptome and epigenome at the cellular level. It can reveal the genetic structure and state of gene expression of individual cells (The uniqueness of a single cell), reflecting heterogeneity between cells. It therefore has become the focus of bioscientific research, as it plays an important role in areas such as cancer, developmental biology, microbiology and microbiology. SnRNA-seq has three main advantages over scRNA-seq. The first point is the abundance of applicable sample types and relatively simple operation steps. Since the nuclear membrane is more stable than the cell membrane, the nuclear membrane will not be destroyed after tissue freezing, so the frozen tissue can be nucleated for snRNA-seq, which improves the sample type and experimental design richness of single-cell sequencing. The second point is to reduce the artificially introduced transcriptional bias. Because the tissue can be nucleated directly from the frozen state, the cell transcriptional activity is already suppressed and fixed in this state, so no further transcriptional state changes will occur and the authenticity of the results is improved. The third point is that the comprehensiveness of cell types is relatively improved. The cells are mechanically or chemically fragmented in the direct lyophilized state without the

dissociation preference introduced by enzymatic digestion, and the comprehensiveness of cell recovery is relatively higher. By comparing and analyzing the differences between ScRNA-seq and SnRNA-seq of kidney tissues using both dropseq and 10x platforms, it was found that the results of snRNA-seq contained many cell types that were not present in scRNA, with a 20-fold increase in the proportion of glomerular foot cells [6].

So, both single-cell sequencing and single-nuclear sequencing have their advantages. We need to choose the appropriate method according to our sample type, cell subpopulation of interest, and other conditions. In the literature, on muscle research, it can be shown that single-cell RNA sequencing and mass flow cytometry studies outline the main mononuclear cell types present in skeletal muscle and the main cell types consistently include the following broad categories: fibroadipogenic progenitor (FAP) cells, tendon cells, endothelial cells, smooth muscle cells, immune cells (B cells, T cells, macrophages, neutrophils), neural/neuroglial cells, and satellite cells. Notably, while mature myofibers are a small part of single-cell studies, they dominate within the tissue because multinucleated cells are not easily isolated in single-cell dissociation methods. In contrast, single-nucleus RNA sequencing can determine the extent of transcriptional diversity within multinucleated skeletal muscle fibers. Single-nuclear sequencing can analyze nuclei of mouse skeletal muscle throughout the life cycle, which reveals the presence of distinct myonuclear populations that emerge in developing and aging muscle after birth [7].

In this thesis, we cover the dataset of the single-cell technique and the dataset of the single nucleus technique. Because single-cell and single-nucleus techniques allow us to reveal the high cellular heterogeneity of FSHD, preliminary results were detected in approximately 1:250 mononuclear myoblasts with an expression of DUX4 or DUX4 markers [8].

1.2 The purpose

In this report, we focus on the analysis of the expression of DUX4. However, endogenous DUX4 expression is extremely low (only 1/1000 myogenic cells or 1/200 myotubular nuclei in primary cells of patients are positive for DUX4 by immunofluorescence [9]). In addition to this, previous studies have reported that there is a temporal sequence in the expression of dux4, which is not a gene that is consistently expressed for a long period of time. These factors pose many difficulties for the detection of DUX4. Although the

small amount of dux4 expression resulted in our inability to explore the entirety of our transcriptional analysis, single-cell/nuclear datasets on FSHD are growing. Although each dataset may have its limitations, together they form a growing repository of information that continues to improve our view (resolution) of the cellular heterogeneity of FSHD-associated transcriptional changes in muscle. To achieve this goal, integrate the results of our initial scRNA-seq (from primary myoblasts) and some newly generated sc/snRNA-seq datasets to improve our insight into transcriptional changes in FSHD-affected muscle cells. The thesis is composed of four main parts. Firstly, generate a full integrated data set. Secondly, perform more targeted integration— our primary dataset and DUX4i dataset. Thirdly concentrate the analysis on single-nuclei data. At last, there is a small shiny app to visualize data.

Chapter 2

Materials and method

2.1 Data source

For this study, five datasets related to FSHD were mainly used. Single-cell dataset sequence numbers were obtained by reviewing FSHD-related literature and are available on the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). NCBI contains a collection of non-redundant sequences representing genomic data, transcripts, and proteins [10]. The datasets used are shown in table 2.1:

1. Single-cell myocyte dataset from the reference article, which is Sequenced data, was analyzed using the 10x Genomics software, Cell Ranger version 1.2.0 (https://www.10xgenomics.com/), and reads were aligned to the Genome Reference Consortium Human Build 38 [8]. This included six samples of primary myogenic cell cultures from muscle biopsies from two healthy control donors, two FSHD1 patients, and two FSHD2 patients, respectively.

2. There are two datasets from the laboratory of Leiden University Medical Center. One dataset is a single cell sequence after induction of DUX4, and the other is single cell nuclear data.

3. Single-cell nuclear data from human healthy controls and FSHD2 myogenic cells from patient's quadriceps and tibia biopsies was captured by fluidigm C1. Raw reads from single nuclear RNA-seq are mapped to hg38 by STAR (version 2.5.1b) using default values, except for a maximum of 10 mismatches per pair with a mismatch ratio read length of 0.07 and a maximum of 10 theaters. RSEM (version 1.2.31) gene annotation

Reference	Sample material	Sample size	Technique	Nr. of cells/nuclei	Platform
van den Heuvel et al, 2019		2 FSHD1			
GSE122873	Single FSHD1/2 myocyte	2 FSHD2	scRNA-seq	7047	Chromium™ Single Cell 3 v1 RNA sequencing specification
		2 CTRL			
Unpublished DUX4i scRNA-seq	Single DUX4i myoblast	2 DUX4i cell lines	scRNA-seq	4184	10X
		3 FSHD			
Unpublished snRNA-seq	Single FSHD Myonucleus		snRNA-seq	9873	10X
		1 CTRL			
Jiang et al, 2020		2 FSHD2			
GSE143493	Single FSHD2 Myonucleus	2 CTRL	snRNA-seq	32273	BioRad ddSeq Single Cell Isolator
Ator Ashoti et al, 2021 GSE156154	Chronic myeloid leukemia cells		scRNA-seq	2930	CEL-seq2-bases scRNA-seq

Table 2.1: Source of the data sets for this study

from GENCODE v28 was used by default for quantification, and the findings were output as transcripts per million (TPM) [11].

4.The last dataset is a chronic myeloid leukemia transgenic cell line in which the addition of doxycycline 23 induces DUX4 expression [12].

2.2 Method

2.2.1 Data cleaning

Due to the different preferences of each researcher and the different sequencing platforms, the data available on the NCBI website are in various formats, with varying forms of naming and a diverse number of decimal places, so we first had to read all the datasets in a uniform format and differentiate each dataset by labeling them.

2.2.2 Eliminate batch effect

For the project, firstly used two integration methods to reduce the batch effect: the CCA method and the harmony method.

CCA

The first method used to eliminate the batch effect was canonical correlation analysis (CCA) combined with a mutual nearest neighbor (MNN). First, use CCA analysis to downscale two datasets to the same low-dimensional space and then apply L2-normalization to the canonical correlation vectors [13]. Because the spatial distance after CCA down-scaling is not similar but correlated, cells of the same type and state can overcome the technical bias by overlapping. Then, the cells have a measurable "distance" in

the low-dimensional space after CCA down-scaling, and the MNN (mutual nearest neighbor) algorithm uses this to find the cells that are "nearest" to each other between the two datasets. Seurat calls these mutual nearest neighbor cells "anchor cells."

Harmony

And another method, Harmony-The principle of this method is to first embed the transcriptome expression profile into a low-dimensional space by PCA analysis, then iterate until convergence to remove dataset-specific effects. Harmony requires the input of coordinate values (embedding) in the low-dimensional space, generally using the dimensionality reduction results of PCA. After Harmony imports the dimensionality reduction data from PCA, the cells are clustered using a soft K-means clustering algorithm. The commonly used clustering algorithms only consider the distance of cells in the lower dimensional space. However, the soft clustering algorithm considers the correction factors we provide, and each cell can be updated by linearly combining multiple coefficients for the soft clustering assignment. It assigns cells to clusters with a certain probability of maximizing the diversity of the dataset in each cluster. The global center of each cluster is computed along with the center of the dataset. In each cluster, Harmony calculates a correction factor for each dataset based on the center. Harmony corrects each cell according to the correction factor. For example, cell c2 is a bit far from cluster1 and cannot be counted as part of cluster1; however, the cells of c2 and cluster1 are from different datasets, and since we expect the different datasets to be mixed, we make an exception and let it join cluster1. After clustering, we first compute the cell in each cluster of each dataset with the centroids. Then we calculate the centroids of each cluster based on these centroids. Finally, the algorithm allows the cells in the clusters to converge to the center, and the outlier cells that really cannot converge are filtered out. For the adjusted data, the process of clustering - computing cluster centroids - fusing cells - clustering is iterated until the clustering effect stabilizes.

2.2.3 Create the pipeline for the single cell analysis

But when we comapared these two methods, it shows CCA always overcorrect the dataset. So for the following analysis, I merely used the Harmony method. And based the method

of harmony, I create a pipeline for data integration of published single-cell/nucleus RNA-sequencing data, it is applied to the following result sections. This pipeline included QC, a few visualizations and downstream analysis for the integration.

Chapter 3

Result

3.1 sc/snRNA transcriptomic atlas in FSHD

First, all data are integrated to show the transcriptome profile of FSHD at the single cell/nucleus level. A before and after integration comparison graph can be observed using the pipeline mentioned in the methods. Each dataset is labeled with different meta-information, such as which sequencing platform the dataset comes from, whether the cell source is a single cell or single cell nucleus, whether the cells are from primary or post-induction cells, etc. The UMAP plots map the high-dimensional data to the low-dimensional space, allowing analysis based on global and local structure [4]. The data obtained from different sequencing platforms (ddseq and 10X) are distributed in the upper and lower parts of the plot; they are divided into single-cell and single-nucleus data according to their cellular origin, which can be seen in the yellow part on the left side of the plot. On the contrary, the green part on the right is mononuclear; according to the cell type, we can divide them into primary and induced cells. Then, from Figure 3.1, we can see the distribution of each sample on the UMAP plot, where each sample from the same dataset is clustered together. Meanwhile, representatives from different datasets are distributed further apart.

After integration, data from different sequencing platforms, different cell sources, different samples, and different cell types are mixed, which is what we would like to see. It is hoped that after downscaling and visualization, there will be more overlap in the data measured from the same cell development to show their connection, rather than directly mixing the

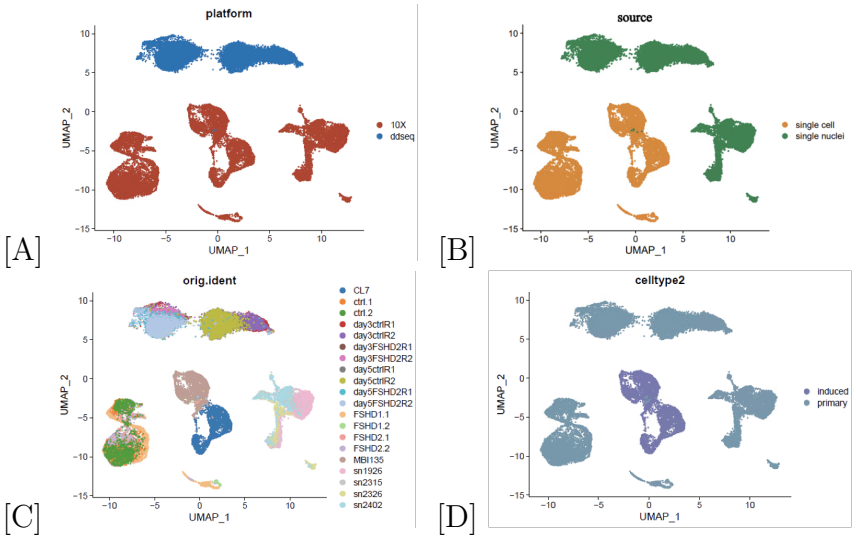

Figure 3.1: The visualizations of all available datasets after directly merging

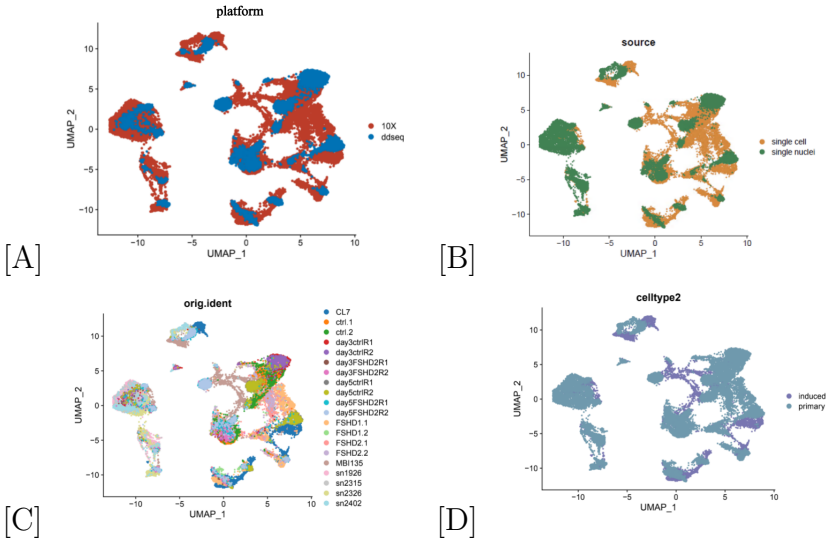

Figure 3.2: The Visualization after integration of all data

data.

3.2 Generate targeted integrated datasets

Based on the assumption that DUX4 may be able to cross cell membranes via homologous domains and thus have an effect on adjacent cells [14], this part of the hypothesis is that introducing DUX4i single-cell RNA-seq may capture cells affected by FSHD without detecting DUX4 target genes. Therefore several more targeted datasets are integrated into this section. A total of three datasets are used here. These three datasets are 1. DUX4i

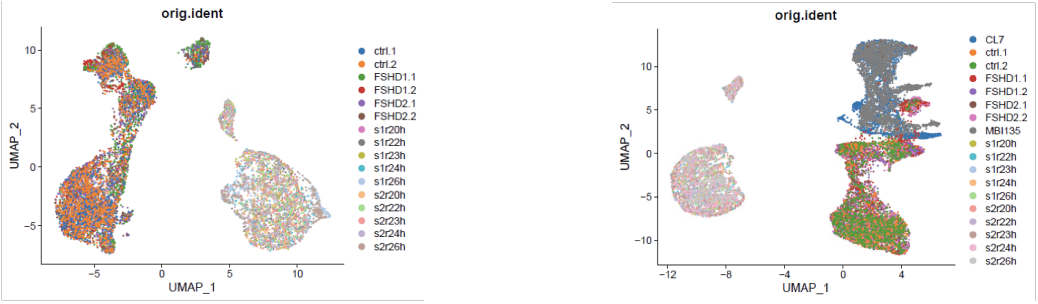

Figure 3.3: Integration of non-myogenic and myocyte datasets

scRNA-seq of non-myogenic cell cultures treated with doxycycline for 0 h, 2 h, 3 h, 4 h, and 6 h, accompanied by duplicates; 2. scRNA-seq of primary myogenic cells containing six samples of 2 CTRL and 4 FSHD, and 3. scRNA-seq of primary myogenic cells containing two samples of CL7 and MBI135 myogenic cell cultures treated with doxycycline for 16 hours for DUX4i scRNA-seq.

First, the non-myogenic cell dataset was used as the induced cell dataset, and the myogenic cell dataset was used as the primary cell dataset. The combined results are shown in Figure 3.2. As shown in the figure, there is no overlap between these two datasets, indicating that they are relatively different. This may be due to the use of variable cell line data. However, it could also be due to the small number of induced cells and the fact that there is no way to reduce this biological variability by integration. So again, another data set based on myogenic was introduced, as shown in Figure 3.2. This figure shows that there is still no overlap between the datasets of different cell lines. Nevertheless, the datasets of different cell types under the same cell can be aggregated together, so the non-myogenic dataset was removed. Only the two primary and induced cells datasets under a uniform cell line were used for subsequent analysis.

So integrate these two datasets separately and classify them. As shown in Figure 3.2, there are 7 clusters in the figure. Figure 3.2 shows the visualization of the different cell types, and based the classification of the clusters, cluster2 and cluster5 contain cells from two other datasets.

To get a clearer picture of what kind of biological information each cluster represents, we identified the top 10 genes of each cluster, and to visualize their intrinsic connections, we did a heat map analysis in Figure 3.5. where MBD3L2, MBD3L2B, KHDC1L, TRIM43, CCNA1, and DUXA are the target genes of DUX4. These genes are more expressed in

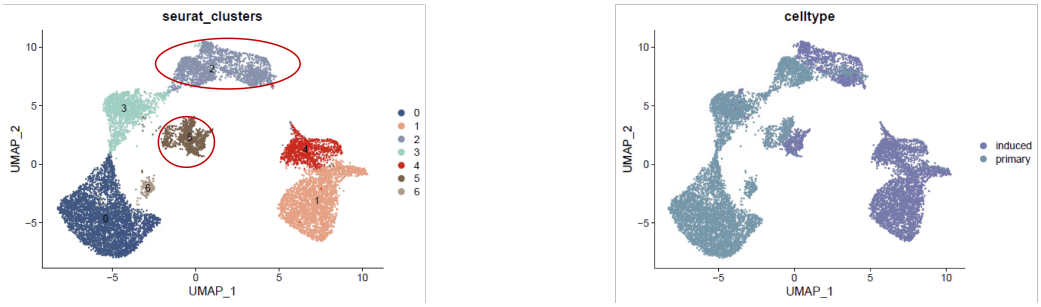

Figure 3.4: Integration of the myoblast and induced cells

clusters 1, 2, 4, and 5. Clusters 1 and 4 are data from induced cells, so we ignore them. We are curious that although cluster 2 and cluster 5 contain cells from both datasets, they do not show similar expressions of DUX4 target genes. So to get a clearer picture of DUX4 target gene expression, the expression of DUX4 in all clusters (Figure 3.6) and the expression of the DUX4 target gene in cluster 5 (Figure 3.7) were done separately. The expression of the DUX4 target gene is mainly expressed in the induced data set and sporadically in the primary cells. But it is worth noting that the expression of the DUX4 target gene in cluster5 is from the induced cells or the primary cells. Then cluster5 was extracted again to observe the DUX4 expression. But the figure then shows that most of the expression of DUX4 is associated with induced cells and has little connection with primary cells. Maybe these Myocytes have been affected by DUX4 but don't have a detectable expression of DUX4 target genes.

This makes us even more curious about what causes primary and induced cells can be clustered together in cluster5. For this purpose, the generated and original datasets with cluster information were obtained by reclassifying the cells according to their type. The genetic markers of cluster5 can then be found and compared with other clusters of the same induced cells. Same as the primary cells. Overlapping several primary cells did not show enrichment of DUX4-pos cells.

And if gene enrichment analysis is done for the essential genes of cluster5, as shown in the figure, the results show that cluster5 is more associated with muscle development.

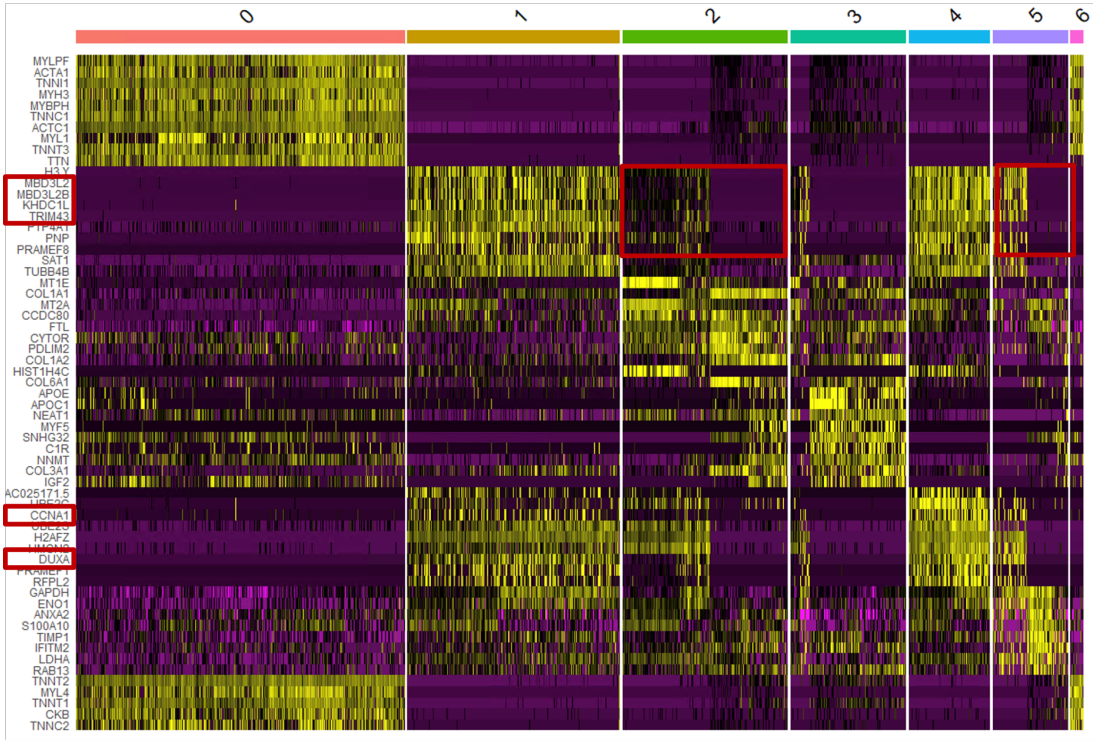

Figure 3.5: Heatmap of top10 genes in each cluster

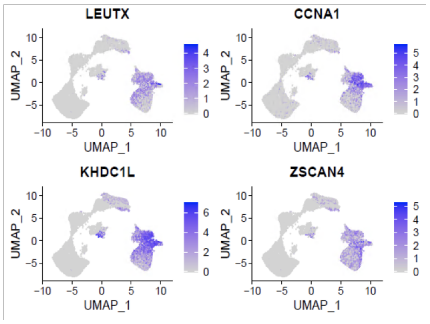

Figure 3.6: Visualization of DUX4 target genes in all clusters

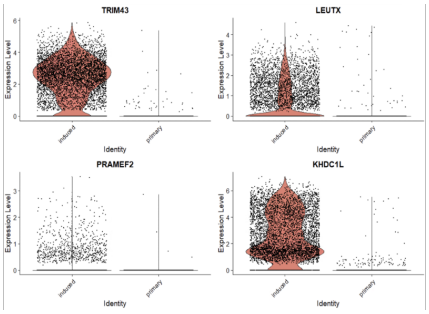

Figure 3.7: Visualization of iDUX4 target genes in cluster 5

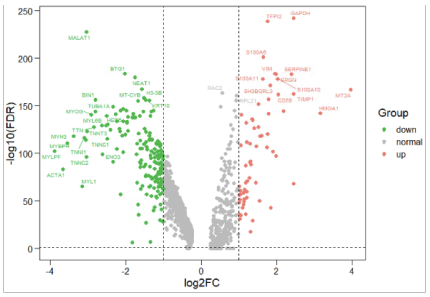

Figure 3.8: Up and down genes in induced dataset

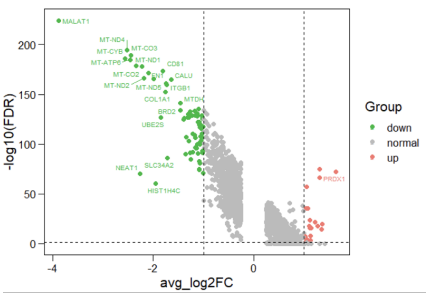

Figure 3.9: Up and down genes in primary dataset

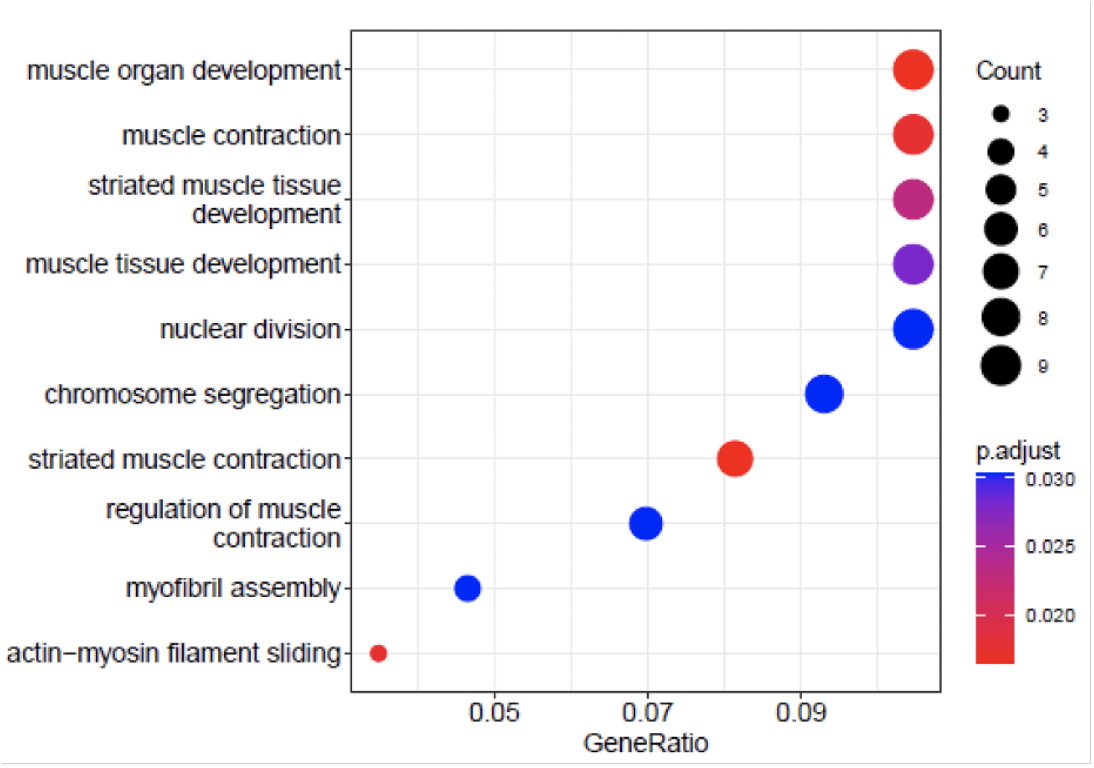

Figure 3.10: GO terms in cluster 5

3.3 Purified snRNA-seq atlas and trajectory analysis

To continue reducing the possibility of introduced variants, we moved our attention to monocyte sequencing and performed the transcriptional analysis at the level of single-nuclei. Two datasets can be used here, they are:

1. snRNA-seq on myotube culture derived from FSHD2 patients. this dataset has a control sample and FSHD2 sample on day 3, a control sample and FSHD2 sample on day 5, and one replicate group for each sample.

2. snRNA-seq on myotube culture derived from FSHD1 patients. There is one control sample and three FSHD samples in this dataset.

Therefore, the two datasets were integrated harmoniously to obtain the figure 3.11. By looking at the left part of the graph reffig:snRNA, it can be seen that it has a good combination of datasets from different sequencing platforms. However, if one focuses on the right part of the graph, one can see that there are too many clusters. For example, the top cluster contains most cells with low fshd, but there are also other cell types. This suggests that there are multiple cell states in the nucleus. This is also more in line with our expectation because some studies have shown that including intron region reads can

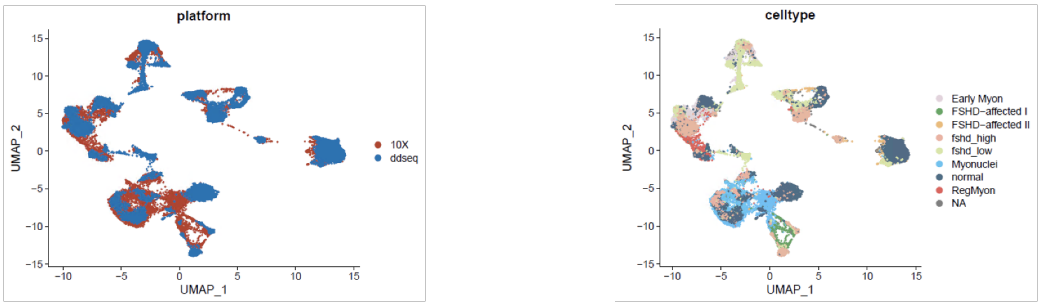

Figure 3.11: Integrated visualization of all snRNA data

improve the sensitivity of snRNA-seq and increase the resolution of identifying cell types [15]. So it offers a lot of diversity in clustering. We can dig deeper into each cell state within the nucleus. However, for my paper, we are more interested in the cells that are labeled as FSHD.

Fshd high, fshd low, fshd affected1, and fshd affected2 are the meta-information of the annotations. fshd high and fshd low are from the reference article [11], so I reproduced a part of the reference article to get the fshd high and fshd low annotations. However, in our lab, the fshd affected1 nucleus and fshd affected2 nucleus have also been labeled regarding fshd expression, and we are interested in the connections and differences between these four. The first hypothesis is whether the already labeled fshd affected1 and fshd affected2 can be matched with fshd high and fshd low. Secondly, another hypothesis is that the integration of nuclei affected by FSHD may provide further insight into FSHD progression to examine the cellular heterogeneity within the nuclei of these cells with different fshd types. Therefore cells containing these four pieces of information were extracted from the complete snRNA data based on the previous step and integrated.

Figure 3.3 shows the visualization with integrated fshd high, fshd low, fshd affected1, and fshd affected2 data. The left panel looks like they blend pretty well. Since we need to check if this is the nuclei cluster we need, the expression of DUX4-related genes is shown in this small dataset, as shown on the right. The expression of DUX4-related genes is relatively distributed throughout the dataset. This indicates that the analysis can be continued to the next step.

Regarding hypothesis 1, we can check the percentage of different FSHD high, FSHD low, FSHD affected1, and FSHD affected2 in each cluster. From the results, if FSHD affected1 and FSHD low are more dominant in a cluster, it is possible to group both roughly. This is

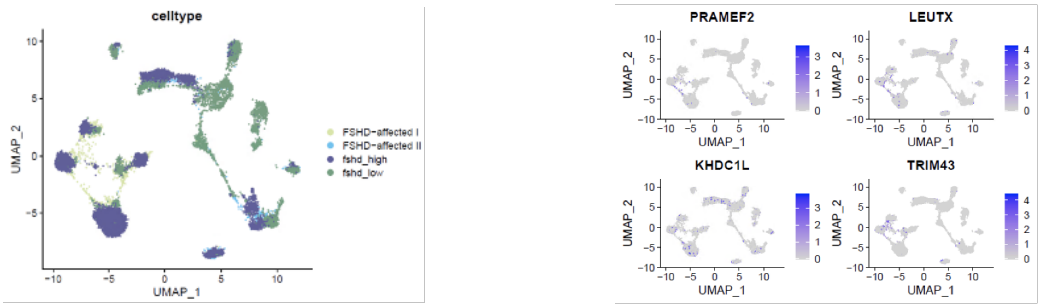

Figure 3.12: Extraction of 4 labeled fshd cells for integration visualization

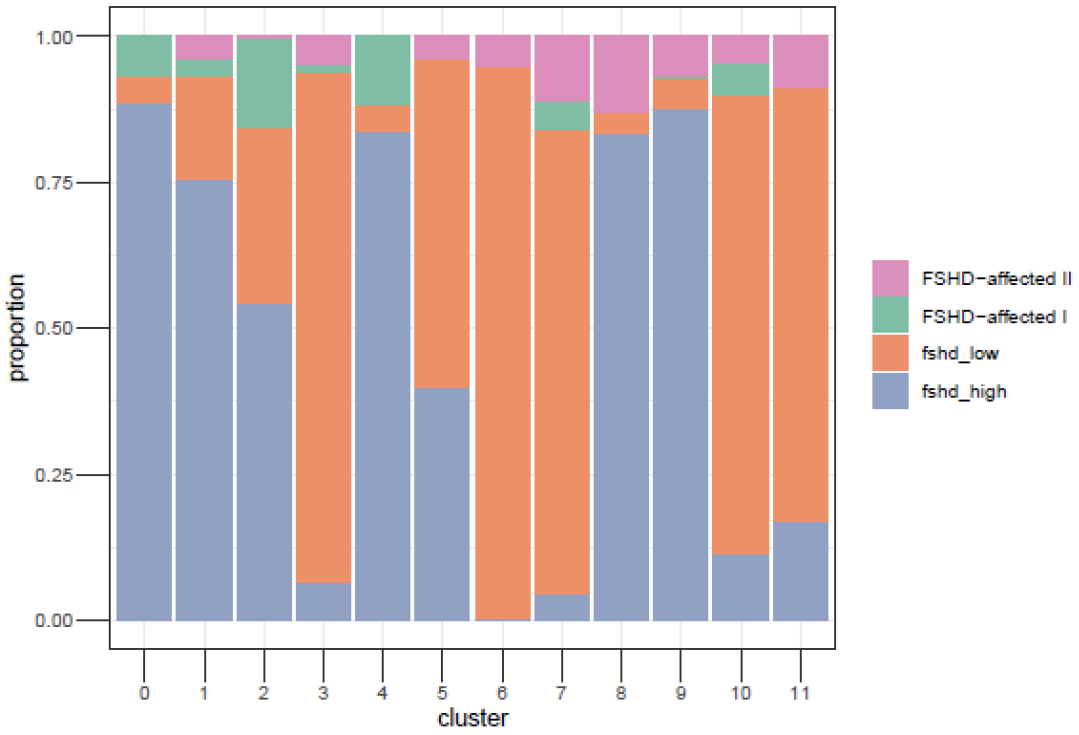

Figure 3.13: Proportion of FSHD nucleus in each cluster

because the division of clusters is related to the similarity of cells. As shown in the figure **??**, it can be divided into these 12 clusters. Based on these clusters, to check whether fshd high, fshd low, fshd affected1, and fshd affected2 can be matched. However, it shows that the distribution of FSHD clusters is reversed in cluster 7 and cluster 8. Therefore, it is impossible to correspond them one to the other.

Then because of hypothesis 2, we wanted to gain more insight into FSHD progression by trajectory analysis within these cell nuclei. The trajectory analysis is mainly based on the expression patterns of essential genes. By learning the sequence of gene expression

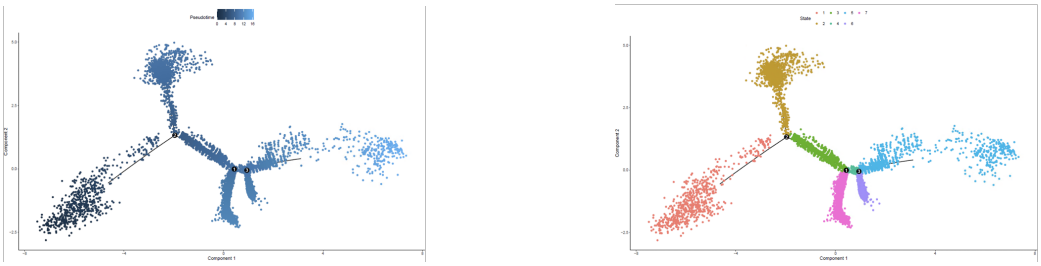

Figure 3.14: Visualization of trajectory analysis and different cell state.

changes that each cell must undergo, the dynamics of the temporal development process are simulated by classifying individual cells according to the proposed time values. It not only stimulates the developmental trajectory of the cells but also clusters them (t-SNE). By obtaining differential genes in different states through clustering, essential genes and their functions influencing branch formation can be analyzed.

The figure3.14 shows that each point represents a cell, and cells with similar states are clustered together. Each branching point represents a decision point for a possible biological process. So the left side of the graph shows that the data starts at the bottom left corner and develops over time. And the right panel shows that the trajectory analysis can be divided into seven cell states.

In addition to observing the pseudo-temporal trajectory and seeing the distribution information of fshd high, fshd low, fshd affected1, and fshd affected2 in the pseudo-temporal wound, we can also check the distribution of the total number of DUX4 and related genes expressed on this trajectory, as shown in the figure 3.15. The left panel shows that our data are mainly distributed at the edges, while their data are concentrated in the center. And the right graph shows that DUX4 expression is more clearly indicated at the edges, while the central part shows less expression. However, this does not match the true biological significance in connection with the start point of the previous graph in the lower-left corner. This leads us to consider the question of determining the starting point of the trajectory analysis. Since temporal trajectory analysis does not represent the real temporal developmental information, we can redefine the starting point according to what we need. Suppose we want to define a pseudo-temporal sequence based on myogenic development. In that case, we can examine several key genes for myogenic development, as shown in fig 3.16. MYF5 is an early myogenic marker. From the figure, we can see that MYF5 is clustered in states 2 and 7. So we decided to define the start of the pseudo-timing

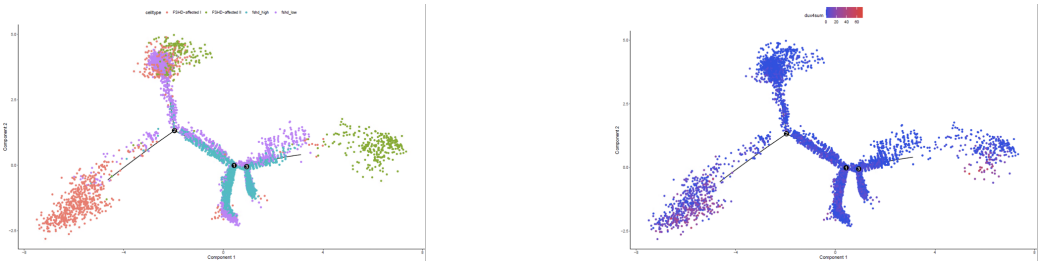

Figure 3.15: Visualization based on FSHD types and the expression of DUX4-related genes

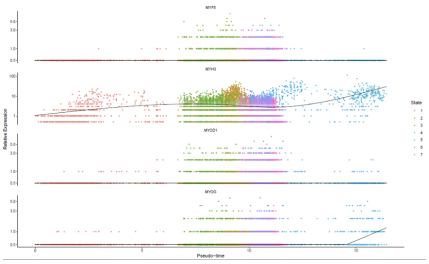

Figure 3.16: Expression of key genes in myogenic development

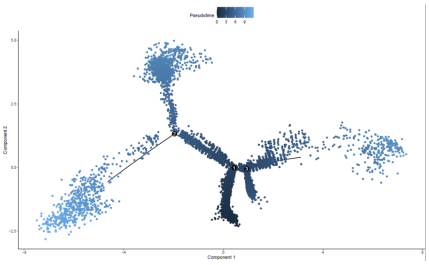

Figure 3.17: Reset the root

sequence at cell state 7.

In fact, the kind of biological function that results in each branch makes the trajectory analysis more attractive to us. These branches occur because cells execute different gene expression programs. During development, branches appear in the trajectory when cells make a fate choice: one developmental lineage follows one pathway while another generates a second. Redefining the root makes this question easier to answer by analyzing the biological significance of the different cell states according to branches 2 and 3. The analysis is based on BEAM, a statistical method used to find regulated branching-dependent genes.

After obtaining the heat map of the branch genes, these genes were extracted for enrichment analysis and selected the three most relevant GO terms to label each cluster. Figure 3.18 shows the analysis of branch 2, the cell states for comparison are cell state 1 and cell state 2. The purple color in the heat map indicates the degree of gene expression, which suggests that their pre branch, i.e., cell state 3, is more likely to exhibit Macrophage differentiation, Peptidyl-cysteine modification, and Bile acid metabolic process are biologically significant. In contrast, the expression of these goes terms gradually decreases to irrelevance as development progresses to cell states 1 and 2. The most relevant GO terms expressed in cell state2 are mRNA catabolic process, RNA catabolic process, and

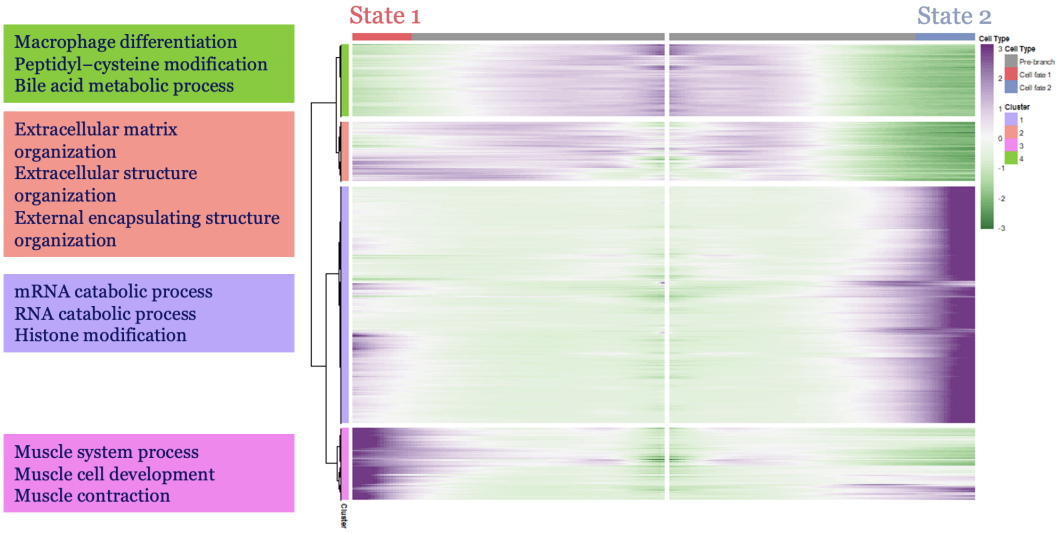

Figure 3.18: Heatmap of branch2

Histone modification, which are also related to the expression of DUX4; the most relevant GO terms expressed in cell state1 are Muscle system process, Muscle cell development, Muscle contraction, which seems to be more related to muscle development.

Figure 3.19 shows the analysis of branch 3, and we can see that the cell states for comparison are cell state 5 and cell state 6. The purple color in the heat map indicates the degree of gene expression, which suggests that their pre branch, i.e., cell state 4, expresses more of the biological significance of Cytoplasmic translation, Ribosome biogenesis, Ribonucleoprotein complex biogenesis. Ribosome biogenesis, Ribonucleoprotein complex biogenesis, and the expression of these go terms gradually decrease to irrelevant as development progresses to cell state 5 and cell state 6. The most relevant GO terms expressed in cell state6 are Regulation of mRNA metabolic process, Cellular response to leukemia inhibitory factor, and Response to leukemia inhibitory factor. The most relevant GO terms expressed in cell state5 are Muscle system process, Muscle cell development, and Muscle contraction, which seem to be more related to muscle development.

3.4 Visualization of shiny app

Finally, a proposed shiny app was created to allow researchers to check their single-cell datasets and find genes of interest freely. As seen in Figure 3.20, the shiny app currently has the ability to downscale visualization of different datasets and check genes.

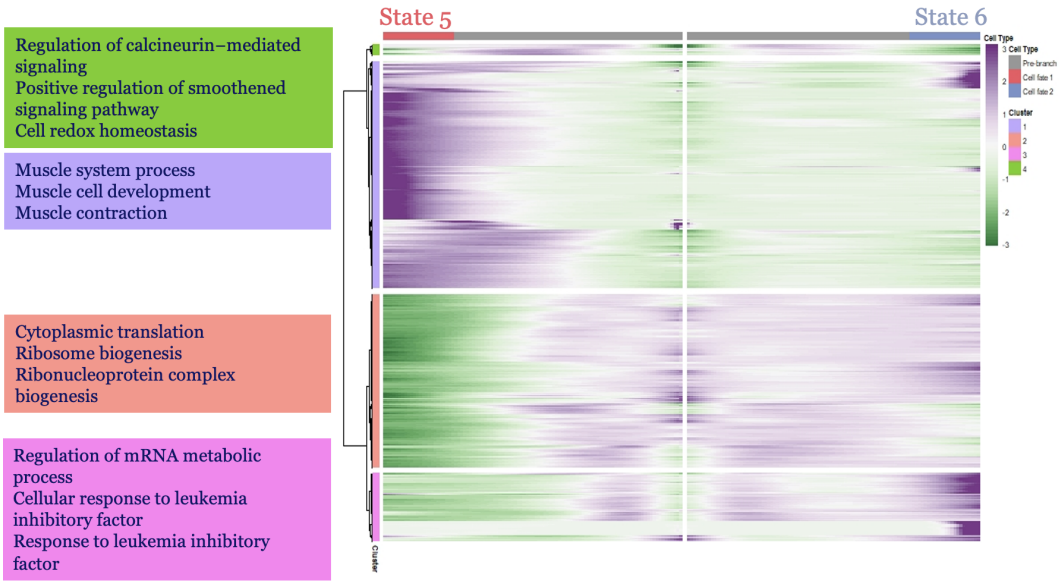

Figure 3.19: Heatmap of branch3

Figure 3.20: Shiny app function

Chapter 4

Discussion

In this project, to improve researchers' insight into transcriptional changes in muscle cells affected by FSHD, multiple data sets were integrated using different FSHD-related datasets available on NCBI as well as related datasets from the LUMC lab. The integrated analysis allows for the discovery of relationships across cellular patterns, understanding the overall representation of cellular states, and the ability to pool datasets generated across individuals and technologies [16]. This is also the reason why we wanted to integrate the data. The experiment was divided into four main parts, first by integrating all existing datasets to present a macroscopic molecular atlas of FSHD muscle cells. Then, a selection of data was integrated to focus more on the cellular heterogeneity in FSHD and the expression of DUX4 and its related genes. This part includes the integration of primary and induced cells from the single cell data as well as the single cell nuclear health control group and the integration of the nuclei of both FSHD1 and FSHD2 patient muscle cell types. And finally the shiny app was introuduced.

The integration of all data in the first part shows the successful atlas of FSHD. However, as seen from the figure, when using the unsupervised integration method, the data are still numerous and redundant despite eliminating some of the batch effects. Although, in general, atlas has broad physiological and medical implications, providing insight into the function, regulation, and interactions of known and new cell types. By drawing an atlas, each cell type can be identified, and each cell cluster can be classified to discover new cell clusters or find new marker genes. In turn, large-scale single-cell integration of certain cancers has been shown to deepen the understanding of the ecosystem of that can-

cer. It suggests ecosystem-based patient classification would facilitate precision medicine approaches to identify individuals for tumors and their immune environment [17]. In the following analysis, we can further categorize each cell cluster in FSHD and try to find previously unknown cell types or cell states. However, this project aims to study the heterogeneity of FSHD and, for this reason, focus more on the biological differences of FSHD. The datasets differ due to sequencing source and cell type masking the heterogeneity of FSHD, making it challenging to integrate precisely.

Therefore, selected portions of the data were used to explore the heterogeneity of FSHD further. Data from both primary and post-induction dux4 are used here. The combined data from the primary and DUX4i scRNA-seq data suggest that the cell states cause the main differences. The integration results show that there is indeed an overlap between the two datasets, which confirms the conjecture - that the introduction of DUX4i single-cell RNA-seq may capture FSHD-affected cells without detecting the feasibility of DUX4 target genes. This hypothesis is based on the possibility that DUX4 may be able to cross cell membranes via homologous domains and thus affect adjacent cells [14]. However, further analysis of the overlap revealed that the transcriptomic features of the overlapping gene markers and pathway GO terms were not associated with the pathology of FSHD. Most of the GO terms we obtained were associated with muscle development, and the cells throughout the overlap appeared in the initial stage of muscle cell development. While the pathways related to DUX4 expression (characteristic of FSHD pathology) are DUX4 activation disrupts RNA metabolism, including RNA splicing, surveillance, and transport pathways. Cell signaling, polarity, and migration pathways are also disrupted [18].

Two possible reasons are speculated here.

1. Maybe the DUX4i cells in Cluster-5 are newly differentiated cells.

2. Maybe the transcriptome change in the overlapping part between the two datasets is mainly caused by the myogenic effect rather than DUX4 activation.

However, the subsequent analysis will require targeted studies of genes within GO terms, where cells are cultured in the laboratory to knock out relevant genes to study the function of specific genes.

Integrating FSHD-affected nuclei created a relatively purified transcriptomic atlas of FSHD at single-nucleus level. The integration of single-cell nuclear data shows a clear double-edged sword. Although the desire to reduce artificially introduced transcriptional

biases leads to the desire to analyze at the single cell nucleus level. However, single-cell nuclear sequencing directly fragments the cells mechanically or chemically without introducing dissociation preferences, recovering as many cell types as possible and thus allowing for a more complete and comprehensive cellular profile. This is reflected by the higher number of cell clusters in figure 3.11. However, in this project, the main focus was on the cell states associated with FSHD, thus neglecting the others. In addition, pseudotemporal analysis can dissect the transcriptomic dynamics of FSHD and perform pathway analysis for different cellular states FSHD high, FSHD low, FSHD affected1, and FSHD affcted2. Through the result, the expression of DUX4 gene did not match well with the Go terms obtained from the marker. the reasons for getting this result may be :

1. Because BEAM analyzes branch-dependent genes, and we only extracted cells affected by FSHD, lacking normal primary cells, so we could only see the expression of some unrelated genes other than DUX4 expression.

2. Because the DUX4 target gene we used is not quite the same as the DUX4 target gene of FSHD high and FSHD low cells, we get different results.

A notable strength of this study is the joint use of multiple data sets to analyze cellular heterogeneity in FSHD. In the previous literature, all single-cell analysis was based on individual data, and few studies integrated all data. An integrated analysis would help to reveal cell type-specific responses to environmental and genetic perturbations and even standardize comparisons of disease phenotype and treatment status between patient samples. The results of the current study could serve to fill in the gaps in the existing literature. It added another research angle based on FSHD atlas to the problem analysis.

Acknowledgements

This research was made possible by support from the LUMC FSHD group and Leiden Univeisity LIACS faculty.

Bibliography

[1] A. M. DeSimone, A. Pakula, A. Lek, and C. P. Emerson Jr, *Facioscapulohumeral muscular dystrophy*, Comprehensive Physiology **7**, 1229 (2011).

[2] L. H. Wang and R. Tawil, *Current therapeutic approaches in FSHD*, Journal of Neuromuscular Diseases **8**, 441 (2021).

[3] P. E. Warburton, D. Hasson, F. Guillem, C. Lescale, X. Jin, and G. Abrusan, *Analysis of the largest tandemly repeated DNA families in the human genome*, BMC genomics **9**, 1 (2008).

[4] A. Mazher, *Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques*, Water **12**, 590 (2020).

[5] C. R. Banerji and P. S. Zammit, *Pathomechanisms and biomarkers in facioscapulohumeral muscular dystrophy: roles of DUX4 and PAX7*, EMBO Molecular Medicine **13**, e13695 (2021).

[6] H. Wu, Y. Kirita, E. L. Donnelly, and B. D. Humphreys, *Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis*, Journal of the American Society of Nephrology **30**, 23 (2019).

[7] M. J. Petrany, C. O. Swoboda, C. Sun, K. Chetal, X. Chen, M. T. Weirauch, N. Salomonis, and D. P. Millay, *Single-nucleus RNA-seq identifies transcriptional heterogeneity in multinucleated skeletal myofibers*, Nature communications **11**, 1 (2020).

[8] A. van den Heuvel, A. Mahfouz, S. L. Kloet, J. Balog, B. G. van Engelen, R. Tawil, S. J. Tapscott, and S. M. van der Maarel, *Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development*, Human molecular genetics **28**, 1064 (2019).

[9] C. Vanderplanck, A. Tassin, E. Ansseau, S. Charron, A. Wauters, C. Lancelot, K. Vancutsem, D. Laoudj-Chenivesse, A. Belayew, and F. Coppée, *Overexpression of the double homeodomain protein DUX4c interferes with myofibrillogenesis and induces clustering of myonuclei*, Skeletal muscle **8**, 1 (2018).

[10] K. D. Pruitt, T. Tatusova, and D. R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins*, Nucleic acids research **33**, D501 (2005).

[11] S. Jiang, K. Williams, X. Kong, W. Zeng, N. V. Nguyen, X. Ma, R. Tawil, K. Yokomori, and A. Mortazavi, *Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei*, PLoS genetics **16**, e1008754 (2020).

[12] A. Ashoti, A. Alemany, F. Sage, and N. Geijsen, *DUX4 induces a homogeneous sequence of molecular changes, culminating in the activation of a stem-cell-like transcriptional network and induction of apoptosis in somatic cells*, bioRxiv (2021).

[13] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, *Comprehensive integration of single-cell data*, Cell **177**, 1888 (2019).

[14] E. J. Lee et al., *Global analysis of intercellular homeodomain protein transfer*, Cell Reports **28**, 712 (2019).

[15] T. E. Bakken et al., *Single-nucleus and single-cell transcriptomes compared in matched cortical cell types*, PloS one **13**, e0209648 (2018).

[16] T. Stuart and R. Satija, *Integrative single-cell analysis*, Nature reviews genetics **20**, 257 (2019).

[17] J. Wagner et al., *A single-cell atlas of the tumor and immune ecosystem of human breast cancer*, Cell **177**, 1330 (2019).

[18] A. M. Rickard, L. M. Petek, and D. G. Miller, *Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways*, Human molecular genetics **24**, 5901 (2015).