UNIVERSITY OF CALIFORNIA,
IRVINE

Investigating mechanisms of pathogenesis in facioscapulohumeral muscular dystrophy

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Developmental & Cell Biology

by

Katherine Elizabeth Williams

Dissertation Committee:
Professor Ali Mortazavi, Chair
Professor Kyoko Yokomori
Professor Lee Bardwell
Assistant Professor Zeba Wunderlich
Assistant Professor Devon Lawson

2021

DEDICATION

To

the Williams family, Robert Jones and Jojo Williams

in recognition of their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My sincerest gratitude to my mentors, Professors Ali Mortazavi and Kyoko Yokomori, for giving me the opportunity to develop into the scientist I am today. Your questions and suggestions have helped to develop my approach to scientific problems. Thank you for challenging me.

I would especially like to thank the FSHD community of patients, caregivers, doctors and researchers. I am constantly amazed at the dedication, cooperation and curiosity in the community. I sincerely thank the patients and other donors who have provided us with precious resources.

To my mentors, Professors Dana Aswad and Michael Cumsky. Thank you for all of your help and guidance. You taught me how to understand results from the micro to the macro scale. My understanding of biotechnology has saved an experiment on more than one occasion.

To my friends, thank you for making UCI a place I belong. I truly enjoyed going to work because I got to be with all of you. You have helped me turn failures to successes and to celebrate both.

Thank you to my family for always supporting me. With you by my side, my dreams have always seemed like inevitabilities. Thank you for your constant encouragement and belief in me. For my Jojo, thank you for staying with me through the long days and reminding me to go outside.

For Robert, I am sorry you do not have a degree for all the work you’ve done these past five years. Being the significant other of a PhD student seems to be a full time job, and one that you have excelled at. You have been there for every bad day, every failed experiment, every success and every milestone. You have kept me balanced and taught me to let go of fruitless anxieties. I am a better scientist and a better person for it.

CURRICULUM VITAE

Katherine Elizabeth Williams

Education

Ph.D. in Developmental & Cell Biology, *University of California, Irvine*	2021
M.S. in Biotechnology, *University of California, Irvine*	2017
B.S. in Biology, Minor in Mathematics, *Georgetown University, Washington, DC*	2014

Publications

Williams, K., Yokomori, K., Mortazavi, A. (2021). Heterogeneous skeletal muscle cell and nucleus populations identified by single-cell and single-nucleus resolution transcriptome assays. Submitted to *Skeletal Muscle*.

Williams, K., Kong, X., Nguyen, N., McGill, C., Tawil, R., Yokomori, K., Mortazavi, A. (2021). Muscle group specific transcriptomic and DNA methylation differences related to developmental patterning influence FSHD. In preparation.

***Williams, K.**, *Jiang, S., Kong, X., Zeng, W., Nguyen, N., Ma, X., Tawil, R., Yokomori, K., Mortazavi, A. (2020). Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. *PLoS Genetics*. DOI: 10.1371/journal.pgen.1008754

Chau, J., Kong, X., Nguyen, N., **Williams, K**., Ball, M., Tawil, R., Kiyono, T., Mortazavi, A., Yokomori, K. (2021) Relationship of DUX4 and target gene expression in FSHD myocytes. *Human Mutation*. DOI: 10.1002/humu.2

*Rebboah, E., *Reese, F., **Williams, K**., Balderrama-Gutierrez, G., McGill, C., Trout, D., Rodriguez, I., Liang, H., Wold, B.J., Mortazavi, A. (2021). Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. *bioRxiv*. DOI: 10.1101/2021.04.26.441522 (Submitted to *Genome Biology*)

Carvalho, K., Rebboah, E., Jansen, C., **Williams, K.**, Dowey, A., McGill, C., Mortazavi, A. (2020). Uncovering the gene regulatory networks underlying macrophage polarization through comparative analysis of bulk and single-cell data. *bioRxiv*. DOI: 10.1101/2021.01.20.427499

Serra, L., Chang, D., Macchietto, M., **Williams, K.**, Murad, R., Lu, D., Dillman, A. R. and Mortazavi, A. (2018). Adapting the Smart-seq2 protocol for robust single worm RNA-seq. *Bio-protocol* 8(4): e2729. DOI: 10.21769/BioProtoc.2729

Awards & Honors

Susan V. Bryant Graduate Fellowship Award *University of California, Irvine*	2021
William D. Redfield Graduate Fellowship Award for excellence in research *University of California, Irvine*	2020
Graduate Fellowship *University of California, Irvine*	2016
Zukowski-Kolleng Fellowship for undergraduate research *Georgetown University, Washington, DC*	2012

Selected Presentations

FSHD International Research Conference *Invited Oral Presentation, FSH Society, Online Format*	2020
FSHD International Research Conference *Poster Presentation, FSH Society, Marseille, France*	2019
MDA Clinical & Scientific Conference *Invited Oral & Poster Presentation, Muscular Dystrophy Association, Orlando, FL*	2019
International *Caenorhabditis elegans* Conference *Poster Presentation, Genetics Society of America, Los Angeles, CA*	2017

Teaching Experience

Teaching Assistant, Genetics for undergraduates *University of California, Irvine*	2020
Teaching Associate, COSMOS for high school students *University of California, Irvine*	2018, 2019
Reader, Systems Biology and Cancer Systems Biology Short Course *University of California, Irvine*	2018, 2019
Teaching Assistant, Laboratory for M.S. in biotechnology *University of California, Irvine*	2016, 2017
Teaching Assistant, Genetics laboratory for undergraduates *Georgetown University*	2012, 2013

Certificates

Course Design from Department of Teaching Excellence and Innovation, *UC Irvine*	2020
Micro-MBA from Rady School of Management, *UC San Diego*	2020
The Complete Management Skills Course, *Eazl*	2020
Business Concepts for STEM Scientists, *GPS-STEM and Beall Applied Innovation, UC Irvine*	2020

Miscellaneous Experience

Biology Research & Development Intern, *Allergan PLC, Irvine, CA*	2016
Animal Care Intern, *CuriOdyssey, San Mateo, CA*	2014
Private Tutor, *Danville, CA*	2014, 2015

ABSTRACT OF THE DISSERTATION

Establishing mechanisms of pathogenesis in facioscapulohumeral muscular dystrophy

by

Katherine Williams

Doctor of Philosophy in Developmental & Cell Biology

University of California, Irvine, 2021

Professor Ali Mortazavi, Chair

Facioscapulohumeral muscular dystrophy (FSHD) is a rare disease with characteristic weakness in facial and periscapular muscles which progresses to additional muscle groups. FSHD is caused by the misexpression of the embryonic transcription factor DUX4 in muscle cells. In 95% of FSHD patients, a series of macrorepeats preceding *DUX4* is contracted and derepressed in part through loss of DNA methylation. In the remaining FSHD patients, the repeats are derepressed but not contracted, and 80% of these patients have a mutation in *SMCHD1*, which regulates DNA methylation in these repeats. DUX4 is involved in zygotic genome activation (ZGA) when it activates a number of transcription factors and chromatin remodelers, such as *DUXA*, *LEUTX* and *ZSCAN4*, as well as long terminal repeats (LTRs), such as ERVL-MaLRs, which are also activated in FSHD. *DUX4* expression in patient muscle cells is sparse (0.5% of myotube nuclei), but its expression in only a few nuclei is sufficient to activate target gene expression in multiple nuclei within a multinucleated muscle cell, which is sustained when DUX4 is no longer present. My work has focused on understanding progression of FSHD at a molecular level both into different muscle groups and following DUX4 activation.

I used single nucleus RNA sequencing to understand the contribution of individual nuclei to gene dysregulation following *DUX4* expression. I identified nuclei with native expression of

DUX4, as well as two populations of nuclei with high and low expression of DUX4-induced genes. The high group appears to perpetuate pathogenesis and has higher expression of genes related to the cell cycle despite the nuclei coming from cells in G0. I also found that *DUX4* is coexpressed with only a subset of its target genes, while the *DUX4* homolog *DUXA* is expressed with a wider set of targets.

To understand why certain muscle groups are commonly or less affected in FSHD, I assayed DNA methylation and gene expression in different muscle groups. Genes induced during myogenesis in FSHD have higher expression in commonly affected muscle groups despite their promoters having high DNA methylation. Muscle groups differ in expression and DNA methylation of transcription factors key to developmental patterning and specification that may contribute to susceptibility to FSHD.

Finally, I explored the role of *DUXA* as a potential regulator of DUX4 target genes following their initial activation. I found that *DUXA* depletion is sufficient to lower expression of DUX4 target genes including LTRs. I also identified a set of genes which are induced along with *DUX4* during myogenesis in FSHD2 that are not induced following *DUXA* depletion. I have thus identified a candidate regulator of FSHD gene dysregulation and candidate contributors to differential muscle group susceptibility in FSHD.

CHAPTER 1

Introduction
Heterogeneous skeletal muscle cell and nucleus populations identified by single-cell and single-nucleus resolution transcriptome assays

Chapter 1

Heterogeneous skeletal muscle cell and nucleus populations identified by single-cell and single-nucleus resolution transcriptome assays

1.1 Abstract

Single-cell RNA-seq (scRNA-seq) has revolutionized modern genomics, but the large size of myotubes and myofibers has restricted use of scRNA-seq in skeletal muscle. For the study of muscle, single-nucleus RNA-seq (snRNA-seq) has emerged not only as an alternative to scRNA-seq, but as a novel method providing valuable insights into multinucleated cells such as myofibers. Nuclei within myofibers specialize at junctions with other cell types such as motor neurons. Nuclear heterogeneity plays important roles in certain diseases such as muscular dystrophies. We survey current methods of high throughput single cell and subcellular resolution transcriptomics, including single-cell and single-nucleus RNA-seq and spatial transcriptomics, applied to satellite cells, myoblasts, myotubes and myofibers. We summarize the major myonuclei subtypes identified in homeostatic and regenerating tissue including those specific to fiber type or at junctions with other cell types. Disease-specific nucleus populations were found in two muscular dystrophies, FSHD and DMD, demonstrating the importance of performing transcriptome studies at the single nucleus level in muscle.

1.2 An overview of myogenesis

Skeletal muscle is the most abundant tissue in our bodies and is crucial for voluntary movement and support. Adult skeletal muscle tissue is composed primarily of mature muscle cells called myofibers and undifferentiated muscle cells called satellite cells. Myofibers can reach up to 30 cm in length in humans and 10 mm in mice and have hundreds of nuclei [1,2].

Muscle differentiation and fusion, called myogenesis, is controlled by a gene regulatory network well studied in mice [3,4]. Skeletal muscle is initially specified in embryonic development when the somite segments into the dermomyotome, and muscle specification begins with the expression of transcriptional regulators Pax3 and Myf5 around embryonic day 9 (E9) [4]. The muscle regulatory factors (MRFs) Myf5, MyoD, Mrf4 and Myog control the transition from cycling cells to postmitotic myocytes that go on to form primary fibers [4]. After the formation of these primary fibers, remaining cycling cells downregulate Pax3 and upregulate Pax7 [4]. These cells either fuse to each other forming new myocytes or to the primary fibers by turning on the transcription factor MyoD and then Myog [4]. Pax7+ cells that do not fuse will continue to cycle and become satellite cells [4]. Satellite cells become activated in embryogenesis by turning on MyoD and turning off Pax7 to form myoblasts which fuse to existing myofibers [4,5]. Following Myog expression at day E11.5 in mouse, Mrf4 (herculin/Myf6) is turned on at day E13.5 and controls final myofiber structure such as myonuclear positioning [6]. The major muscle specification and patterning is complete at this point [6]. Myoblasts continue to fuse to myofibers after birth to build the muscle until postnatal day 21 in mice [5]. At this point, the satellite cells also become quiescent, and muscle structure is established [4,5]. Several myofibers group together to form fascicles, and multiple fascicles make up the total muscle [7]. Each individual myofiber is surrounded by a matrix of connective proteins termed the basal lamina [8]. Satellite cells reside under the basal lamina in direct contact with the myofiber [4].

Myogenesis in adult muscle follows a similar trajectory as seen in development. Regeneration is stimulated upon injury when satellite cells become activated [9]. Quiescent satellite cells expressing Pax7 become activated, express MyoD and proliferate [9]. Some myoblasts continue to proliferate while others commit to differentiation [5]. These myoblasts turn on Myog after exiting the cell cycle such that early myotubes are marked by Myog expression [5], which is important for activation of numerous myogenesis genes and fusion [9]. The cells can fuse to existing myofibers but more commonly fuse to each other to form nascent myotubes (early differentiated, postmitotic muscle cells) and eventually myofibers [10]. Myog expression wanes and Mrf4 is expressed to control intracellular structure as in embryogenesis [9]. Understanding regeneration is important to the study of muscular diseases such as dystrophies or sarcopenia (the loss of muscle with age) that involve defects in muscle repair [11].

The myogenesis process up to myotube formation can be mimicked *in vitro* for human and mouse using primary myoblasts or stem cells isolated from tissue or from induced pluripotent stem cells (iPSCs) differentiated to form myoblasts. *In vitro* myoblasts are triggered to differentiate using serum or with ITS (insulin-transferrin-selectin) [12]. MYOD+ myoblasts then elongate and begin to fuse. MYOG+ myotubes make up a majority of cells by 72 hours post-induction with a number of mononuclear cells still present. Past 72 hours, myotubes will continue to grow and add more nuclei, but these myotubes cannot form full myofibers in 2D culture. A number of studies have worked on 3D culture to form something more akin to myofibers [13]. *In vitro* culture has provided an invaluable tool for studying skeletal muscle myogenesis. However, the systems lack interaction with other cell types such as neurons and tendons which are present in tissue and help to shape the muscle cells.

Neurons and tendons make direct contact with myofibers at specialized junctions. The nuclei along the length of the myotube are able to respond to local signals at the junctions thereby making nuclei within a myofiber heterogenous in terms of incoming signals and transcriptional output [14]. Neurons interact with muscle at the neuromuscular junction (NMJ). Nuclei clustered around NMJ transcribe specialized genes such as *ACHE* (acetylcholinesterase), which serves as a receptor for acetylcholine signals from the motor neuron [8,15]. Tendons interact with muscle at the myotendinous junction (MTJ) where myonuclei express collagens including *Col22a1* [16]. The MTJ is crucial for skeletal muscle and tendon development as the tendon guides the muscle to attach and muscle contraction maintains the tendon cells [17]. Myonuclei are therefore specialized within the myofiber.

The transcriptome of skeletal muscle has been relatively well studied from *in vitro* to *in vivo* systems. Several studies have assessed the transcriptome of muscle cells using microarray and RNA sequencing (RNA-seq) from mouse and human from both cell lines as well as biopsies [3]. These studies have provided valuable insights into myogenesis. For example, RNA-seq on iPSC derived myocytes found TWIST1 to be important for maintaining *PAX7* expression in satellite cells [18]. Additionally, the transcription factor Tead4 was found to regulate differentiation in the mouse skeletal myoblast line C2C12 [19]. Transcriptome studies have been especially valuable for disease studies such as muscular dystrophies and sarcopenia [20–22]. These studies have been crucial to our understanding of skeletal muscle biology and molecular mechanisms of disease. However, these studies have been limited in their findings due to limitations of profiling multiple heterogenous cells together. Pooling of cells in different states of differentiation for example can lead to averaging of expression from subsets of cells. To identify the timing of MRF expression, satellite cells were isolated by hand to avoid pooling cells in different states of myogenesis [23]. Recent advances in transcriptomics have enabled high throughput single-cell and single-nucleus RNA-seq and spatial transcriptomics to assay transcriptional heterogeneity at high resolution. In this review, we will go over current applications of high throughput, high-resolution transcriptomic techniques to the different stages of skeletal muscle cell differentiation in human and mouse. We will cover their uses in understanding basic biology of muscle cells as well as application to development, regeneration, aging and disease.

1.3 High-resolution transcriptome methods

Single-cell RNA-seq (scRNA-seq) revolutionized the field of transcriptomics, and multiple studies have applied this to assay mononucleated satellite cells [24–26]. However, application of scRNA-seq to skeletal muscle is challenging due to the size of myotubes and myofibers as most methods of scRNA-seq involve microfluidics that restrict the input cell size. To assay these larger cell types, researchers have turned to hand-picking cells or to single-nucleus RNA-seq (snRNA-seq). Alternatively, spatial transcriptomics can be used to preserve the spatial context of the transcriptomic data. We will briefly discuss the advantages and disadvantages of these technologies for use in skeletal muscle (Table 1.1).

Single-cell RNA-seq (scRNA-seq) has already been widely used for studying mononucleated muscle cells [27]. Both droplet and microfluidic based platforms have been used with sorted and unsorted mononucleated muscle cells from cell lines and satellite cells derived from tissues. This approach is relatively fast, and FACS sorting can help by filtering out low quality material such as debris, doublets and dead cells. ScRNA-seq has been useful in understanding myogenesis in more detail as cells in different states of differentiation are distinguishable at single cell resolution [28,29]. ScRNA-seq from tissue has the advantage of capturing multiple cell types within a certain size. Analyzing multiple skeletal muscle resident cell types, such as fibroadipogenic progenitors (FAPs) or tenocytes, gives us a better understanding of interactions, perturbations and responses of the whole tissue [24,25,30]. For mononuclear muscle cells, scRNA-seq is a fast, high-throughput and high-resolution way to profile the transcriptome.

Profiling the transcriptomes of mature muscle cells such as myotubes or myofibers can be done at the whole cell or nucleus level. Mature muscle cells can also be used for scRNA-seq, but isolation of individual cells has to be done manually due to their size which makes the method low throughput. SnRNA-seq is a high throughput alternative that involves lysing cells to isolate nuclei and to subsequently use them for RNA-seq. SnRNA-seq has been widely used for neurons since their long and intricate morphologies can cause them to clump or be too large for scRNA-seq microfluidics systems [31–34]. However, the adaptation of snRNA-seq to other cell types, such as skeletal muscle, has been slow. Since the first application of snRNA-seq in muscle in 2016 [35], only five papers have used snRNA-seq for skeletal muscle [36–40]. SnRNA-seq can be used for multiple cell types, for example isolated from a tissue, and is therefore able to profile both mononuclear and multinucleated cells together. For multinucleated cell types, snRNA-seq offers the additional advantage of resolving unicellular nuclear heterogeneity. This has proven useful for studies looking at transcriptomes of nuclei near neuromuscular junctions (NMJ) or myotendinous junctions (MTJ) [38,39]. However, snRNA-seq in multinucleated cells has the distinct complexity that we cannot determine which nuclei originate from the same cell. Additionally, isolating the nucleus means only a fraction of the transcripts within a cell are surveyed, so nuclear resident transcripts such as lncRNAs and pre-mRNA are enriched [35]. Nevertheless, snRNA-seq is a high-throughput, high-resolution method for transcriptome studies in mature skeletal muscle cell types.

Single-cell and nucleus RNA-seq have the important limitation that spatial information is lost as the cells or nuclei are removed from their native contexts for RNA isolation. The relative locations of cells and nuclei can be important for understanding cell type interactions and response. For example, the relative location of activated or quiescent satellite cells to pathological features such as fat deposits or immune infiltrates can be informative in disease context [41,42]. Myofiber nuclei are known to specialize based on their relative locations to non-muscle cell types [37,38]. Spatial transcriptomics enables us to associate RNA expression with morphological and physiological differences. Lower throughput *in situ* hybridization (ISH) methods have been used extensively in muscle to tag anywhere from one to four genes, such as with conventional RNA FISH and the original RNAscope. New medium throughput RNAscope methods are now available to assay tens of genes at a time [43]. High throughput methods have not yet been used in muscle. These include combinatorial barcoding FISH techniques, such as seqFISH and multiplexed error-robust fluorescence ISH (MERFISH), that are able to image thousands of genes from a single sample, and another method that involves imaging the sample and preparing it for sequencing while preserving the relative location of the transcripts, such as with 10X's Spatial Transcriptomics platform [44–46]. Spatial transcriptomics has an important advantage over single-cell and single-nucleus RNA-seq in preserving spatial information which is crucial to understand skeletal muscle structure and function. Whether from tissue or *in vitro* culture, muscle cells are heterogenous in size, number of nuclei and type including myofibers, myotubes and mononucleated cells. By preserving spatial information, expression data can be confidently paired with individual cell types. For multinucleated cells, spatial transcriptomics enables the identification of transcriptionally distinct nuclei originating from the same myofiber. This is an advantage no other method currently offers. Not only can we identify cells of origin, but how nuclei differ within a cell and how neighboring cells or environment affect transcriptional heterogeneity.

1.4 Current single-cell RNA-seq analyses of mononuclear muscle cells

Mononuclear muscle cells, namely satellite cells and myoblasts, are of great interest in understanding embryonic development of skeletal muscle and regeneration in adult tissue. Regeneration of muscle tissue through the activation of satellite cells is vital to restore damaged muscle, and its dysregulation is important to understand skeletal muscle diseases and muscle loss in aging, called sarcopenia. Several studies profiled skeletal muscle resident mononuclear cells with multiple cell types but do not make many conclusions with regards to muscle cells themselves [24,25]. However, scRNA-seq of muscle cells have identified significant heterogeneity (Table S1.1).

During mouse embryogenesis, muscle cells are specified and differentiate into *Pax7*+ cells which will form satellite cells in the adult as well as myocytes that fuse to form myofibers [4]. The satellite cell precursors express *Pax7* and its target *Msc* (Figure 1.1) [47]. The mature myocytes express *Tnnc* [47]. A set of cells in the embryonic limb bud express markers of both fibroblasts, *Col1a1* and *Osr1/2*, and muscle cells, *Myod1* and *Myog*, and may give rise to interstitial muscle fibroblasts (IMFs) which are able to differentiate to muscle [47].

During adult skeletal muscle regeneration, *Pax7*-positive (*Pax7+)* satellite cells exit quiescence and become activated [10,23]. They then proliferate with some cells retaining satellite cell identity while others differentiate to committed myoblasts that continue to divide. Myoblasts differentiate further to myocytes to rebuild lost muscle [10]. A number of studies have examined the regeneration of adult muscle cells and have identified *Pax7*+ satellite cells in both mouse and human using scRNA-seq (Reviewed in [27]). Satellite cells in quiescence express *Pax7* and *Btg2*, but upon activation express *Myod1* and *Myf5* as early activated cells [30,48,49].

Of note, quiescent satellite cells transcribe *Myod1* but do not translate it until activated [50]. When measured by RNAscope, 71% of satellite cells attached to myofibers expressed *Myod1* [51]. Quiescent satellite cells which resist activation were found in human muscle marked by expression of *CAV1* [52]. After satellite cell activation, primary myoblasts express cell cycle related genes and can progress to one of two populations [30]. Myoblasts which differentiate activate *Myog* and *Tnnt2*, while myoblasts which continue to proliferate express *Ccnd1/2* and *Ccnb2* [30,48]. Myoblast proliferation is affected by interactions with the cell surface receptor family Syndecans (Sdcs) that are expressed in a subset of quiescent and cycling muscle cells [48]. *Sdc4* is expressed in 100% of satellite cells attached to myofibers when measured by RNAscope, while *Pax7* and *Myf5* were found in 99% of satellite cells [51]. Most but not all of the transcriptional heterogeneity in mononucleated muscle cells is attributable to myogenesis.

During aging, muscle mass is lost and not regenerated leading to loss in strength [53]. Aged satellite cells have reduced differentiation abilities and therefore cannot restore lost fibers [54]. In spite of this, aged satellite cells are also more likely to exit quiescence [29,55]. Retinoic acid receptors help to maintain satellite cell quiescence but are lost with age [29]. The aged satellite cells follow the normal regeneration trajectory but are delayed in activation [28]. Upon activation, they upregulate genes related to stress, inflammation and immune response [28,29]. Transcription in aged satellite cells is uncoordinated possibly due to stochastic methylation differences between aged cells [55]. The variability in expression between cells leads to dysregulation of genes for interaction with the cell-niche [55].

Single-cell RNA-seq has enabled identification of new markers or pathways important to subsets of cells that were not observable with bulk sequencing methods. Future studies combining perturbations with scRNA-seq could validate the importance of these pathways in the subset of cells that are affected.

1.5 Single-cell and single-nucleus transcriptome analyses in differentiating myocytes

The single-cell studies described above provide valuable insights into satellite cells and myoblasts but do not look at differentiated cells. Myofibers from a mouse can be anywhere from 2 to 10 mm in length [2], while cultured differentiated myotubes can be over 100 um in length [13]. Most single-cell sequencing platforms can cells accommodate cells up to 60 um [56–58]. The restriction on size is mostly due to the microfluidics on these platforms that limit the size within the system to ensures that doublets or clumps of cells aren't isolated together, which would produce a false mixed "single cell" signal. These size limitations make high throughput scRNA-seq of differentiated muscle cells difficult. Yet, a number of creative solutions are available for high resolution RNA-seq from muscle cell types depending on the question of interest. In order to use microfluidics based methods, some researchers have used single nucleated "mature" myocytes formed by blocking fusion during differentiation using a calcium chelator [59]. These cells are smaller than myotubes and myofibers and therefore are easily captured on technologies such as 10X. For investigations into native mature muscle cells, the non-microfluidics based RNA-seq methods, snRNA-seq and spatial transcriptomics, are available.

Differentiation of human myoblasts followed by scRNA-seq was used to study myogenesis *in vitro* [60]. *In vitro* differentiation is asynchronous with cells differentiating at different rates such that markers of myotubes are present in some cells as early as 24 hours. Ordering cells by pseudotime arranged cells into a differentiation trajectory based on their transcriptome profiles rather than the differentiation time. This enabled the discovery that *ID1* has switch-like inactivation, which is followed by activation of *MYOG* [60]. *CUX1* and *USF1* were also identified as novel regulators of myogenesis [60]. scRNA-seq was able to parse out signatures of early myocytes, while larger, more differentiated myotubes were assessed using snRNA-seq.

Single-nucleus RNA-seq (snRNA-seq) in muscle cells was validated initially by comparing the transcriptomes of whole myoblasts to myoblast nuclei from in vitro culture [35]. Overall, the nuclear transcriptomes were found to faithfully recapitulate those of whole cells with the exception of enrichment for nuclear resident transcripts such as long non-coding RNAs [35]. snRNA-seq of in vitro human myoblasts differentiated into myotubes for 72 hours revealed a subset of nuclei expressing ID1, ID3, PDGFRA and SPHK1 which appear mesenchymal [35,60]. These nuclei were from mononucleated muscle cells (MNCs) that failed to fuse similar to the bifurcation of myoblast differentiation in vivo [35,60]. From mouse C2C12 culture differentiated for 72 hours, snRNA-seq identified 8 clusters of nuclei that express Pax7 and exhibit heterogeneity, which most likely represent unfused MNCs [61]. Two of the 8 clusters represent proliferating cells with high expression of the cell cycle gene Top2a in one and Lix1, which is important for satellite cell proliferation, in the other [61]. Another cluster of nuclei with significant expression of collagens and Fn1 appears similar to a population of stem cells found in vivo that remodel their extracellular matrix and trigger proliferation in neighboring satellite cells [61,62]. This ECM cluster was distinct from another cluster that expressed Itm2a and Pax7 and may be similar to activated satellite cells [61]. Through RNAscope, Myog expression was detected in MNCs in addition to multinucleated myotubes [61]. Nuclei from the most differentiated cells appear to either express Myog or Mef2c and could represent specialized myonuclei in culture [61]. Nuclear heterogeneity in in vitro culture provides evidence that myonuclei specialization is inherent and not solely due to contact with non-muscle cell types that is seen in tissue. In vitro studies may also provide evidence for the role of individual nuclei within multinucleated cells as the process of this specialization is poorly understood.

Single-cell and single-nucleus transcriptome assays in mature muscle

Single-cell sequencing of mature muscle cells is low throughput as myofibers need to be isolated by hand. However, scRNA-seq of myofibers was able to parse gene expression arising specifically from myofibers as opposed to other cell types in the muscle tissue, such as the satellite cell marker *Pax7* and the fibroblast marker *Col1a1* [63]. Comparison of whole myofibers isolated from old and young mice identified dysregulation of genes related to muscle growth and structure, such as *Actc1* and *Myl1*, collagen synthesis and metabolism which may contribute to age related muscle function decline [63]. scRNA-seq of myofibers is not scalable and blurs the transcriptional heterogeneity within the cells.

In the past two years, snRNA-seq has been applied to myofibers enabling discoveries into myofiber type and intranuclear heterogeneity. Muscle groups are made up of different types of muscle fibers, generally called fast and slow twitch. Each of these function slightly differently and express distinct genes and transcripts. Fast twitch fibers generally express *MYH2* (Type 2A), *MYH1* (Type 2X), *MYH4* (Type 2B) while slow fibers express *MYH7* (Figure 1.2) [8]. Single nucleus RNA-seq on muscle tissue have recovered myonuclei from each of these fiber types expressing the respective *Myh* [37–40]. Nuclei expressing different isoforms of *Myh4*, A, B or C, have distinct expression profiles and are found in differing proportions in different muscle groups [39]. Most nuclei express one *Myh* gene, as confirmed by RNAscope [39]. Nuclei expressing *Myh1* and *Myh2* are similar transcriptionally similar and are occasionally expressed in the same nucleus and often from the same allele [37,39]. These nuclei are mostly limited to the soleus whereas the majority of EDL myonuclei express only one *Myh* [39]. The expression of *Myh* genes is coordinated in nuclei across the length of the myofiber in quadricep and EDL, but not in soleus [39]. Innervation by motor neurons is required for coordinated *Myh* expression, and this coordination is activated in early postnatal development [39].

Myofibers interact with other cell types within the muscle tissue, notably neurons and tendons. Nuclei under NMJ and MTJ are transcriptionally distinct. NMJ nuclei express *Ache* (acetylcholinesterase) which process acetylcholine from neurons [15]. MTJ nuclei are known to express the collagen *Col22a1* [16]. snRNA-seq has revealed additional markers for these populations such as *Etv5*, *Etv4*, *Chrne*, *Colq*, *Musk*, *Ufsp1*, *Lrfn5*, *Ano4*, *Vav3* for NMJ nuclei and *Maml2*, *Ankrd1*, *Slc24a2*, *Adamts20* for MTJ nuclei [38,39] (Table S1.2). Additionally, *Ufsp1* and *Gramd1b* were found to regulate the specification of NMJ nuclei [38]. MTJ nuclei are also heterogeneous. MTJ nuclei expressing *Tigd4*, *Itgb1*, *Col24a1* and *Col22a1* were present in every fiber from adult mouse tibialis anterior. Only some fibers had MTJ nuclei which express *Pdgfrb*, *Ebf1*, *Col1a2*, *Col6a1*, and *Col6a3* [37]. Overall, NMJ nuclei make up 0.8% of the myonuclei from adult mouse tibialis anterior, while MTJ nuclei are about 3.6% [38].

Skeletal muscles contain myofibers, called intrafusal or spindle fibers, innervated by sensory neurons that are responsible for proprioception [64]. Spindle fiber nuclei can be marked by expression of *Calb1* [37]. Spindle fibers are classified as either bag or chain fibers based on the arrangement of nuclei in the cell [64]. Bag fiber nuclei express *Myh7b* and *Tnnt1*, while chain fiber nuclei are heterogeneous with expression of either *Myh13* or *Tnnt3* [37]. Spindle fibers also contain NMJ nuclei under motor neurons and MTJ nuclei. Areas in contact with sensory neurons contain densely packed nuclei which express *Calcrl* and are distinct from nuclei innervated by motor neurons [37].

Myonuclei from adult homeostatic muscle were found to have mild transcriptional differences beyond those attributable to fiber type or proximity to non-muscle cell types in mice [38]. However, some studies have identified heterogeneous populations of myonuclei present during development. One such population is found in the highest abundance in P21 mice, which is when fusion stops [38]. It is marked by expression of *Myh9, Flnc* and possibly *Runx1*, *Nrap*, *Fhod3*, *Enah*, *Myh10*, *Ifrd1*, *Nfat5*, *Mef2a*, *Ell*, *Creb5*, *Zfp697* with no expression of the fiber type specific *Myh4* and *Myh1* [38,39]. These nuclei are referred to as "sarcomere assembly states" due to the expression of genes related to "pre-myofibrils" used before mature myosins are in place [38]. The upregulation of the transcription factor *Atf3* and many of its target genes suggesting a role for Atf3 in these myonuclei during development [38]. Two other distinct myonucleus populations are present in P21 mouse muscle marked by expression of *Meg3* or *Nos1* [38]. In developing tissue at P10 when cells are still fusing, myocytes express *Myog* and *Mymk* which is crucial for fusion [38]. Some myonuclei appear to represent a transcriptional transition away from other cell types with expression of both myogenic markers *Ckm*, *Tnni2*, *Tnnt3* and ECM genes *Col1a1*, *Col3a1*, *Col5a3*, *Col6a1*, and *Dcn* [38]. These make up 4.7% of P10 myonuclei which are still developing and 0.8% of P21 myonuclei which have finished fusing [38]. Myofiber nuclear heterogeneity persists beyond muscle development to aging mouse muscle. The sarcomere assembly population present abundantly in P21 mice are also present in aged mice [38]. A subset of aged nuclei express *Ampd3* as well as genes for immune response and apoptosis [38]. These nuclei may represent dysfunction due to denervation [38].

Interestingly, some transcriptionally distinct nuclei express lncRNAs from the *Dlk1-Dio3* locus, specifically *Rian* and *Meg3* [37,38]. RNA FISH for *Rian* found that these nuclei are dispersed throughout myofibers [37]. Found on the outer edge of fibers near the perimysium are nuclei expressing *Muc14* and *Gucy2e* which may be specialized to help in adhesion [37]. Additional populations of heterogeneous nuclei have been identified but poorly described with expression of *Gssos2*, *Suz12* or *Bcl2* and possibly relating to the ER, epigenome or steroid synthesis, respectively [37].

Current snRNA-seq studies in myofibers have only been done on mouse. Studies in human biopsies could reveal new nuclear heterogeneity specific to humans. Additionally, skeletal muscle groups throughout the body have different fiber type compositions and other differences that could be surveyed using snRNA-seq. For example, Dos Santos, et al. found different levels of *Myh* co-expression in the soleus than in the EDL [39]. Muscular dystrophies often affect some muscle groups more severely than others. Thus, snRNA-seq analyses in multiple muscle groups may reveal factors contributing to susceptibility.

1.6 Use of high-throughput and high-resolution transcriptome methods to study muscular dystrophies

1.6.1 Facioscapulohumeral muscular dystrophy (FSHD)

Cellular heterogeneity is known to play an important role in some disease contexts, such as facioscapulohumeral muscular dystrophy (FSHD). FSHD is linked to the misexpression of an embryonic transcription factor, *DUX4* [65,66]. DUX4 misexpression causes downstream dysregulation of embryonic genes and retrotransposons such as ERVLs [66–69]. Previous studies have sought to identify the patient-specific transcriptome using bulk RNA-seq using multiple cells [70,71]. However, *DUX4* is rarely detected at the protein or RNA level in patient muscle (0.5% of patient myotube nuclei) [72]. Bulk RNA-seq averages out the signal from the few muscle cells that express *DUX4*. Other studies have used artificial expression of *DUX4* to identify a DUX4 gene signature, but these systems do not replicate the expression dynamics seen in native patient cells [20,67]. With its fine resolution, scRNA-seq and snRNA-seq can be used to look at native expression of *DUX4* and its downstream genes without averaging out over non-expressing cells or nuclei.

With scRNA-seq on fusion inhibited 72 hour differentiated myocytes, between 0.2 to 0.9% of FSHD cells were found to express *DUX4,* which is higher than the reported DUX4 protein in 0.29 to 4.28% of fusion inhibited myocytes [59,71]. These results may be entirely due to the high amount of variability in different patient cell lines in expressing *DUX4*, technical factors such as dropout, or as the result of additional stress from fusion inhibition [73]. *DUX4* expression has been suggested to be burst-like and to cause immediate cell death, which may account for its rare detection [71,74]. DUX4 target genes were more readily detectable than *DUX4* which may suggest a transient burst of *DUX4* expression followed by more sustained activation of downstream pathways [59]. Comparison of these affected cells to other non-affected FSHD cells identified dysregulation of transcriptional regulators and confirmed the dysregulation of pathways previously identified to be affected by DUX4 from bulk RNA-seq studies [70,71]. scRNA-seq enabled identification of transcriptional regulators that could activate *DUX4* expression or aid in gene dysregulation following DUX4 expression.

With snRNA-seq of native multinucleated FSHD patient myotubes, *DUX4* transcript was detected in 0.1% of nuclei, which is much higher than the results of fusion-inhibited scRNA-seq [36]. Interestingly, RNAscope detection of *DUX4* revealed its accumulation in the nucleus as foci [36,75]. Thus, sequencing the nuclear population of RNA might have made it easier to detect *DUX4* transcript. DUX4-affected nuclei made up 3.7% which is much higher than DUX4-expressing nuclei and higher than the 0.55% of DUX4-affected fusion inhibited cells from scRNA-seq [36,59]. This further supports the spreading of DUX4 protein to multiple nuclei in the same myotube to activate target genes [71,72,75]. However, *in situ* RNA detection revealed the expression of target genes without detectable DUX4 transcript or protein in some of the patient myotubes, raising the possibility that, once activated by DUX4, target gene expression may be maintained in these myotubes even in the absence of DUX4 [36,75]. This sustained target gene expression suggests the contribution of additional transcriptional regulators in perpetuating target gene expression. Indeed, the *DUX4* homolog and target gene, *DUXA*, is expressed in many more nuclei than *DUX4* in patient myotube nuclei and can regulate the expression of at least two DUX4 target genes, *LEUTX* and *ZSCAN4* [36]. Depletion of DUXA or LEUTX inhibited the expression of these genes (for DUXA) and *KDM4E* (for both) in late but not early differentiation, indicating that DUX4 target genes themselves participate in the maintenance of the DUX4 gene network [36]. Additionally, two populations of FSHD nuclei are apparent by snRNA-seq and RNAscope separated by high (FSHD-Hi) or low (FSHD-Lo) DUX4 target gene expression [36]. The FSHD-Hi nuclei appear to inappropriately activate cell cycle genes despite the myotubes having entered G0. Despite sparser DUX4 target gene expression in FSHD-Lo nuclei, their transcriptome is distinct from that of control nuclei, suggesting that patient myocytes are altered even in the absence of DUX4 and target gene expression [36,59]. The identification of these populations of nuclei and the precise co-expression of *DUX4* with its target genes in native, fused myotube nuclei was only possible with nuclear resolution afforded by snRNA-seq and spatial transcriptomics. Overall, the use of single-cell and single-nucleus

RNA-seq in conjunction with *in situ* RNA FISH helped identify pathogenic populations of nuclei that express *DUX4* and the factors contributing to its expression and disease progression.

1.6.2 Duchenne's muscular dystrophy (DMD)

Duchenne's muscular dystrophy is the most common form of muscular dystrophy arising from nonsense mutations in the structural protein dystrophin leading to its loss of function [76,77]. The altered dystrophin mainly affects myofibers in which dystrophin connects the center of the cell to the membrane [77]. The MDX mouse model of DMD produces a truncated form of dystrophin and is a popular model for studying DMD [78]. This model provides a way to assess the disease pathology in active muscle which includes cell death and muscle repair [41,79,80]. Accordingly, snRNA-seq and RNAscope of MDX myofibers found that a subset of nuclei near sites of damage appear to activate repair with co-expression of *Flnc* and *Xirp1* [37]. Notably, these nuclei appear similar to nuclei identified previously in P21 mice and aged muscle and are thought to be involved in sarcomere assembly [38]. This subset is also observed in biopsies from DMD and patients with mutations in dysferlin which plays a role in myofiber membrane repair [37,81]. MDX myofibers also have nuclei that are associated with dying myofibers with leaky membranes [37]. These nuclei express an abundance of noncoding transcripts, and infiltrating cells, thought to be macrophages, are nearby [37]. Macrophage infiltration into skeletal muscle can exacerbate DMD [82]. Interestingly, MDX myofibers lack substantial NMJ nuclei which is consistent with the disruption of NMJ structure in the MDX mouse model [37].

The identification of apparently dystrophic specific nuclear populations holds promise for future investigation into their roles in pathogenesis. These subcellular signatures provide biomarkers for active disease that are otherwise not observable at the whole cell level.

Understanding that multinucleated cells contain disease associated specialized myonuclei has important implications for conclusions drawn from bulk RNA-seq studies as signatures of these nuclei are lost. Additionally, single-cell RNA-seq on mononucleated cells from disease models or patients is not sufficient to fully understand disease pathology in myofibers. Due to heterogeneity of the nuclei, high resolution transcriptome assays in skeletal muscle should be performed with nucleus level resolution.

1.7 Future prospects

Single nucleus resolution transcriptome methods in muscle have the advantage of being able to answer fundamental questions about skeletal muscle that are unanswerable by other methods. Muscle development is known to be mutually reliant on tenocyte development with mechanical stress stimulating differentiation [83,84]. High resolution transcriptomic methods can help reveal how mechanical force is translated to gene expression changes in specific nuclei to stimulate maturation.

Disorders such as sarcopenia, FSHD and DMD affect the transcriptomes of subsets of cells and nuclei that are important for understanding pathogenesis [36,37,55]. Additional neuromuscular diseases are known to affect myofiber nuclei. Myofibers from spinal muscular atrophy (SMA) have centrally localized nuclei and loss of innervation [85]. Understanding how denervation affects subpopulations of nuclei, such as in the NMJ, in SMA would require single nucleus resolution techniques [86].

To address the mechanism of nucleus specialization, single nucleus transcriptome methods need to be combined with additionally assays. Assays for single-cell ATAC-seq, DNA methylation sequencing and ChIP-seq are already available, and some of these assays can be combined with RNA-seq to be done from the same cell or nucleus [87–92]. SnATAC-seq on myofiber nuclei has already identified potential transcriptional regulators of fiber type [39]. Single cell DNA methylation has identified stochastic methylation changes which result in alterations to transcriptional networks in aged muscle cells [55]. Using additional high-resolution methods in combination with snRNA-seq can help to identify heterogenous gene regulatory networks acting across nuclei within the same cell.

1.8 Conclusions

Skeletal muscle is composed of heterogeneous mononuclear cells and myofibers with transcriptionally specialized nuclei. This level of diversity is finally being discovered using high throughput, high resolution transcriptome studies. Mononucleated cells are mostly heterogeneous due to their state in myogenesis, and sarcopenia appears to affect activation of myogenesis. Upon differentiation *in vitro*, sets of cells do not fuse to form myotubes but instead remain mononucleated. Nuclei in fused myotubes begin to show specialization in *in vitro* culture. Nuclei from mature myofibers show a high degree of specialization depending on fiber type and proximity to other cell types. However, not all nuclear heterogeneity is attributable to these differences. Sets of nuclei appear to be transcriptionally distinct with specific sets of marker genes such as *Flnc* or lncRNAs. Diseases such as FSHD and DMD also cause myofiber nuclei to specialize for example by activation of pathogenic genes or in response to damage. These studies have only begun to unlock the heterogeneity of myonuclei and demonstrate the importance of surveying myofibers at the level of the nucleus.

In **Chapter 2**, I apply snRNA-seq to 3-day and 5-day differentiated primary FSHD myotubes to identify transcriptional signatures of affected nuclei. First, I use a time course of differentiation to identify genes specifically upregulated in FSHD over time including DUX4 target genes. From the snRNA-seq, I characterize two populations of FSHD nuclei as Hi and Lo based on expression of these genes. The FSHD-Hi population have higher expression of cell cycle related genes, which is unusual as differentiated myotubes have entered G0. We notably recover 13 nuclei that express DUX4 and have limited coexpression with its target genes. We detect 120 nuclei that express the *DUX4* homolog *DUXA*, which is also coexpressed with a much larger set of FSHD-induced genes. We also find that DUXA depletion results in decreased expression of two DUX4 target genes LEUTX and ZSCAN4, suggesting that DUXA may play a role in perpetuating gene dysregulation following initial activation by DUX4.

In **Chapter 3**, I identify differences in gene expression and DNA methylation associated with muscle groups with differential susceptibility to FSHD. I use primary myoblasts from the quadricep and tibialis anterior (TA) of healthy individuals and FSHD2 patients to perform RNA-seq and to investigate changes in DNA methylation using bisulfite sequencing over a time course of differentiation. I find that TA and quadricep retain differences in DNA methylation and expression of key TFs related to developmental patterning. We also find that more commonly affected muscles such as TA and bicep have higher expression of FSHD-induced genes than muscles affected later or less frequently such as quadricep and deltoid. TA also has higher expression of LTRs such as ERVL-MaLR, which are also hypomethylated. Additionally, more commonly affected groups differ from less affected in expression of developmentally associated TFs and genes involved in WNT signaling.

In **Chapter 4**, I assess the contributions of DUX4 target genes in regulating FSHD gene dysregulation. I find DUXA is able to regulate a substantial portion of FSHD-induced genes including LTRs. Additionally, I identify inherent differences in chromatin accessibility and gene expression between FSHD2 and control preceding *DUX4* expression. These include accessibility changes related to SMCHD1 mutations as well as for a lncRNA that regulates PRC2 binding. Genes induced following DUX4 expression are not more accessible before its expression, and regions less accessible in FSHD2 are enriched for motifs of transcriptional regulators activated by DUX4 including DUX4 itself. I also used gene network analysis to find a set of genes with similar upregulation in FSHD2 during myogenesis that may be targets of DUX4 and that are not significantly upregulated upon *DUXA* depletion.

Finally, in **Chapter 5**, I discuss further questions with regards to susceptibility and progression of FSHD. I propose additional experiments to assess the regulatory networks active in different muscle groups. In light of my findings from chapters 2 and 4, I discuss the potential therapeutic limitations of targeting DUX4 in FSHD. Importantly, I acknowledge the limitations of our *in vitro* system and suggest experiments to resolve the contributions of extracellular signals and non-muscle cell types. This includes the use of high resolution assays for the study of FSHD and skeletal muscle in general in which spatial resolution is key.

1.9 Figures

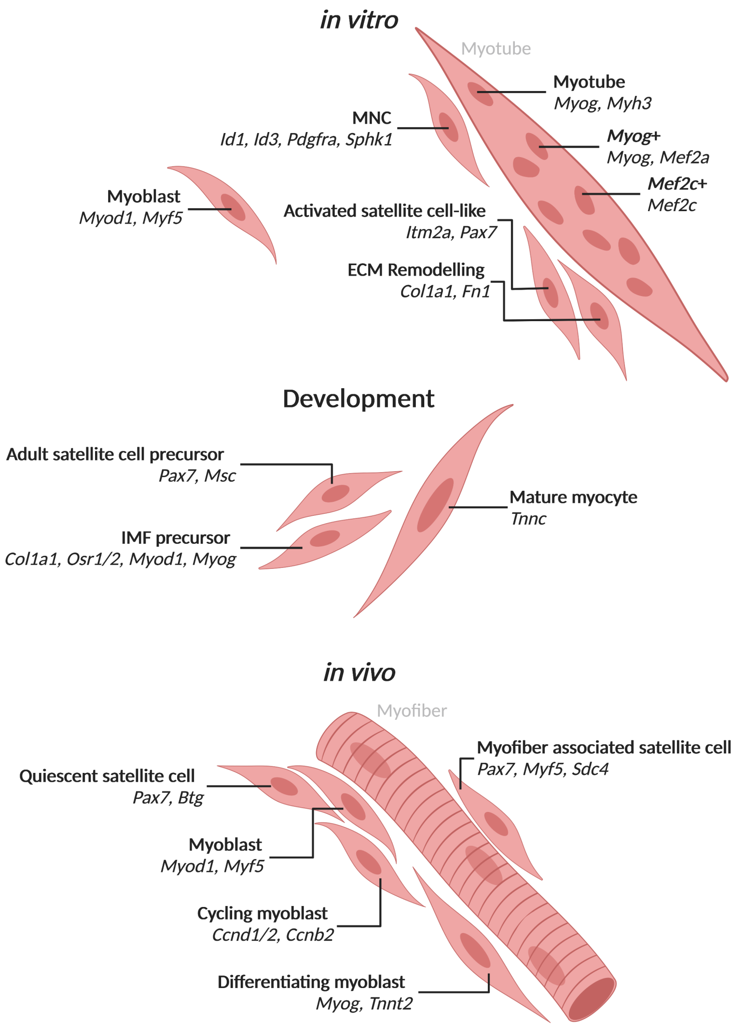


Figure 1.1: Populations of skeletal muscle cells and nuclei with associated markers identified by single-cell and single-nucleus RNA-seq Populations of cells and nuclei identified from [30,35,47–49,51,60,61]**.** Created with BioRender.com.

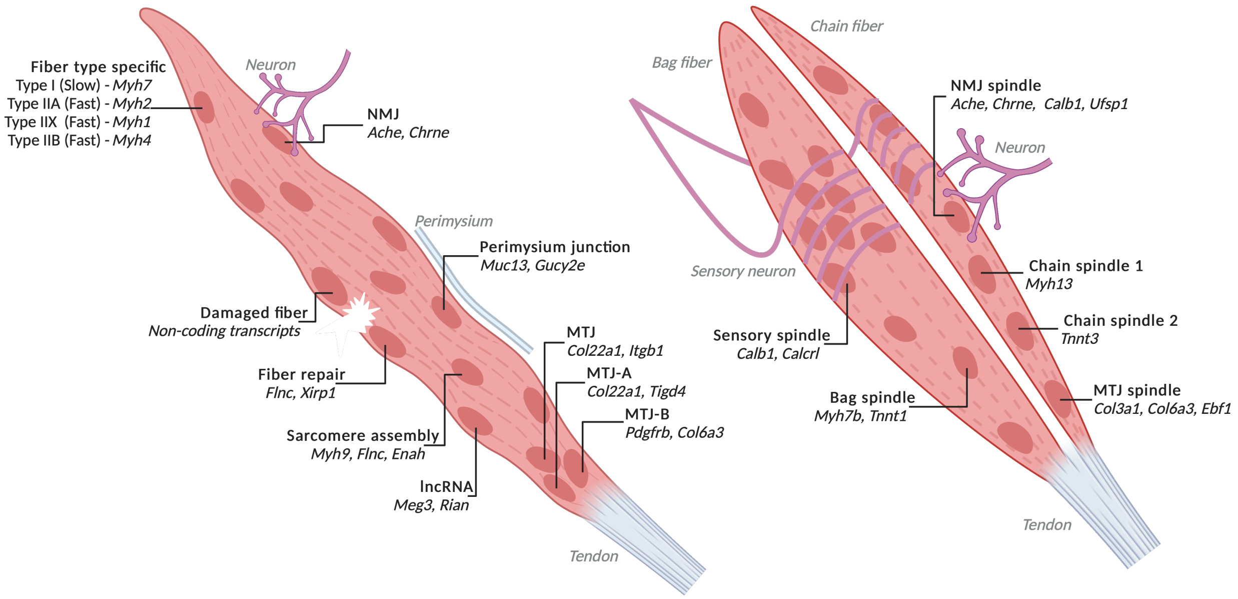


Figure 1.2: Heterogenous myofiber nuclei with associated markers identified by single-nucleus RNA-seq and RNAscope Populations of nuclei identified from [37–39]. Created with BioRender.com.

1.10 Tables

Table 1.1: Advantages and limitations of high throughput transcriptome methods for skeletal muscle cell types

	Mononucleated cells (Satellite cells and myoblasts)			Myotubes			Myofiber	
	scRNA-seq	snRNA-seq	Spatial transcriptomics	scRNA-seq	snRNA-seq	Spatial transcriptomics	scRNA-seq	snRNA-seq
Muscle tissue	Yes[1]	Yes[1]	Yes	Yes[1]	Yes	Yes	Yes[3]	Yes
Cell lines	Yes	Yes	Yes	Yes[2]	Yes	Yes	Yes[3]	Yes
Throughput	High	High	Low to medium[4]	High	High	Low to medium[4]	Low	High
Resolution	High	High	Low to high[4]	High	High	Low to high[4]	High	High
Assay multiple cell types at once	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Obtain spatial expression information	No	No	Yes	No	No	Yes	No	No
Assay nuclear heterogeneity	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Size restricted	Yes	No	No	Yes	No	No	Yes	No
Survey cytoplasmic transcripts	Yes	No	Yes	Yes	No	Yes	Yes	No
Associate nuclei with cell of origin	Yes	No	Yes	*NA*	No	Yes	*NA*	No
References	(24–34)	(30,35–38)	(36,39,40)	(40,41)	(29,30)	(29,42)	(43)	(35–38)

[1] Due to size restrictions or to isolate specific cell type, sample must be specially prepared by filtering or FACS sorting.
[2] Due to size restrictions or to isolate specific cell type, sample must be specially prepared by filtering or FACS sorting alone or in combination with fusion inhibition.
[3] To survey the whole myofiber or fiber bundle (from 3D culture), fibers must be hand picked.
[4] Throughput and resolution depend on the method used.

Table S1.1: Populations of heterogeneous mononuclear cells and myotube nuclei

	Top Markers	**References**
in vitro Populations		
Myoblast	*Myod1, Myf5*	[35]
MNCs	*Id1, Id3, Pdgfra, Sphk1*	[35,60]
ECM remodeling	*Col1a1, Fn1*	[61]
Myotube nucleus	*Myog, Myh3*	[35,60]
Myog+ myotube nucleus	*Myog, Mef2a*	[61]
Mef2c+ myotube nucleus	*Mef2c*	[61]
Development Populations		
Adult satellite cell precursor	*Pax7, Msc*	[47]
IMF precursor	*Col1a1, Osr1/2, Myod1, Myog*	[47]
Mature myocyte	*Tnnc*	[47]
in vivo Populations		
Quiescent satellite cell	*Pax7, Btg*	[30,48]
Human specific quiescent satellite cell	*CAV1, SPRY1, HEY1*	[52]
Myoblast	*Myod1, Myf5*	[23,30]
Cycling myoblast	*Ccnd1/2, Ccnb2*	[30,48,49]
Differentiating myoblast	*Myog, Tnnt2*	[30,48,49]
Myofiber associated satellite cell	*Pax7, Myf5, Sdc4*	[51]

Table S1.2: Populations of heterogeneous myofiber nuclei

	Top Markers	**Additional Markers**	**References**
Fiber Type Specific Populations			
Type I (Slow)	*Myh7*	NA	[39]
Type IIA (Fast)	*Myh2*	NA	[39]
Type IIX (Fast)	*Myh1*	NA	[39]
Type IIB (Fast)	*Myh4*	NA	[39]
Subcellular Localized Populations			
NMJ	*Ache, Chrne*	*Etv5, Musk, Lrp4, Colq, Chrna1, Prkar1a, Etv4, Ufsp1, Lrfn5, Ano4, Vav3, Ablim2, Phldb2, Irf8*	[37–39]
MTJ	*Col22a1, Itgb1*	*Slc24a2, Adamts20, Ankrd1, Maml2, Col24a1, Tigd4, Col1a2, Col6a1, Col6a3, Pdgfrb, Ebf1*	[37–39]
MTJ-A	*Tigd4, Col22a1*	*Itgb1, Col24a1, Smad3*	[37]
MTJ-B	*Pdgfrb, Col6a3*	*Ebf1, Col1a2, Col6a1*	[37]
Perimysium junction	*Muc13, Gucy2e*	*NA*	[37]
Homeostatic Populations			
Sarcomere assembly	*Myh9, Flnc, Enah*	*Runx1, Nrap, Fhod3, Myh10, Ifrd1, Nfat5, Mef2a, Ell, Creb5, Zfp697, Atf3*	[38,39]
lncRNA	*Meg3, Rian*	*Mirg*	[37,38]
Damage Associated Populations			
Damaged fibers	*Gm10801, Gm10717*	NA	[37]
Fiber repair	*Flnc, Xirp1*	NA	[37]
Spindle Fiber Populations			
Bag spindle	*Myh7b*	*Tnnt1, Piezo2*	[37]
Chain spindle 1	*Myh13*	NA	[37]
Chain spindle 2	*Tnnt3*	NA	[37]
NMJ spindle	*Chrne, Ache, Calb1*	*Ufsp1, Piezo2*	[37]
MTJ spindle	*Col3a1, Col6a3, Ebf1*	*Calb1, Col6a1*	[37]

Sensory spindle	*Calb1, Calcrl*	NA	[37]

1.11 References

1. Konno T, Suzuki A. Myofiber Length and Myofiber Arrangement in the Antebrachial and Leg Muscles of Sheep. Okajimas Folia Anat Jpn. 2000;77(1):5–10.
2. Griffin GE, Williams PE, Goldspink G. Region of longitudinal growth in striated muscle fibres. Nat New Biol. 1971;232(27):28–9.
3. Blais A. Myogenesis in the Genomics Era. Vol. 427, Journal of Molecular Biology. Academic Press; 2015. p. 2023–38.
4. Chal J, Pourquié O. Making muscle: Skeletal myogenesis in vivo and in vitro. Dev. 2017;144(12):2104–22.
5. Relaix F, Zammit PS. Satellite cells are essential for skeletal muscle regeneration: The cell on the edge returns centre stage. Vol. 139, Development (Cambridge). Oxford University Press for The Company of Biologists Limited; 2012. p. 2845–56.
6. Huang AH. Coordinated development of the limb musculoskeletal system: Tendon and muscle patterning and integration with the skeleton. Vol. 429, Developmental Biology. Elsevier Inc.; 2017. p. 420–8.
7. Jorgenson KW, Phillips SM, Hornberger TA. Identifying the Structural Adaptations that Drive the Mechanical Load-Induced Growth of Skeletal Muscle: A Scoping Review. Cells. 2020;9(7).
8. Schiaffino S, Reggiani C. Fiber Types in Mammalian Skeletal Muscles. Physiol Rev. 2011 Oct;91(4):1447–531.
9. Zanou N, Gailly P. Skeletal muscle hypertrophy and regeneration: Interplay between the myogenic regulatory factors (MRFs) and insulin-like growth factors (IGFs) pathways. Vol. 70, Cellular and Molecular Life Sciences. Springer; 2013. p. 4117–30.
10. Rochlin K, Yu S, Roy S, Baylies MK. Myoblast fusion: When it takes more to make one. Vol. 341, Developmental Biology. Academic Press Inc.; 2010. p. 66–83.
11. Robinson DCL, Dilworth FJ. Epigenetic Regulation of Adult Myogenesis. In: Current Topics in Developmental Biology. Academic Press Inc.; 2018. p. 235–84.
12. van der Wal E, Herrero-Hernandez P, Wan R, Broeders M, in 't Groen SLM, van Gestel TJM, et al. Large-Scale Expansion of Human iPSC-Derived Skeletal Muscle Cells for Disease Modeling and Cell-Based Therapeutic Strategies. Stem Cell Reports. 2018 Jun 5;10(6):1975–90.
13. Jones JM, Player DJ, Martin NRW, Capel AJ, Lewis MP, Mudera V. An Assessment of Myotube Morphology, Matrix Deformation, and Myogenic mRNA Expression in Custom-Built and Commercially Available Engineered Muscle Chamber Configurations. Front Physiol. 2018 May 8;9(MAY):483.
14. Wu H, Xiong WC, Mei L. To build a synapse: Signaling pathways in neuromuscular junction assembly. Development. 2010;137(7):1017–33.
15. Jacobson C, Côté PD, Rossi SG, Rotundo RL, Carbonetto S. The dystroglycan complex is necessary for stabilization of acetylcholine receptor clusters at neuromuscular junctions and formation of the synaptic basement membrane. J Cell Biol. 2001 Apr 30;153(3):435–50.
16. Charvet B, Guiraud A, Malbouyres M, Zwolanek D, Guillon E, Bretaud S, et al. Knockdown of col22a1 gene in zebrafish induces a muscular dystrophy by disruption of the myotendinous junction. Dev. 2013 Nov 15;140(22):4602–13.
17. Subramanian A, Schilling TF. Tendon development and musculoskeletal assembly: emerging roles for the extracellular matrix. Development. 2015 Dec 15;142(24):4191–

204.

18. Choi IY, Lim H, Cho HJ, Oh Y, Chou BK, Bai H, et al. Transcriptional landscape of myogenesis from human pluripotent stem cells reveals a key role of TWIST1 in maintenance of skeletal muscle progenitors. Elife. 2020 Feb 1;9.
19. Benhaddou A, Keime C, Ye T, Morlon A, Michel I, Jost B, et al. Transcription factor TEAD4 regulates expression of Myogenin and the unfolded protein response genes during C2C12 cell differentiation. Cell Death Differ. 2012 Feb 24;19(2):220–31.
20. Jagannathan S, Shadle SC, Resnick R, Snider L, Tawil RN, van der Maarel SM, et al. Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. Hum Mol Genet. 2016 Aug 17;ddw271.
21. Gonorazky H, Liang M, Cummings B, Lek M, Micallef J, Hawkins C, et al. RNAseq analysis for the diagnosis of muscular dystrophy. Ann Clin Transl Neurol. 2016 Jan 1;3(1):55–60.
22. Su J, Ekman C, Oskolkov N, Lahti L, Ström K, Brazma A, et al. A novel atlas of gene expression in human skeletal muscle reveals molecular changes associated with aging. Skelet Muscle. 2015 Oct 9;5(1).
23. Cornelison DDW, Wold BJ. Single-cell analysis of regulatory gene expression in quiescent and activated mouse skeletal muscle satellite cells. Dev Biol. 1997;191(2):270–83.
24. Rubenstein AB, Smith GR, Raue U, Begue G, Minchev K, Ruf-Zamojski F, et al. Single-cell transcriptional profiles in human skeletal muscle. Sci Rep. 2020;10(1):1–15.
25. Giordani L, He GJ, Negroni E, Sakai H, Law JYC, Siu MM, et al. High-Dimensional Single-Cell Cartography Reveals Novel Skeletal Muscle-Resident Cell Populations. Mol Cell. 2019 May 2;74(3):609-621.e6.
26. Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727):367–72.
27. Saber J, Lin AYT, Rudnicki MA. Single-cell analyses uncover granularity of muscle stem cells. Vol. 9, F1000Research. F1000 Research Ltd; 2020.
28. Kimmel JC, Hwang AB, Scaramozza A, Marshall WF, Brack AS. Aging induces aberrant state transition kinetics in murine muscle stem cells. Dev. 2020;147(9).
29. Shcherbina A, Larouche J, Fraczek P, Yang BA, Brown LA, Markworth JF, et al. Dissecting Murine Muscle Stem Cell Aging through Regeneration Using Integrative Genomic Analysis. Cell Rep. 2020;32(4):107964.
30. Dell’Orso S, Juan AH, Ko KD, Naz F, Perovanovic J, Gutierrez-Cruz G, et al. Single cell analysis of adult mouse skeletal muscle stem cells in homeostatic and regenerative conditions. Dev. 2019 Jun 1;146(12).
31. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017 Oct 1;14(10):955–8.
32. Lake BB, Codeluppi S, Yung YC, Gao D, Chun J, Kharchenko P V., et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Sci Rep. 2017 Dec 1;7(1):1–8.
33. Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. Soriano E, editor. PLoS One. 2018 Dec 26;13(12):e0209648.
34. Krishnaswami SR, Grindberg R V., Novotny M, Venepally P, Lacar B, Bhutani K, et al.

Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat Protoc. 2016 Mar 1;11(3):499–524.

35. Zeng W, Jiang S, Kong X, El-Ali N, Ball AR, Ma CIH, et al. Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. Nucleic Acids Res. 2016 Dec 1;44(21).

36. Williams K, Jiang S, Kong X, Zeng W, Nguyen NV, Ma X, et al. Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. PLoS Genet. 2020;16(5):1–26.

37. Kim M, Franke V, Brandt B, Lowenstein ED, Schöwel V, Spuler S, et al. Single-nucleus transcriptomics reveals functional compartmentalization in syncytial skeletal muscle cells. Nat Commun. 2020;11(1):1–14.

38. Petrany MJ, Swoboda CO, Sun C, Chetal K, Chen X, Weirauch MT, et al. Single-nucleus RNA-seq identifies transcriptional heterogeneity in multinucleated skeletal myofibers. Nat Commun. 2020;11(1).

39. Dos Santos M, Backer S, Saintpierre B, Izac B, Andrieu M, Letourneur F, et al. Single-nucleus RNA-seq and FISH identify coordinated transcriptional activity in mammalian myofibers. Nat Commun. 2020;11(1).

40. Orchard P, Manickam N, Varshney A, Rai V, Kaplan J, Lalancette C, et al. Human and rat skeletal muscle single-nuclei multi-omic integrative analyses nominate causal cell types, regulatory elements, and SNPs for complex traits. bioRxiv. 2020;21(1):1–9.

41. Morgan JE, Prola A, Mariot V, Pini V, Meng J, Hourde C, et al. Necroptosis mediates myofibre death in dystrophin-deficient mice. Nat Commun. 2018 Dec 1;9(1):1–10.

42. Wang LH, Friedman SD, Shaw D, Snider L, Wong CJ, Budech CB, et al. MRI-informed muscle biopsies correlate MRI with pathology and DUX4 target gene expression in FSHD. Hum Mol Genet. 2019 Feb 1;28(3):476–86.

43. Phatak J, Lu H, Wang L, Zong H, May C, Rouault M, et al. The RNAscope ® multiplex in situ hybridization technology enables the incorporation of spatial mapping and confirmation of gene signatures into single cell RNA sequencing workflows. ACD Bio; 2019. p. 1–10.

44. Marx V. Method of the Year 2020: spatially resolved transcriptomics. Nat Methods. 2021;18(1):1.

45. Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019 Apr 11;568(7751):235–9.

46. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science (80-). 2015 Apr 24;348(6233):aaa6090–aaa6090.

47. He P, Williams BA, Trout D, Marinov GK, Amrhein H, Berghella L, et al. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. Vol. 583, Nature. Springer US; 2020. 760–767 p.

48. De Micheli AJ, Laurilliard EJ, Heinke CL, Ravichandran H, Fraczek P, Soueid-Baumgarten S, et al. Single-Cell Analysis of the Muscle Stem Cell Hierarchy Identifies Heterotypic Communication Signals Involved in Skeletal Muscle Regeneration. Cell Rep. 2020;30(10):3583-3595.e5.

49. De Micheli AJ, Spector JA, Elemento O, Cosgrove BD. A reference single-cell transcriptomic atlas of human skeletal muscle tissue reveals bifurcated muscle stem cell populations. Skelet Muscle. 2020;1–13.

50. De Morrée A, Van Velthoven CTJ, Gan Q, Salvi JS, Klein JDD, Akimenko I, et al. Staufen1 inhibits MyoD translation to actively maintain muscle stem cell quiescence. Proc Natl Acad Sci U S A. 2017;114(43):E8996–9005.
51. Kann AP, Krauss RS. Multiplexed RNAscope and immunofluorescence on whole-mount skeletal myofibers and their associated stem cells. Dev. 2019 Oct 15;146(20).
52. Barruet E, Garcia SM, Striedinger K, Wu J, Lee S, Byrnes L, et al. Functionally heterogeneous human satellite cells identified by single cell RNA sequencing. Elife. 2020 Apr 1;9.
53. Naranjo JD, Dziki JL, Badylak SF. Regenerative Medicine Approaches for Age-Related Muscle Loss and Sarcopenia: A Mini-Review. Gerontology. 2017;63(6):580–9.
54. Carlson ME, Suetta C, Conboy MJ, Aagaard P, Mackey A, Kjaer M, et al. Molecular aging and rejuvenation of human muscle stem cells. EMBO Mol Med. 2009;1(8–9):381–91.
55. Hernando-Herraez I, Evano B, Stubbs T, Commere PH, Jan Bonder M, Clark S, et al. Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. Nat Commun. 2019;10(1):1–11.
56. 10X Genomics. What is the range of compatible cell sizes? [Internet]. 2018 [cited 2020 Aug 27]. Available from: https://kb.10xgenomics.com/hc/en-us/articles/218170543-What-is-the-range-of-compatible-cell-sizes-
57. Fluidigm. The Single-Cell Preparation Guide. 2014.
58. Illumina Inc., Bio-Rad Laboratories Inc. The Illumina Bio-Rad Single-Cell Sequencing Solution. 2016.
59. van den Heuvel A, Mahfouz A, Kloet SL, Balog J, van Engelen BGM, Tawil R, et al. Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. Hum Mol Genet. 2018 Nov 16;
60. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381–6.
61. Rebboah E, Reese F, Williams K, Gutierrez GB-, McGill C, Trout D, et al. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. bioRxiv. 2021;
62. Bentzinger CF, Wang YX, Dumont NA, Rudnicki MA. Cellular dynamics in the muscle satellite cell niche. EMBO Rep. 2013;14(12):1062–72.
63. Blackburn DM, Lazure F, Corchado AH, Perkins TJ, Najafabadi HS, Soleimani VD. High-resolution genome-wide expression analysis of single myofibers using smart-seq. J Biol Chem. 2019;294(52):20097–108.
64. Kröger S, Watkins B. Muscle spindle function in healthy and diseased muscle. Skelet Muscle. 2021;11(1):1–13.
65. Lemmers RJLF, van der Vliet PJ, Klooster R, Sacconi S, Camaño P, Dauwerse JG, et al. A unifying genetic model for facioscapulohumeral muscular dystrophy. Science. 2010 Sep 24;329(5999):1650–3.
66. Vanderplanck C, Ansseau E, Charron S, Stricwant N, Tassin A, Laoudj-Chenivesse D, et al. The FSHD Atrophic Myotube Phenotype Is Caused by DUX4 Expression. Chadwick BP, editor. PLoS One. 2011 Oct 28;6(10):e26820.
67. Young JM, Whiddon JL, Yao Z, Kasinathan B, Snider L, Geng LN, et al. DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. PLoS

Genet. 2013 Nov;9(11).
68. Geng LN, Yao Z, Snider L, Fong AP, Cech JN, Young JM, et al. DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy. Dev Cell. 2012 Jan 17;22(1):38–51.
69. Knopp P, Krom YD, Banerji CRS, Panamarova M, Moyle LA, den Hamer B, et al. DUX4 induces a transcriptome more characteristic of a less-differentiated cell state and inhibits myogenesis. J Cell Sci. 2016 Oct 15;129(20):3816–31.
70. Yao Z, Snider L, Balog J, Lemmers RJLF, Van Der Maarel SM, Tawil R, et al. DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. Hum Mol Genet. 2014 Oct 15;23(20):5342–52.
71. Rickard AM, Petek LM, Miller DG. Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. Hum Mol Genet. 2015 Jun 5;24(20):5901–14.
72. Tassin A, Laoudj-Chenivesse D, Vanderplanck C, Barro M, Charron S, Ansseau E, et al. DUX4 expression in FSHD muscle cells: How could such a rare protein cause a myopathy? J Cell Mol Med. 2013 Jan;17(1):76–89.
73. Himeda CL, Jones TI, Jones PL. Facioscapulohumeral muscular dystrophy as a model for epigenetic regulation and disease. Antioxid Redox Signal. 2015 Jun 1;22(16):1463–82.
74. Feng Q, Snider L, Jagannathan S, Tawil R, van der Maarel SM, Tapscott SJ, et al. A feedback loop between nonsense-mediated decay and the retrogene DUX4 in facioscapulohumeral muscular dystrophy. Elife. 2015 Jan 7;2015(4).
75. Chau J, Kong X, Viet Nguyen N, Williams K, Ball M, Tawil R, et al. Relationship of DUX4 and target gene expression in FSHD myocytes. Hum Mutat. 2021;
76. Wilson K, Faelan C, Patterson-Kane JC, Rudmann DG, Moore SA, Frank D, et al. Duchenne and Becker Muscular Dystrophies: A Review of Animal Models, Clinical End Points, and Biomarker Quantification. Vol. 45, Toxicologic Pathology. SAGE Publications Inc.; 2017. p. 961–76.
77. Gao QQ, McNally EM. The dystrophin complex: Structure, function, and implications for therapy. Compr Physiol. 2015 Jul 1;5(3):1223–39.
78. McGreevy JW, Hakim CH, McIntosh MA, Duan D. Animal models of Duchenne muscular dystrophy: From basic mechanisms to gene therapy. Vol. 8, DMM Disease Models and Mechanisms. Company of Biologists Ltd; 2015. p. 195–213.
79. Sreenivasan K, Ianni A, Künne C, Strilic B, Günther S, Perdiguero E, et al. Attenuated Epigenetic Suppression of Muscle Stem Cell Necroptosis Is Required for Efficient Regeneration of Dystrophic Muscles. Cell Rep. 2020;31(7).
80. Chang NC, Chevalier FP, Rudnicki MA. Satellite Cells in Muscular Dystrophy - Lost in Polarity. Trends Mol Med. 2016;22(6):479–96.
81. Bansal D, Miyake K, Vogel SS, Groh S, Chen CC, Williamson R, et al. Defective membrane repair in dysferlin-deficient muscular dystrophy. Nature. 2003;423(6936):168–72.
82. Kharraz Y, Guerra J, Mann CJ, Serrano AL, Muñoz-Cánoves P. Macrophage plasticity and the role of inflammation in skeletal muscle repair. Mediators Inflamm. 2013;2013.
83. Valdivia M, Vega-Macaya F, Olguín P. Mechanical control of myotendinous junction formation and tendon differentiation during development. Vol. 5, Frontiers in Cell and Developmental Biology. Frontiers Media S.A.; 2017. p. 26.
84. Subramanian A, Schilling TF. Tendon development and musculoskeletal assembly:

emerging roles for the extracellular matrix. Development. 2015 Dec 15;142(24):4191–204.

85. Dastur DK, Razzak ZA. Possible neurogenic factor in muscular dystrophy: Its similarity to denervation atrophy. J Neurol Neurosurg Psychiatry. 1973;36(3):399–410.

86. Swoboda KJ, Prior TW, Scott CB, McNaught TP, Wride MC, Reyna SP, et al. Natural history of denervation in SMA: Relation to age, SMN2 copy number, and function. Ann Neurol. 2005 May;57(5):704–12.

87. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat Biotechnol. 2019 Aug 1;37(8):916–24.

88. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat Genet. 2019 Jun 1;51(6):1060–6.

89. Linker SM, Urban L, Clark SJ, Chhatriwala M, Amatya S, McCarthy DJ, et al. Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. Genome Biol. 2019 Feb 11;20(1):30.

90. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014 Jul 20;11(8):817–20.

91. Clark SJ, Lee HJ, Smallwood SA, Kelsey G, Reik W. Single-cell epigenomics: Powerful new methods for understanding gene regulation and cell identity. Genome Biol. 2016 Apr 18;17(1).

92. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. Nat Commun. 2018 Dec 1;9(1).

CHAPTER 2

Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei

Notes: (1) This chapter is adapted from the following publication.
*Williams, K., *Jiang, S., Kong, X., Zeng, W., Nguyen, N., Ma, X., Tawil, R., Yokomori, K., Mortazavi, A. (2020). Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. PLoS Genetics. DOI: 10.1371/journal.pgen.1008754
(2) Dr. Shan Jiang contributed to the conceptualization, data curation, analysis and writing of the original draft.
(3) Dr. Xiangduo Kong contributed to the investigation, validation and visualization of this work
(4) Dr. Weihua Zeng contributed to the conceptualization and review of this work.
(5) Nam Viet Nguyen contributed to the validation of this work.
(6) Xinyi Ma contributed to the sequencing in this work.
(7) Dr. Rabi Tawil provided the resources for this work.
(8) Professor Kyoko Yokomori and Professor Ali Mortazavi contributed to the conceptualization, resources, supervision, writing and review of this work.

Chapter 2

Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei

2.1 Abstract

FSHD is characterized by the misexpression of *DUX4* in skeletal muscle. Although *DUX4* upregulation is thought to be the pathogenic cause of FSHD, *DUX4* is lowly expressed in patient samples, and analysis of the consequences of *DUX4* expression has largely relied on artificial overexpression. To better understand the native expression profile of *DUX4* and its targets, we performed bulk RNA-seq on a 6-day differentiation time-course in primary FSHD2 patient myoblasts. We identify a set of 54 genes upregulated in FSHD2 cells, termed FSHD-induced genes. Using single-cell and single-nucleus RNA-seq on myoblasts and differentiated myotubes, respectively, we captured, for the first time, *DUX4* expressed at the single-nucleus level in a native state. We identified two populations of FSHD myotube nuclei based on low or high enrichment of *DUX4* and FSHD-induced genes ("FSHD-Lo" and "FSHD Hi", respectively). FSHD-Hi myotube nuclei coexpress multiple DUX4 target genes including *DUXA, LEUTX* and *ZSCAN4*, and also upregulate cell cycle-related genes with significant enrichment of E2F target genes and p53 signaling activation. We found more FSHD-Hi nuclei than *DUX4*-positive nuclei and confirmed with *in situ* RNA/protein detection that *DUX4* transcribed in only one or two nuclei is sufficient for DUX4 protein to activate target genes across multiple nuclei within the same myotube. *DUXA* (the *DUX4* paralog) is more widely expressed than *DUX4*, and depletion of *DUXA* suppressed the expression of *LEUTX* and *ZSCAN4* in late, but not early, differentiation. The results suggest that the DUXA can take over the role of DUX4 to maintain target gene expression. These results provide a possible explanation as to why it is easier to

detect DUX4 target genes than *DUX4* itself in patient cells and raise the possibility of a self-sustaining network of gene dysregulation triggered by the limited *DUX4* expression.

2.2 Introduction

Facioscapulohumeral muscular dystrophy (FSHD) is one of the most common inherited muscular dystrophies and is characterized by progressive wasting of facial, shoulder and upper arm musculature [1]. The most common form of FSHD, FSHD1 (>95% of cases), is linked to the mono-allelic contraction of the D4Z4 macrosatellite repeat array on chromosome 4q from 11-100 units to 1-10 units, with each 3.3 kb repeat containing the open reading frame for the double-homeobox transcription factor DUX4 [2–4]. In contrast, FSHD2 (<5% of FSHD cases) has no contraction of the chromosome 4q repeat array. Approximately 80% of FSHD2 cases are characterized by recurring mutations in the chromatin modifier SMCHD1 (Structural Maintenance of Chromosomes flexible Hinge Domain-containing protein 1) on chromosome 18 [5]. SMCHD1 is important for maintenance of DNA methylation and epigenetic silencing of multiple genomic loci, including the D4Z4 repeat array [5]. Studies have also found that SMCHD1 mutations can act as disease modifiers in severe cases of FSHD1 [6, 7].

FSHD is associated with the expression of the full-length *DUX4* transcript (*DUX4fl*) which is stabilized by a specific single-nucleotide polymorphism in the chromosomal region distal to the last D4Z4 repeat creating a canonical polyadenylation signal [8–10]. *DUX4fl* encodes a transcriptional activator with a double-homeobox domain that binds to a specific sequence motif upstream of its target genes in the genome [3, 4]. Normal expression of *DUX4* is restricted to brief expression in 4-cell human embryos when it activates genes for zygote genome activation (ZGA), and in the testis [11–13]. In muscle cells, overexpression of *DUX4fl* causes differentiation defects and cytotoxicity in human and mouse myoblasts [14, 15]. However, the endogenous *DUX4fl* is expressed at extremely low levels in FSHD and DUX4 protein is only detected in 0.1% and 0.5% of patient myoblasts and myotubes, respectively, *in vitro* [16]. The relationship of DUX4-positive and -negative cells and whether DUX4-negative patient cells contribute to the disease is unclear. The regulation of *DUX4* expression is controlled by multiple epigenetic processes. D4Z4 repeats are normally heterochromatic with DNA hypermethylation and histone H3 lysine 9 trimethylation (H3K9me3), which are significantly reduced in FSHD1 and FSHD2 [17, 18]. The depletion of SMCHD1, which binds to D4Z4 repeats in an H3K9me3-dependent fashion [2], results in *DUX4fl* upregulation and mutations throughout the gene correlate with CpG hypomethylation in D4Z4 repeats [19].

Here we focused on the *SMCHD1*-mutated FSHD2 subtype in order to characterize the heterogeneity of *DUX4* and FSHD-induced target gene expression at the single-cell level using *in vitro* differentiation of primary FSHD2 patient-derived myoblasts into myotubes. Although FSHD2 represents a minor population of FSHD cases, patient cells exhibit comparable clinical and gene expression phenotype as FSHD1 [20]. We used two FSHD2 patient samples with defined genetic mutations of *SMCHD1* and significant DNA hypomethylation of D4Z4 (Table S2.1). Using bulk RNA-seq, we profiled gene expression patterns during a differentiation time-course and identified candidate disease-related key genes (i.e. FSHD-induced genes) that are upregulated specifically in FSHD cells by comparing expression profiles between FSHD2 and control. We then used single-cell RNA-seq in myoblasts and single-nucleus RNA-seq [21] in day 3 and day 5 post-differentiation myotubes to characterize the expression patterns of *DUX4* and other FSHD-induced genes. We successfully detected the first set of single nuclei with endogenous *DUX4* expression (*DUX4*-detected) from FSHD myotubes. We found that *DUX4* transcript-positive nuclei do not necessarily co-express all the FSHD-induced genes whereas a much larger set of FSHD myotube nuclei express multiple FSHD-induced genes. We performed cluster analyses and identified multiple subpopulation of FSHD nuclei with distinct gene expression signatures. In particular, we found that FSHD nuclei can be subcategorized into two populations based on high or low FSHD-induced gene expression levels (termed FSHD-Hi and FSHD-Lo, respectively). Further analyses of these two populations revealed expression of distinct sets of transcription factors related to cell cycle regulation in the FSHD-Hi nuclei, indicating their distinct cellular states. Interestingly, we found that the DUX4 target and paralog, DUXA, is widely expressed and maintains other DUX4 target gene expression, which may provide insight into how rare expression of *DUX4* results in a wide-spread dystrophic phenotype.

2.3 Results

2.3.1 Upregulation of FSHD-induced genes during FSHD2 myotube differentiation

Previous studies indicated that *DUX4* is upregulated during FSHD patient myoblast differentiation [22]. In order to understand the temporal expression differences between FSHD2 patient-derived and control myoblasts, we differentiated these *in vitro* to measure the dynamics of gene expression in a 6-day time-course using conventional bulk RNA sequencing (RNA-seq) (Figure 2.1A and S2.1) (Methods). We used two independent primary control myoblast samples from tibialis anterior, Control-1 and Control-2, and two from quadricep, Control-3 and Control-4, and two independent primary FSHD2 myoblast samples from tibialis anterior, FSHD2-1 and FSHD2-2, which have known *SMCHD1* mutations (Table S2.1). After sequencing two biological replicate RNA samples for each of the six cell lines every day for six days, we filtered out lowly expressed genes and kept 10,827 genes for downstream analysis. We do not detect *DUX4* from the RNA-seq probably due to few nuclei expressing *DUX4*, but we detect the induction of *DUX4-fl* via RT-qPCR (S2.2 Figure). We looked for differences between the control and FSHD2 myoblasts from the tibialis anterior using principal component analysis (PCA) (S2.3 Figure) and for all the samples (S2.4 Figure). We observed that the days of differentiation aligned to each other across cell lines following a clear trajectory of myogenesis (PC1, 51.9% variance in expression; PC2, 13.2% variance in expression). We also found that the two FSHD2 cell lines diverge from the two tibialis anterior control cell lines for days 3 to 5 in two principal components with known genes upregulated in FSHD driving the variance (PC3, 5.9% variance in expression; PC4, 4.0% variance in expression) (S2.1 Figure). Thus, FSHD2 patient-derived myotubes can be distinguished from control cells by day 3 of differentiation when profiling transcriptomes at the population level.

In order to identify temporal patterns of expression, we used maSigPro [23] to cluster genes into three clusters based on expression over time (Figure 2.1B and S2.5) (Methods). A set of 54 genes are specifically upregulated in FSHD2 starting at day 2 (Cluster 3) (Figure 2.1B and 2.1C). We define these 54 upregulated genes along with *DUX4* as "FSHD-induced genes" (Figure 2.1B and 2.1C). Genes in this cluster were highly enriched in GO terms for negative regulation of cell differentiation (p=1x10$^{-12.9}$) and methylation-dependent chromatin silencing (p=1x10$^{-7.17}$) (Table S2.2). Of these 54 genes, 53 were previously identified as possible DUX4 targets from myoblasts with inducible DUX4 [24], endogenous DUX4 [22] or FSHD biopsies [20] (S2.6 Figure). While these genes overlap with those upregulated in response to *DUX4* expression, they may not be direct DUX4 target genes since DUX4 turns on other transcriptional regulators. For this reason we refer to these as "FSHD-induced genes". These genes were upregulated in waves starting at day 2, such as *LEUTX* and *ZSCAN4*, followed by day 3, such as

CCNA1 and *DUXA*, and day 4, such as *DUXB* (Figure 2.1C and S2.7). After being significantly upregulated, most FSHD-induced genes remained upregulated through the end of the time-course, including two DUX4 paralogs, *DUXA* and *DUXB* (Figure 2.1C and S2.7) [25].

The other two clusters of genes identified from maSigPro represent genes increasing (Cluster 2) or decreasing (Cluster 1) in expression in both FSHD2 and control across the timecourse (S2.5 Figure, Table S2.2). GO terms for these clusters include muscle system process ($p=1x10^{-67.0}$) and muscle structure development ($p=1x10^{-47.1}$) for cluster 2, and RNA splicing ($p=1x10^{-11.5}$) for cluster 1 (Table S2.2). Myogenesis genes, such as ACTA1 and MYOG, are in cluster 2. Both FSHD2 and control samples have similar expression levels in both these clusters across time (S2.5A and S2.5B Figure), suggesting that the control and FSHD2 samples seem to differentiate at similar efficiencies. We also monitored the differentiation of Control-2 and FSHD2-2 by differentiation index and MYH1 staining (S2.5C Figure). The differentiation index of FSHD2-2 is statistically lower than that of Control-2 at day 3, but the two are not statistically different by day 5. Altered myogenesis in FSHD cells has been shown in previous studies [26]. Recently, a study showed upregulation and incorporation of alternate histones H3.X and H3.Y following DUX4 expression [27]. In this study, *H3.Y* (AKA *RP11-432M8.17*) has increased expression in FSHD2 cells and is included in our FSHD-induced genes. *H3.X (RP11-321E2.13)* is classified as a pseudogene in the reference we use and was therefore not included in our analysis. In summary, we found a set of genes significantly upregulated in differentiating FSHD2 myotubes by day 3 which we term FSHD-induced genes along with *DUX4*.

2.3.2 Detection of nuclei with *DUX4* expression from FSHD2 myotubes using single-nucleus full-length RNA-seq

Although we failed to detect *DUX4* in our bulk RNA-seq, the upregulation of FSHD-induced genes was nevertheless observed during myotube differentiation specifically in FSHD2 samples. We wondered whether the expression of FSHD-induced genes is seen in every cell and whether the expression of *DUX4* and DUX4-target genes were indeed present only in a subset of cells. We therefore performed single-cell RNA-seq on undifferentiated myoblasts and single-nucleus RNA-seq on myotubes using the Smart-Seq protocol on the Fluidigm C1 platform [21] at day 3 of differentiation using control and FSHD2 primary cells (S2.8A Figure). Day 3 was chosen as it was the first day of robust FSHD-induced gene expression in the differentiation time-course thereby allowing us to observe early transcriptional changes. Additionally, we selected FSHD2-2 based on the higher expression level of FSHD-induced genes compared to FSHD2-1 during differentiation (Figure 2.1B and 2.1C). The Fluidigm C1 platform enables us to prepare full-length cDNA libraries from up to 96 cells or nuclei at a time. We captured a total of 317 cells and nuclei with an average read depth of 2,624,274 per cell or nucleus and kept cells and nuclei with at least 500 genes detected (S2.8A Figure). As quality control that our single cell data matched our bulk time-course, we first pooled reads from all single cells/single nuclei for each cell type and performed incremental PCA with the bulk time-course RNA-seq samples for these cell lines (S2.8B and S2.8C Figure). As expected, the pooled single cell myoblasts clustered with day 0 samples in both control and FSHD2. For the pooled myotube single nuclei, FSHD2 replicate 1 (FSHD2 R1) aligned with day 3 of the FSHD2 time-course, but FSHD2 replicate 2 (FSHD2 R2) located between control and FSHD2 day 3 in the time-course (S2.8C Figure). This suggests variable differentiation efficiencies for the two replicates, which could be caused by subtle differences in seeding density.

Importantly, we found that 3 out of 79 (3.8%) nuclei in FSHD2 R1 showed high expression of *DUX4* (11.24 TPM, 34.15 TPM and 68.49 TPM) while we found no *DUX4*-detected nuclei in FSHD2 R2, revealing the high level of heterogeneity in the FSHD2 cell population with *DUX4* only expressed in a small fraction of nuclei. We then analyzed the global profiles of the single-cell and single-nucleus transcriptomes using PCA analysis and found that all 3 *DUX4*-detected nuclei as well as other FSHD2 R1 nuclei clearly separated from FSHD2 R2 and control myotube nuclei (Figure 2.2A). Co-clustering of both *DUX4*-positive and negative nuclei of FSHD2 R1 suggests that they might come from the same myotubes as cell fusion was not blocked during differentiation in our study. Diffusion of the DUX4 protein to multiple nuclei was demonstrated previously despite *DUX4* mRNA transcription in only a few nuclei of the same myotube [22]. We further confirm this by RNA-protein costaining of DUX4 (Figure S2.11B). We analyzed the 55 genes, which includes *DUX4* and FSHD-induced genes, genes specifically upregulated at day 3 or later during our bulk time-course of FSHD2 differentiation (Figure 2.1B and 2.1C), and observed that these genes showed significant enrichment in FSHD2 R1 myotube nuclei compared with control myotube nuclei ($p<2e-16$). Nuclei with the highest enrichment clustered with the 3 *DUX4*-detected nuclei, and thus we labeled this group of nuclei “FSHD-induced genes high” (FSHD-Hi) (Figure 2.2B and S2.9). The FSHD2 R2 myotube nuclei also showed significantly higher enrichment of FSHD-induced genes than control myotube nuclei ($p<2e-16$) but had fewer FSHD-induced genes expressed than the FSHD-Hi group, and therefore this group of nuclei was labeled “FSHD-induced genes low” (FSHD-Lo) (Figure 2.2B and S2.9). We found that all myoblast cells and control myotube nuclei rarely express more than 2 FSHD-induced genes (Figure 2.2C), whereas FSHD-Lo nuclei coexpress between 1 to 6 and at most 9 of the FSHD-induced genes. However, all FSHD-Hi nuclei express at least 6 of these genes with most coexpressing at 12 and up to 22 genes (Figure 2.2C). In summary, we detected two different patient myotube nuclei populations: (1) a set of 79 nuclei that express FSHD-induced genes (FSHD-Hi), 3 of which express endogenous *DUX4* (*DUX4*+); (2) 60 nuclei that are clearly different from control nuclei but with no *DUX4* detected and significantly lower FSHD-induced gene expression (FSHD-Lo).

Interestingly, we observed the expression of DUX4 paralogs *DUXA* and *DUXB* expressed in FSHD2 myotube nuclei. *DUXA* was expressed exclusively in the FSHD-Hi nuclei population. We found that 34 FSHD-induced genes were expressed in both FSHD-Hi and FSHD-Lo populations, including reported DUX4 targets *LEUTX*, *ZSCAN4*, *MBD3L2*, *TRIM43*, *KHDC1L* and *CCNA1* [4, 20, 25] indicating that they may perform as a core set of responsive and interactive genes during FSHD progression (Figure 2.2D). We observed that FSHD-Hi and FSHD-Lo have distinct coexpression patterns which indicates different cell states. Within the FSHD-Hi nuclei, a large number of the FSHD-induced genes are coexpressed with transcription factors, such as *LEUTX* and *DUXA*, but not *DUX4* (Figure 2.2D). Taken together, two identified patient myotube nuclei populations, FSHD-Hi with a small set of *DUX4*-detected nuclei and FSHD-Lo, exhibit distinct co-expression patterns of FSHD-induced genes including DUX4-target transcription factor genes.

To assess whether these groups of nuclei have distinct expression of FSHD-induced genes, we determined the coexpression patterns between a subset of FSHD-induced genes which had variable expression in the single cells and nuclei. To determine expression profiles of *DUX4*-detected nuclei, we examined genes coexpressed with *DUX4*. We found that *DUX4* was coexpressed with 23 FSHD-induced genes including two transcription factors, *LEUTX* and *ZSCAN4*, which have been reported as DUX4 targets in FSHD (Figure S2.6 and S2.10) [22, 24].

DUX4 and *ZSCAN4* were expressed in all three *DUX4*-detected nuclei while *DUX4* and *LEUTX* were only simultaneously expressed in one *DUX4*-detected nuclei. FSHD-induced genes coexpressed in all three *DUX4*-detected nuclei include *KHDC1L*, *PRAMEF25*, *PRAMEF9*, *RFPL4B*, *RP11-432M8.17*, *SLC34A2*, *SLC38A1* and *ZSCAN4*, while genes like *CTB-25J19.*1, *TRIM49*, *RFPL1*, *MBD3L2*, *MBD3L3* and *MBD3L5* are coexpressed with *DUX4* in two of the *DUX4*-detected nuclei. Additionally, the nucleus with *DUX4*, *LEUTX* and *ZSCAN4* also expressed *KDM4E*, *TRIM43*, *TRIM43B*, *MBD3L3*, *MBD3L5*, and *RFPL2*. Taken together, the genes expressed in the *DUX4*-detected nuclei may represent early targets of DUX4 which initiate a pathogenic gene regulatory network.

To substantiate the co-expression of *DUX4* and/or DUX4-target genes, we performed RNA FISH on *DUX4* and two representative FSHD-induced genes, *LEUTX* and *SLC34A2*, in day 3 differentiated FSHD2-2 myotubes (Figure 2.2E). Probes were designed to hybridize to the two regions unique to the *DUX4fl* transcript to ensure the specificity, and we support the specificity with staining for DUX4 protein along with *DUX4* RNA FISH (Figure S2.11A and S2.11B). Our *DUX4* probe detected the *DUX4* transcript primarily in the nucleus, possibly reflecting the de novo RNA transcription with some weak signals in the cytoplasm (Figure 2.2E, S2.11A and S2.11B). We observed that ~7% of myotubes have at least 1 *DUX4*-detected nucleus, and that *DUX4*-positive myotubes contain on average 2 *DUX4*-detected nuclei (among on average 15 nuclei per myotube), indicating that even in the permissive patient myotubes, very few nuclei actually express *DUX4*. In these myotubes, however, DUX4 protein spreads to almost all the nuclei (Figure S2.11B). In contrast to the limited expression of *DUX4* RNA, *LEUTX* and *SLC34A2* RNA transcripts are abundantly present in the cytoplasm in addition to multiple nuclei (Figure 2.2E). These results are in agreement with snRNA-seq results in which a higher number of nuclei expressing FSHD-induced genes were detected compared to the small number of *DUX4* RNA-positive nuclei (Figure 2.2A). Taken together, these results suggest that once expressed, DUX4 protein may transcribe target genes in multiple nuclei in the same myotube. Interestingly, we also found that some FSHD myotubes contain *DUX4* transcript but no *LEUTX*, whereas others contain no detectable *DUX4* transcript with abundant signals of *LEUTX* and *SLC34A2* transcripts (Figure 2.2E and S2.11C). These results raise the possibility that FSHD-induced gene expression may persist even after *DUX4* transcript is no longer detectable.

2.3.3 Single-nucleus 3' end RNA-seq on FSHD2 and control early and late myotubes

We identified two distinct populations of FSHD patient nuclei, FSHD-Hi and FSHD-Lo. Since we analyzed a limited number of nuclei using Smart-Seq, we decided to perform additional single-nucleus sequencing in a larger set of nuclei and over two time points in order to address whether the two populations simply reflect different stages of differentiation. We performed 3' end RNA-seq on two biological replicates of FSHD2-2 and two of Control-2 nuclei from day 3 and day 5 of differentiation using the Illumina SureCell WTA 3' protocol using the BioRad ddSeq Single Cell Isolator (referred to from now on as "ddSeq"), which allows us isolate thousands of nuclei at a time (Methods). We have 32,273 nuclei which pass our quality filters with an average of 14,139 reads/cell (Figure S2.12). We performed the UMAP dimensionality reduction using Seurat on 19,615 genes (Figure 2.3A). Nuclei separate across the first dimension by disease, and to a lesser extent by differentiation in the second dimension (Figure 2.3A). To distinguish subpopulations, we cluster the nuclei using shared nearest neighbors (SNN) and find 22 clusters across FSHD2 and control nuclei (Figure 2.3D). These clusters contain a mix of FSHD2 and control nuclei across differentiation (Figure 2.3G). We plot the expression of *MYH3* to check that the nuclei are originally from myotubes (Figure 2.3E). As expected, the majority of nuclei express *MYH3* and were therefore differentiated. However, clusters 15 and 7 have little or no *MYH3* detected and we presume these are either mononuclear cells that did not differentiate given the expression of *MYOD1, MYF5 and DES* (Figure S2.18) or contaminating non-myogenic cells. We see a similar pattern when looking at expression of other myogenic markers as well (Figure S2.18). FSHD2 nuclei seem to have somewhat lower expression of *MYH3* than control across both days of differentiation, which may be biologically significant as was previously noted in that FSHD cells have transcriptome profiles of less differentiated cells [28].

We detect *DUX4* in 13 FSHD2 nuclei, 3 nuclei (0.05%, 3/6152) from day 3 and 10 nuclei (0.1%, 10/9396) from day 5, and they are found spread across multiple clusters (Figure 2.3B). Higher number of *DUX4*-positive nuclei on day 5 is consistent with the previous studies reporting the increased frequency of *DUX4* expression upon differentiation [16]. Interestingly, the *DUX4*+ nuclei do not cluster with nuclei expressing the highest number of FSHD-induced genes (Figure 2.3B and 2.3C). We find a much larger number of nuclei that express *DUXA* and some that express *DUXB*, and these nuclei cluster with nuclei expressing high number of FSHD-induced genes (Figure 2.3B). Except for one nucleus coexpressing *DUXA* and *DUXB*, the three DUX genes are never coexpressed (Figure 2.3B).

To identify similar FSHD-Hi and FSHD-Lo populations as found in the full-length RNA-seq data from the Fluidigm C1, we mapped the number of FSHD-induced genes detected per nuclei. Nuclei with 2-5 FSHD-induced genes coexpressed are spread across both day 3 and day 5 FSHD2 myotube nuclei (Figure 2.3C). Cluster 16 and neighboring clusters have the highest proportion of nuclei with more than 6 FSHD-induced genes detected (Figure 2.3C). *ZSCAN4* expression follows a similar pattern to that of the number of FSHD-induced genes detected, with its highest expression in cluster 16 (Figure 2.3F). We found *ZSCAN4* to be significantly upregulated starting at day 2 of differentiation in our bulk RNA-seq time-course and therefore its wide spread expression is not surprising. The expression patterns of *ZSCAN4*, particularly in the day 3 FSHD2 nuclei, and the other FSHD-induced genes shows the heterogeneity in the activation of FSHD-induced genes across different nuclei, especially as the day 5 nuclei express *ZSCAN4* more robustly (Figure 2.3F and S2.19). Looking at the average gene expression of all the nuclei for each ddSeq cluster, clusters 16, 17, 1, 4 and 11 have the highest expression of the FSHD-induced genes and are made up primarily of FSHD2 nuclei (Figure 2.3G and 2.3H). These ddSeq clusters are akin to the FSHD-Hi cluster from Smart-Seq, and we refer to the FSHD2 nuclei in them collectively as FSHD-Hi (Figure 2.3H). ddSeq clusters 13, 2, 9, 6, 5 and 18 have moderate expression of the FSHD-induced genes, and cluster separately from the FSHD-Hi clusters (Figure 2.3H). They also have a large proportion of FSHD2 nuclei and nuclei with 2-5 FSHD-induced genes coexpressed (Figure 2.3G, 2.3C). These ddSeq clusters are similar to the Smart-Seq FSHD-Lo group identified from the Fluidigm nuclei, and we therefore label the FSHD2 nuclei in them FSHD-Lo (Figure 2.3H). Thus, using ddSeq with a larger population of nuclei, we confirmed the presence of two different states of FSHD nuclei "FSHD-Hi and FSHD-Lo". Importantly, our FSHD-Hi and FSHD-Lo groups includes mixes of both day 3 and day 5 myotube nuclei, suggesting that the differences are not simply attributable to differentiation status (Figure 2.3G).

2.3.4 Day 3 FSHD2 myotube nuclei expression patterns are similar across full-length RNA-seq and 3' end RNA-seq

To make sure that the nuclei from the two sequencing technologies, Smart-Seq and ddSeq, are comparable, we plotted them together on one UMAP (Figure 2.4A). The nuclei from both technologies overlap, and FSHD-Hi and FSHD-Lo nuclei still separate (Figure 2.4A). The six *DUX4*+ nuclei from these day 3 FSHD2 samples do not cluster together, nor do they cluster with nuclei with high numbers of FSHD-induced gene detected (Figure 2.4B, S2.13). In this set, no nuclei coexpress *DUX4*, *DUXA* or *DUXB*, perhaps because *DUXB* is expressed later in differentiation as is seen in the bulk timecourse (Figure 2.1C, S2.7). To see if nuclei separate by expression of FSHD-induced genes, we plot the number of FSHD-induced genes and find that nuclei expressing six or more FSHD-induced genes separate to one side of the UMAP, but do not form a distinct cluster (Figure 2.4C, S2.13). Nuclei from the different technologies mix regardless of the number of FSHD-induced genes detected. Given that the nuclei do not separate based on technology, we continue with comparative analysis with the ddSeq data only.

A recent single-cell RNA-seq study also identified a small population of *DUX4* transcript-positive cells in both FSHD1 and FSHD2 patient-derived primary myocytes [29]. In that study, however, myoblast differentiation was induced but myotube fusion was artificially blocked by the use of a calcium chelator [29]. This is in contrast to our study, in which we examined nuclei from unperturbed myotubes using snRNA-seq. Importantly, our approach enables us to uniquely address how *DUX4* expression, even in a single nucleus, results in target gene activation in other nuclei in the same myotube (due to the DUX4 protein spreading) under native condition to distinguish the FSHD-Hi and FSHD-Lo population of cells. We analyzed the expression of 67 DUX4 target genes used in Heuvel, et al. [20, 27] in our FSHD-Hi and FSHD-Lo myotube single nucleus populations. For the Smart-Seq nuclei, all FSHD-Hi nuclei and about 3.3% of FSHD-Lo nuclei highly express at least 5 of these genes (Figure S2.14). For the ddSeq nuclei, 5.2% of FSHD-Hi nuclei and 1% of FSHD-Lo nuclei express at least 5 of these genes (Figure S2.14). Interestingly, even 1.5% of our ddSeq FSHD2 nuclei excluded from the High and Low populations based on apparent differentiation status express at least 5 of those genes. These percentages are much higher than that in single cell myocyte data (0.2-0.9%) (Figure S2.14) [29]. As we confirmed by RNA FISH with DUX4 protein co-staining (Figure 2.2E and S2.11), higher percentages of nuclei expressing more target genes in our study is due to DUX4 protein spreading and target gene activation in multiple nuclei in native myotubes, which is blocked in single nucleus myocytes [29].

We identified that 0.05% of our ddSeq day 3 and 0.1% of our day 5 myotube nuclei express *DUX4*, which is consistent with frequencies observed in other studies (Figure S2.14A) [16]. In our Smart-Seq data, 2.12% of the day 3 nuclei express *DUX4* at high levels, which is higher than the percentage reported in single cell myocytes (0.2-0.9%) [29] (Figure S2.14A). Currently unclear is whether blocking myotube fusion interferes with the normal course of myotube differentiation and affects frequency of *DUX4* expression. Taken together, our snRNA-seq analysis captured the extent of target gene expression by the limited expression of *DUX4* in patient myotubes. Our higher-sensitivity Smart-Seq data allowed us to identify the FSHD-Hi and FSHD-Lo populations, and our more robust number of nuclei from the ddSeq data enables us to distinguish the differences between these two populations, possibly representing two different states of patient myotube nuclei.

2.3.5 FSHD-Hi myotube nuclei turn on cell cycle programs

To identify genes marking the Low and FSHD-Hi populations, we performed differential expression analysis on 19,615 genes for 6,210 FSHD-Lo nuclei and 8,135 FSHD-Hi nuclei

(Figure 2.5A). We found 1,557 genes significantly more highly expressed in FSHD-Hi and 111 genes more highly expressed in FSHD-Lo (t-test, Benjamini-Hochberg FDR < 0.05) (Figure 2.5B). Of the 54 FSHD-induced genes, 42 were more highly expressed in FSHD-Hi. SMCHD1 has been shown to regulate the *PCDH* gene clusters, and we find four *PCDH* genes differentially expressed; *PCDH10* and *PCDHGA6* were higher in FSHD-Hi, while *PCDHGB4* and *PCDHGB5* were higher in FSHD-Lo. We also find 149 transcription factors (10% of the FSHD-Hi genes) in FSHD-Hi including 87 zinc fingers and 16 homeobox genes, many of which are important in embryogenesis including several *HOX* genes (Table S2.3). We also see 84 cofactors (5% of the FSHD-Hi genes) upregulated including six cyclin genes; *CCNA1*, *CCNA2*, *CCNE1*, *CDK1*, *CDK2*, *CDKN1C* (Table S2.3). In contrast, the FSHD-Lo group has 2 transcription factors (2% of the FSHD-Lo genes) and 4 cofactors (4% of the FSHD-Lo genes) upregulated, including *NOTCH3* and *TGFB1* (Table S2.3). The myogenic regulator, *MYOD1*, whose expression decreases during myogenesis, is more highly expressed in the FSHD-Hi group, while *ACTA1*, whose expression increases during myogenesis, is higher in FSHD-Lo. This suggests that although FSHD-Lo has a higher percentage of day 3 FSHD2 myotube nuclei, the FSHD-Hi group has expression of key genes indicative of a less differentiated transcriptomic state.

Additionally, the genes more highly expressed in FSHD-Hi have gene ontology (GO) terms related to cell division and replication (Figure 2.5E). Included in these categories are many chromatin remodelers and transcription factors involved during the cell cycle. As these myotubes are no longer cycling, these cell cycle related gene products could be altering the genomic landscape in lieu of DNA replication. Additionally, FSHD cells have been shown to have transcriptomes of less differentiated cell states [28]. Activation of these cell cycle genes in the G0 myotubes could be responsible for the less differentiated transcriptomes of FSHD cell. The

GO term “signal transduction by p53 mediator” is also enriched in FSHD-Hi (Figure 2.5E), and previous studies have shown that DUX4 requires p53 to cause cytotoxicity [30]. These FSHD-Hi nuclei could be activating the p53 pathway and therefore be the disease-driving nuclei in FSHD. GO terms enriched in the FSHD-Lo group include those related to extracellular structures which has been shown previously to be downregulated in *DUX4* expressing cells (Figure S2.15) [22].

To identify regulators key to the genes upregulated in the FSHD-Hi population, we looked for enrichment of transcription factors and other DNA binding proteins that bind these genes based off of ChIP-seq data from two genomic databases, ENCODE and ChEA (Figure 2.5C). Five transcription factors, E2F1, E2F4, FOXM1, NFYA and NFYB, and one corepressor, SIN3A, are statistically enriched for regulating the FSHD-Hi genes. All of these are involved in cell cycle gene regulation, which is consistent with the GO terms identified for these genes. *FOXM1* and *E2F1* are both upregulated in FSHD-Hi nuclei as well (Figure 2.5B). The target genes for five of these transcriptional regulators, all but E2F1, show a significant difference in expression between FSHD-Hi and FSHD-Lo (Figure 2.5D). E2F4 represses genes which are upregulated by E2F1 during the G1 to S phase transition, which may explain why we see E2F4 target genes as significantly different between the two groups but not E2F1 [31]. Additionally, we do not detect this upregulation of cell cycle genes other than *CCNA1* in the bulk RNA-seq time-course (Figure 2.1B and S2.5), which emphasizes that this upregulation is specific to these FSHD-Hi nuclei.

2.3.6 DUXA regulates FSHD-induced genes

Given that *DUX4* expressing nuclei did not cluster with the nuclei expressing the highest amount of FSHD-induced genes (Figure 2.3B and 2.3C), we searched for other widespread transcriptional regulators that could be regulating the FSHD-induced genes in a wider set of nuclei. A DUX4 target gene, *DUXA*, is highly upregulated in FSHD2 and detected in a large number of nuclei and we therefore looked for binding sites of these two transcription factors, DUX4 and DUXA, in the promoter regions (-1.5 kb to +0.5 kb) of *DUX4*, *DUXA*, *ZSCAN4* and *LEUTX* to see if they could be regulating themselves and other FSHD-induced genes. A ChIP-seq binding motif is available for DUX4, and an HT-SELEX motif is available for DUXA (Figure S2.16A and S2.16B) [32]. Not surprisingly, DUX4 has one binding site in each of the FSHD-induced genes we looked at, two for *LEUTX*, and one for itself. DUXA has one binding site for itself, one in the promoter of *LEUTX*, and two for *ZSCAN4*. The DUX4 and DUXA binding sites overlap in the *DUXA* promoter, and for one of the sites for *LEUTX* and one for *ZSCAN4* (Figure S2.16C). Since the binding sites overlap, DUXA, once expressed, may regulate these DUX4 target genes after DUX4 is no longer present.

To further analyze the relationship between *DUX4*, *DUXA* and other FSHD-induced genes, we look at the coexpression of DUX4 and DUXA with the FSHD-induced genes in the FSHD-Hi and FSHD-Lo populations (Figure 2.6A and 2.6B). In the FSHD-Lo, we see *DUX4* coexpressed with *CCNA1* only (Figure 2.6A). However, *DUXA* is coexpressed with a 26 FSHD-induced genes in the low population. In the FSHD-Hi population, we see *DUX4* coexpressed with 10 FSHD-induced genes, while *DUXA* is coexpressed with 41 FSHD-induced genes (Figure 2.6B). Assuming that the nuclei in which we detect *DUX4* are the first to express *DUX4* in their respective myotubes, these ten genes coexpressed with *DUX4* may be its early targets. However, we cannot rule out that these differences could be attributable to the detection sensitivity of the technology and to the difference between the number of *DUX4* and *DUXA* expressing nuclei detected. *ZSCAN4* is coexpressed with *DUX4* and to a larger extent with *DUXA* (Figure 2.6B).

LEUTX appears to be coexpresssed primarily with *DUXA* in both the FSHD-Hi and FSHD-Lo populations (Figure 2.6A and 2.6B). As described earlier, we observed that some myotubes express *LEUTX* with apparent lack of *DUX4* transcript (Figure 2.2E and S2.11C). However, this may be due to persistent DUX4 protein. Thus, we performed DUX4 protein immunostaining combined with *LEUTX* RNA FISH (Figure 2.6G and S2.11B). While DUX4 protein can be detected in multiple nuclei within the same myotube expressing *LEUTX* (Figure S2.11B), we also found that in some myotubes the levels of DUX4 protein and *LEUTX* transcript expression are discordant (Figure 2.6G). Indeed, in some nuclei with *LEUTX* expression, no significant DUX4 protein was detected, raising the possibility that *LEUTX* may be transcribed in the absence of DUX4 protein (Figure 2.6G).

To further assess the relationship between DUX4 target transcription factor genes, *DUXA*, *LEUTX* and *ZSCAN4*, we transfected FSHD2-2 myoblasts with shRNAs against *DUXA*, *LEUTX* and *ZSCAN4* and measured their gene expression after 4 and 6 days of differentiation (Figure S2.17). Interestingly, RT-qPCR results reveal that while depletion of *DUXA* has no significant effect on *LEUTX* and *ZSCAN4* expression on day 4, it significantly suppressed their expression on day 6 (Figure 2.6C, 2.6D and 2.6E). Depletion of *LEUTX* or *ZSCAN4* did not have any significant effect on *DUXA* expression on either day 4 or day 6 (Figure 2.6C). The results demonstrate that in addition to DUX4, DUXA can regulate the expression of *LEUTX* and *ZSCAN4* (Figure 2.6F). Differential effects on days 4 and 6 strongly suggest that these genes are initially activated by DUX4, but once sufficient amount of DUXA is induced, their expression is further promoted by DUXA. Thus, DUXA may function to amplify and sustain the DUX4 signal in this way, providing a self-supporting network of gene dysregulation that can lead to pathogenesis regardless of the temporary expression of *DUX4* consistent with the long-standing observation in previous studies that FSHD-induced gene expression is easier to detect in patient muscle cells than the *DUX4* transcript itself.

2.4 Discussion

Using our time-course bulk RNA-seq analysis of control and FSHD2 patient myoblast differentiation, we defined a set of 54 genes that are specifically induced in FSHD2 as "FSHD-induced genes". Those genes largely overlap with previously defined downstream targets of DUX4 [22, 25] though we cannot rule out the possible contribution of the *SMCHD1* mutation. Single-cell and single-nucleus RNA-seq on two different platforms revealed that FSHD2 myotube nuclei express higher FSHD-induced genes than myoblasts or control myotube nuclei. Importantly, we were able to identify *DUX4* transcript-positive nuclei, which were not detected in our bulk RNA-seq. We further identified two populations of FSHD2 myotube nuclei, FSHD-Hi and FSHD-Lo, based on the expression of FSHD-induced genes. We found that FSHD-Hi nuclei also upregulate cell cycle genes and identified a set of transcriptional regulators that may contribute to this upregulation. We found evidence that DUXA affects expression of the DUX4 target genes *LEUTX* and *ZSCAN4* in later days, which raises the possibility that DUXA may be important for DUX4 signal amplification by contributing to the regulation of DUX4 target genes.

While FSHD-Hi nuclei express markers of differentiated myotubes and have a higher proportion of day 5 nuclei than FSHD-Lo, they exhibit higher expression of *MYOD1* and lower expression of *ACTA1* than FSHD-Lo. Thus, FSHD-Hi nuclei appear to have transcriptomic markers of a less differentiated state, which may be consistent with a previous observation in a mouse model of FSHD [28]. Accordingly, we found that cell cycle genes are specifically upregulated in FSHD-Hi nuclei, and five transcription factors, E2F1, E2F4, FOXM1, NFYA and

NFYB, and one corepressor, SIN3A, are statistically enriched for regulating these genes. Interestingly, some of these factors have been previously linked to FSHD-related gene regulation. SIN3A complexes with HDACs and TET proteins and appears to be involved in *DUX4* repression [33, 34]. NF-Y, made up in part by NFYA and NFYB, binds to HERV LTR repeats which are activated in FSHD [3, 35]. E2F4, E2F1 and FOXM1 are all part of the DREAM complex which regulates cell cycle genes [31]. E2F1 activates a DUX4-target gene *CCNA1*, and both E2F1 and FOXM1 are regulated by phosphorylation by CDK2 complexed with cyclin A [31, 36], which are both upregulated in FSHD-Hi nuclei. Thus, these cell cycle transcriptional regulators may contribute to FSHD-associated gene dysregulation. How these cell cycle-related genes in a subset of post-mitotic, multinucleated myotubes contribute to pathogenesis in FSHD is currently unclear.

Small populations of DUX4-positive myotubes are thought to drive pathogenesis in FSHD [2–4]. We found 0.1% (16/15,687) of nuclei expressed *DUX4* using single-nucleus RNA-seq which is lower than the reported 0.5% (1/200 in myotube nuclei) [8, 10, 16, 27] possibly due to variability in expression levels of *DUX4* between individuals. However, the percentage of DUX4-affected nuclei we found (3.7%, 583/15,687) is higher than that reported in FSHD single myocytes (0.55%, 27/4902) [29]. Our high-resolution single-cell and single-nucleus dataset is the first to observe the endogenous expression of *DUX4* in a small number of FSHD2 myotube nuclei with wider expression of target genes. Our snRNA-seq and immuno-RNA FISH results demonstrate that one or two nuclei expressing *DUX4* transcripts appears to produce sufficient DUX4 protein to spread to multiple nuclei, consistent with a previous study [16, 22], and possibly initiate target gene expression.

Previous studies suggested a feedback loop between DUX4 and its target genes to further increase DUX4 expression via (1) DUX4-mediated proteolytic degradation of UPF1 and inhibition of nonsense-mediated RNA decay resulting in stabilization of DUX4 mRNA [37], and (2) the DUX4 target MBD3Ls binding to D4Z4 and relieving DUX4 repression [34]. Additionally, alternate histones which are targets of DUX4 have been shown to continue the expression of DUX4 target genes [27]. In the current study, we provide support for DUXA as another regulator of DUX4 target genes which may amplify the DUX4 gene network. Although DUXA is a DUX4 paralog [25] and has been identified as a DUX4 target gene in human patient muscle cells [20], no study reports its specific functions in FSHD. DUXA is a transcription factor with two homeobox domains like DUX4, and it binds to a 10 bp motif similar to DUX4 [24, 29]. Importantly, our results indicate that *DUXA* upregulates *LEUTX* and *ZSCAN4* in late but not early in differentiation, suggesting a possible two-step mechanism for upregulation of these DUX4-target transcription factors; first by DUX4 then DUXA. In support of this, we found that *LEUTX* is present in nuclei with no significant DUX4 protein present. Given that *DUXA* is much more widely expressed together with FSHD-induced genes in our analysis, we propose that DUXA may drive an FSHD-specific pathogenic program by binding and activating a pool of DUX4 targets, therefore reinforcing the DUX4-induced gene network in patient myotube nuclei.

Previous studies indicated that DUX4 expression leads to p53-dependent apoptosis within 20 hours of initial expression [15, 22, 30]. We observed, however, continuous upregulation of DUX4 target genes over 6 days without any overt cytotoxic phenotype or apoptotic transcriptomic signature. This suggests that DUX4 upregulation may not be immediately toxic in the endogenous context. We hypothesize that sporadic endogenous *DUX4* expression may be relatively short-lived, and that downstream DUX4 target genes, such as DUXA, may amplify and/or reinforce the FSHD-induced gene network in addition to or in place of DUX4 eventually leading to myotoxicity and dystrophy. If this is the case, it is possible that therapeutics targeting DUX4 or *DUX4* expression may limit initiation of FSHD in new tissue but may not stop muscle wasting in already disease-activated tissue, and targeting transcription factors downstream of DUX4 may be necessary.

Our time-course bulk RNA-seq and single-cell/-nucleus RNA-seq of primary control and FSHD2 myoblasts and myotubes addressed FSHD-specific gene expression during differentiation. Single-nucleus RNA-seq demonstrated the heterogeneity and disease-specific transcriptomic changes of patient myotube nuclei with high or low expression of *DUX4* and FSHD-induced genes. Our results provide strong evidence that *DUX4* transcript expression in one or two nuclei can result in a high expression of downstream target genes in the entire myotube, which may be mediated by DUX4 protein spreading to multiple nuclei as well as signal amplification by downstream target transcription factors, such as DUXA. Although our current study is limited to FSHD2 primary myocytes from tibialis anterior, our strategy should be effective in further analyzing FSHD pathophysiology during different stages of muscle differentiation and in biopsies and muscle types with different sensitivities to the disease at a single nucleus resolution.

2.5 Materials & Methods

2.5.1 Human myoblast culture and differentiation

Human control and FSHD2 myoblast cells from patient quadricep and tibia biopsies were grown on dishes coated with collagen in high glucose DMEM (Gibco) supplemented with 20% FBS (Omega Scientific, Inc.), 1% Pen-Strep (Gibco), and 2% Ultrasor G (Crescent Chemical

Co.) [21]. Upon reaching 80% confluence, differentiation was induced by using high glucose DMEM medium supplemented with 2% FBS and ITS supplement (insulin 0.1%, 0.000067% sodium selenite, 0.055% transferrin; Invitrogen). Fresh differentiation medium was changed every 24hrs.

2.5.2 Bulk, single-nucleus and single-cell RNA-seq library preparation and sequencing

For bulk RNA-seq for the time-course, total RNA was extracted by using the RNeasy kit (QIAGEN). Between 19 and 38 ng of RNA were converted to cDNA using the Smart-Seq 2 protocol [38]. Libraries were constructed with the Nextera DNA Library Prep Kit (Illumina) for control-3, control-4, and FSHD2-2 libraries, and the Nextera DNA Flex Library Prep Kit (Illumina) for control-1, control-2 and FSHD2-1 libraries. Libraries were sequenced on the Illumina NextSeq500 platform using paired-end 43 bp mode with 15 million reads per sample.

Full-length single-cell and single-nucleus RNA-seq was performed according to [21] using the Fluidigm C1 with the following modifications. Myotube single nuclei were isolated from mononucleated cells (MNCs) by washing a 6 cm dish once with trypsin, then adding trypsin for about 5 min until myotubes lifted off the plate and MNCs were still attached. Cells were initially pelleted at 2000 rpm for 2 min and resuspended in lysis buffer with 0.02% IGEPAL CA-630. Lysis was done for 3 minutes, filtered and spun at 4000 rpm for 1 minute. Nuclei were captured on medium IFCs (10-17 um) at a density between 340 and 640 nuclei/ul in a volume of 10 ul. Visual confirmation was aided with the LIVE/DEAD kit (Thermo Fisher Scientific), and cDNA was normalized to approximately 0.1 ng/ul for tagmentation and library prep. Libraries were quality-controlled prior to sequencing based on Agilent 2100 Bioanalyzer profiles and normalized using the KAPA Library Quantification Kit (Illumina). Libraries were sequenced on the Illumina NextSeq500 platform using paired-end 75 bp mode with 1-3 million reads per sample for full-length RNA-seq single-cell and single-nucleus libraries.

Single-nucleus 3' end RNA-seq libraries from nuclei isolated on the ddSeq Single Cell Isolator (BioRad) were prepared as follows. Myotubes from day 3 or day 5 of differentiation were isolated from mononucleated cells (MNCs) by washing a 35 mm dish once with trypsin, then adding trypsin for about 5 min until myotubes lifted off the plate and MNCs were still attached. Cells were washed once in 1X PBS + 0.1% BSA, and the nuclei were prepared according to [39] with the following modifications. We used 0.2 U/ul RNasin Plus RNase Inhibitor (Promega) for the cell lysis buffer and nuclei storage buffer, and nuclei were filtered through a 40 um filter (Falcon) after isolation. Nuclear isolation and quality were assessed by staining with ethidium homodimer. Nuclei were loaded onto the ddSeq Single Cell Isolator (BioRad) for droplet generation, and libraries were prepared using the SureCell WTA 3' Library Prep Kit (Illumina). Libraries were sequenced on the Illumina NextSeq500 platform using PE 68 bp for read 1 and 75 bp for read 2 with a custom primer with around 370 million reads for four samples.

2.5.3 RNA FISH (Fluorescent *in situ* hybridization targeting ribonucleic acid molecules) by RNAScope

FSHD2-2 myoblasts were seeded in micro-slide eight-well plates at ~8x10^4 cells per well, and differentiation was initiated ~20hrs later. After 3 or 7 days, as indicated, of differentiation, cells were fixed with 10% neutral buffered formalin at room temperature for 30 min, and the RNA FISH experiments were performed using the RNAScope fluorescent Multiplex system (Advanced Cell Diagnostic Inc.) according to the manufacturer's instructions.

For costaining of immunofluorescent (IF) staining and RNAScope, cells were permeabilized with 0.5% Triton X-100 for 5 min at 4°C between fixation and dehydration process, then DUX4 (Abcam, ab124699) IF was performed as previously described [40]. Probe-Hs-DUX4-C1, Probe-Hs-LEUTX-C2, were custom-designed to avoid crossreactivity to related homologs (for *DUX4* probe set, see Figure S2.11A). Probe-Hs-SLC34A2-C3 was also used. Images were acquired using a Zeiss LSM 510 META confocal microscope. A technical consideration should be made that due to the process of IF and RNAScope costaining that much of the cytoplasmic RNAScope signal is washed out.

2.5.4 Quantification of differentiation index in myosin heavy chain 1 (MYH1) stained control and FSHD2 cells

Control-2 and FSHD2-2 cells were fixed with 4.0% paraformaldehyde in PBS for 10 min at room temperature, and cells were permeabilized with 0.5% Triton X-100 for 5 min at 4°C. Then MYH1 (ABclonal, Inc., A6935) IF was performed as previously described [40]. Differentiation index is defined as the number of nuclei in myotubes expressing MYH1 divided by the total number of nuclei in a field. We determined the differentiation index by counting at least 600 nuclei from 3 random fields on the coverslip which was fixed at each time point of differentiation.

2.5.5 shRNA depletion of DUX4 target genes

Lentiviruses carrying shRNA plasmids for each DUX4 target gene: *DUXA* (5'-CTAGATTACTTCTCCAGAGAA-3', TRCN0000017664), *LEUTX* (5'-CCTGGAATCTCTGATGCAAAT-3', TRCN0000336862), *ZSCAN4* (5'-

CCCAAGATACTTCCTTAGAAA-3’, TRCN0000016848) and an shRNA non-targeting control (Sigma-Aldrich, SHC002) were made in 293T cells using Lipofectamine 3000. The cells were transfected with 2 ug of shRNA plasmids, 1.5 ug of pCMV plasmids, and 0.5 ug of pMP2G plasmids. The media was changed after 24 hours. The lentiviruses were harvested at 48 hours and 72 hours post-transfection. FSHD2-2 myoblasts were infected once at 32 hours and once at 8 hours prior to addition of differentiation media. The myoblasts were selected for plasmid integration using puromycin. RNA was extracted using RNeasy kit (Qiagen) at days 4 and 6 of differentiation. Approximately 16 ng of RNA was converted to cDNA using SuperScript VILO (Invitrogen), and expression quantitation of *DUXA*, *LEUTX*, *ZSCAN4*, and *GAPDH* was done via RT-qPCR using SYBR green (Invitrogen) and the primers listed in Table 2.1.

2.5.6 RNA-seq data processing

Raw reads from both bulk RNA-seq and single-cell and single-nucleus RNA-seq were mapped to hg38 by STAR (version 2.5.1b) [41] using defaults except with a maximum of 10 mismatches per pair, a ratio of mismatches to read length of 0.07, and a maximum of 10 multiple alignments. Quantitation was performed using RSEM (version 1.2.31) [42] by defaults with gene annotations from GENCODE v28, and results were output in transcripts per million (TPM). Myoblast cells were kept for downstream analysis if *desmin* expression was >=1 TPM, *MYOG* <1 TPM, number of expressed genes was more than 500 and expression level of *GAPDH* was higher than 100 TPM. Myotube nuclei were kept for downstream analysis if *MYOG* expression was >=1 TPM, number of expressed genes was more than 500 and expression level of *GAPDH* was higher than 100 TPM. We only kept cells and nuclei with a uniquely mapped efficiency higher than 45%. For differential gene expression analysis in differentiation time-course, protein coding and long non-coding RNA genes with greater than 5 TPM in both replicates in at least one timepoint and with greater than 1 TPM for both reps for both cell lines of the same disease and day were kept. Genes were TMM normalized using edgeR (version 3.18.1) [43] and log2-transformed. For the bulk RNA-seq time-course, Batch correction was performed using ComBat from sva (version 3.32.1) and scaled for two batches which used different library prep kits; control-3, control-4, FSHD2-2 for one batch, and control-1, control-2, FSHD2-1 for the second. LogFC and p-values of FSHD-induced genes was calculated using edgeR with p-value <0.05. Clustering of genes across the time-course was done by using maSigPro using an r-squared of 0.66 [23]. Comparisons in Figure 2.5D, 2.6C, 2.6D, 2.6E and S2.9 were done using a t-test, and FDR was used where indicated (stats package version 3.6.1).

Sequencing data from 3' end RNA-seq was demultiplexed using ddSeekR [44]. Nuclei with at least 500 UMIs detected were mapped using STAR (version 2.5.1b) [41] and quantitated using RSEM (version 1.2.31) [42] with the *rsem-calculate-expression* with options *--star* and *--estimate-rspd*. We kept nuclei with ≥150 genes detected and <20% mitochondrial reads. Genes detected in at least 5 nuclei were kept for downstream analysis. The data was loaded into Seurat (version 3.1.0) and normalized using the SCTransform function [42, 43]. Seurat was also used to create UMAPs, determine clusters and calculate average expression. Heatmap of average expression was created using ComplexHeatmap (version 2.0.0) [45]. For overlap of full-length RNA-seq data with 3' end RNA-seq data, we apply SCTransform to both sets individually, then use the integration pipeline in Seurat to combine the datasets [46,47]. Differentially expressed genes were called using a t-test and FDR calculated from the stats (version 3.6.1) package with an FDR cutoff of 0.05 and a log2FC cutoff of 1. Fold change between the groups was calculated using average expression calculated in Seurat. Gene ontology analysis was done by using

Metascape [48] with the whole genome as the background set and an FDR <0.05. Transcription factor and DNA binding protein enrichment was done using enrichR (version 2.1) [49] with an adjusted p-value cutoff of 0.05. Transcription factors and cofactors identified from AnimalTFDB (version 3.0) [50]. Gene coexpression networks were plotted by using Cytoscape [51] using counts or TPM >0.

2.5.7 Binding site analysis of DUXA and DUX4

We used binding motifs from HOCOMOCO v11 [32] for DUX4 and DUXA as input into HOMER (version 4.10) using the scanMotifGenomeWide.pl command for hg38 [52]. Motif logos were generated using LogOddsLogo [53].

2.5.8 Reanalysis and comparisons of previously published data

Fastq files from [20, 22, 25] (Table 2.2) were obtained from SRA and mapped and quantitated as described above. We kept genes with greater than 1 TPM either for all experimental or FSHD samples or control samples. Genes with a logFC >2 and p-value <0.01 as calculated by edgeR were considered differentially expressed. For comparisons with [29], we report the 95% confidence interval calculated using prop.test from stats (version 3.6.1). We use the DUX4-affected cell counts found in Supplemental table 4 of [29].

2.5.9 Data availability

All bulk, single-cell, and single-nucleus RNA-seq data along with their associated meta data are deposited in the GEO database, series reference number GSE143493 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143493).

2.6 Figures

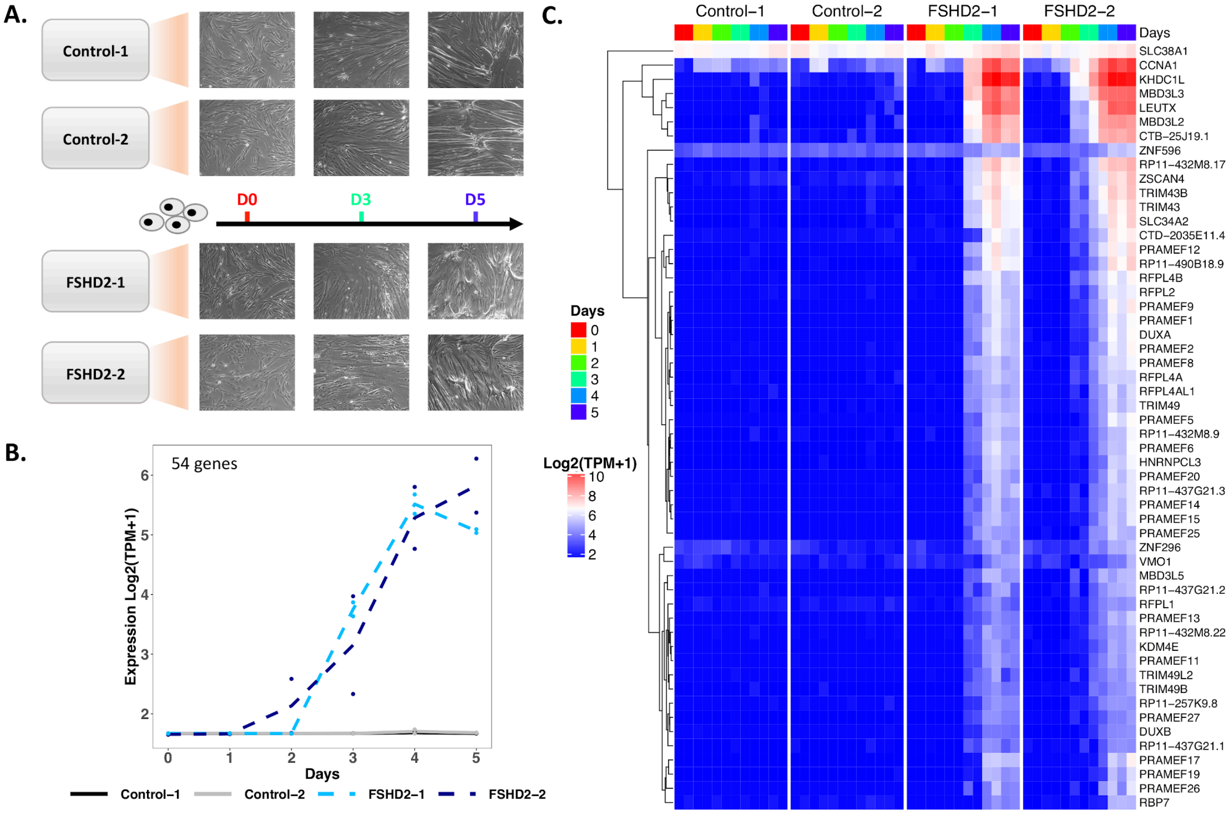


Figure 2.1: Upregulation of FSHD-induced genes starting at day 2 identified in bulk RNA-seq time-course (A) Differentiation time-course of control and FSHD2 patient-derived myoblasts to myotubes. Morphology changes are shown for days 0, 3 and 5 of differentiation. (B) Average expression profile of 54 genes upregulated in FSHD2 cells starting at day 2 of differentiation. maSigPro clustered 10,827 genes into three clusters based on their expression patterns during control and FSHD2 differentiation time-course. (C) Hierarchical heatmap of gene expression values of the 54 genes from (B). Expression values in transcripts per million (TPM) are TMM and log normalized. We refer to these 54 genes and *DUX4* as “FSHD-induced genes”.

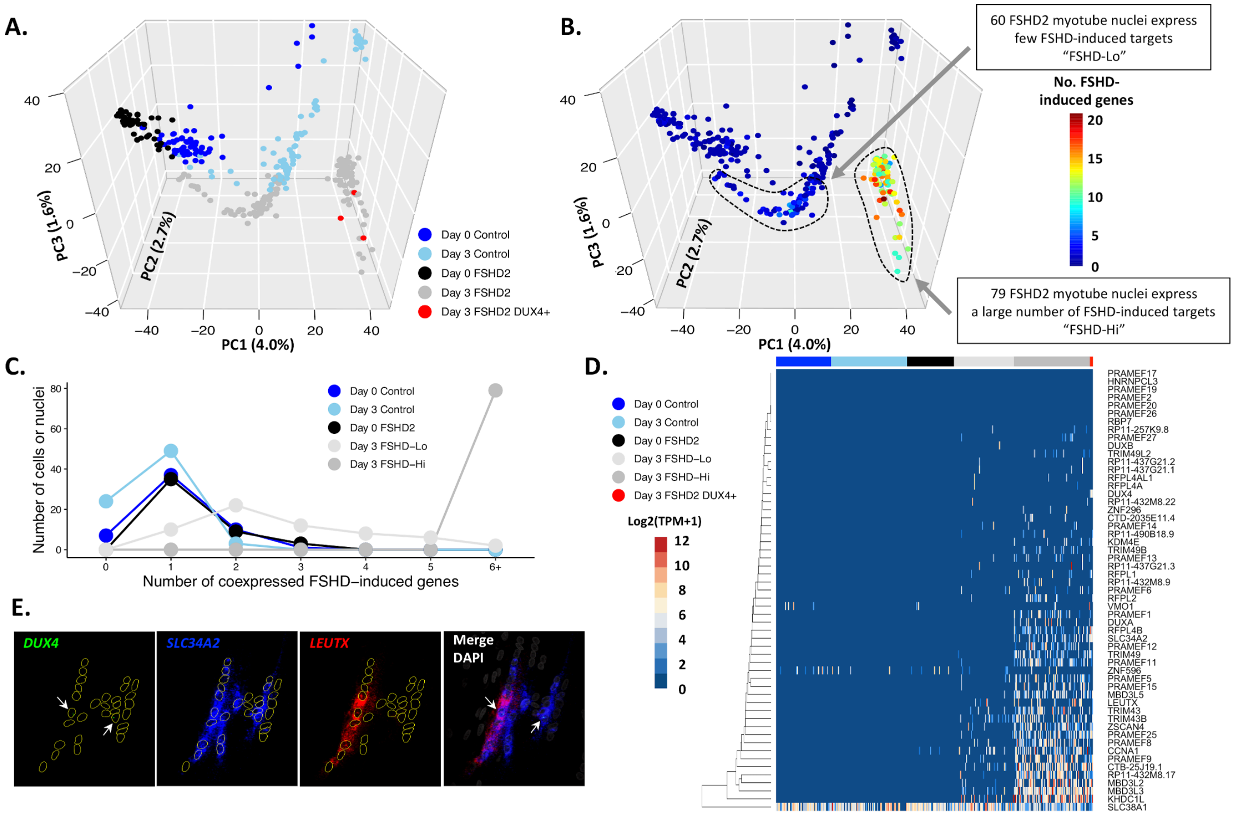


Figure 2.2: FSHD2 myotube nuclei can be separated into two clusters with differential expression of FSHD-induced genes (A) PCA of single-cell (for myoblast) and single-nucleus (for myotube) RNA-seq data for control-3 and FSHD2-2. Cell types are labeled by color, and three DUX4-detected FSHD2 myotube nuclei are specifically labeled in red. (B) PCA from panel (A) colored by the number of FSHD-induced genes detected (TPM >1) defined in Figure 2.1. (C) Summary of the number of FSHD-induced genes coexpressed (TPM >0) in different cell types. Cell lines and days are labeled by color. (D) Heatmap of the expression of FSHD-induced genes in single-cell myoblasts and single-nuclei from myotubes. The bar is colored by cell line and day. (E) RNA FISH (RNAScope) of *DUX4*, *LEUTX* and *SLC34A2* in FSHD2 myotubes at day 3 of differentiation. *DUX4*, green; *LEUTX*, red; *SLC34A2*, blue; DAPI, white. Arrow indicate *DUX4* spots in green. We examined 240 myotubes, of which 11 myotubes were found to be *DUX4*-positive and 7 of them co-expressed both *LEUTX* and *SLC34A2* while 2 co-expressed *SLC34A2* only. Two additional myotubes expressed *LEUTX*/*SLC34A2* without detectable *DUX4* signal, and 4 appear to express *SLC34A2* only.

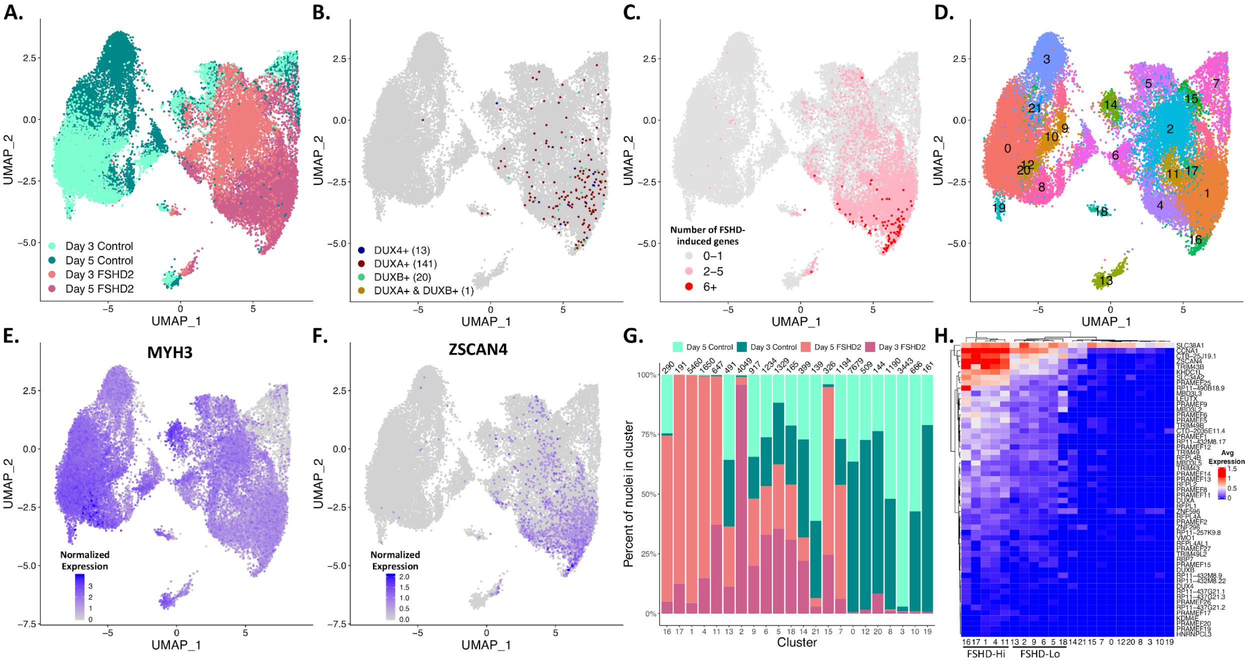


Figure 2.3: Day 3 and day 5 FSHD2 myotube nuclei cluster based on expression of FSHD-induced genes (A) Day 3 and day 5 myotube nuclei from control-2 and FSHD2-2 plotted on a UMAP based off of expression values from 3' end sequencing from libraries prepared using the BioRad's ddSeq. Control and FSHD2 nuclei separate across component 1 (UMAP_1). Day 3 and day 5 nuclei separate within cell type with some mixing. (B) UMAP from (A) colored by detection of *DUX4*, *DUXA* and/or *DUXB*. The number of nuclei in which we detect (counts >0) the indicated gene is in parentheses. (C) UMAP from (A) colored by the number of FSHD-induced genes detected (counts >0). (D) UMAP from (A) colored by cluster determined by shared nearest neighbors (SNN). Each cluster is colored and labeled by its respective number. (E) UMAP from (A) colored by expression of the myogenic marker *MYH3*. (F) UMAP from (A) colored by expression of the FSHD-induced gene *ZSCAN4*. (G) The percent of control and FSHD2 nuclei from day 3 or day 5 in each cluster from (D). The total number of nuclei in each cluster is indicated above each bar. Colored by cell line and day of differentiation. (H) Average expression profiles of the FSHD-induced genes in each cluster. Rows and clusters are ordered and dendrogram is calculated using Euclidean distance.

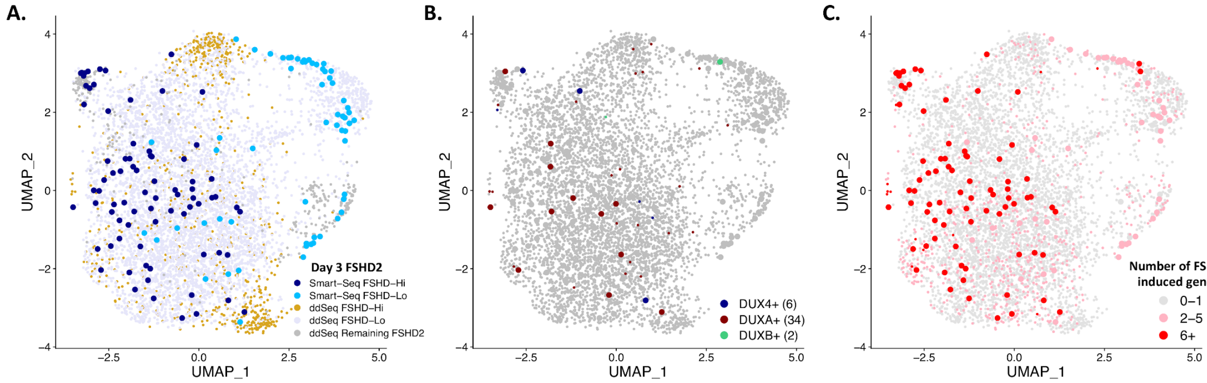


Figure 2.4: Day 3 FSHD2-2 nuclei from Fluidigm and ddSeq mix (A) UMAP with day 3 FSHD2-2 myotube nuclei from Smart-Seq and ddSeq. Nuclei are colored by technology and classification as FSHD-Hi or FSHD-Lo. (B) UMAP from (A) colored by detection (counts >0) of *DUX4*, *DUXA* and/or *DUXB*. The number of nuclei in which we detect the indicated gene is in parentheses. (C) UMAP from (A) colored by the number of FSHD-induced genes detected (counts >0).

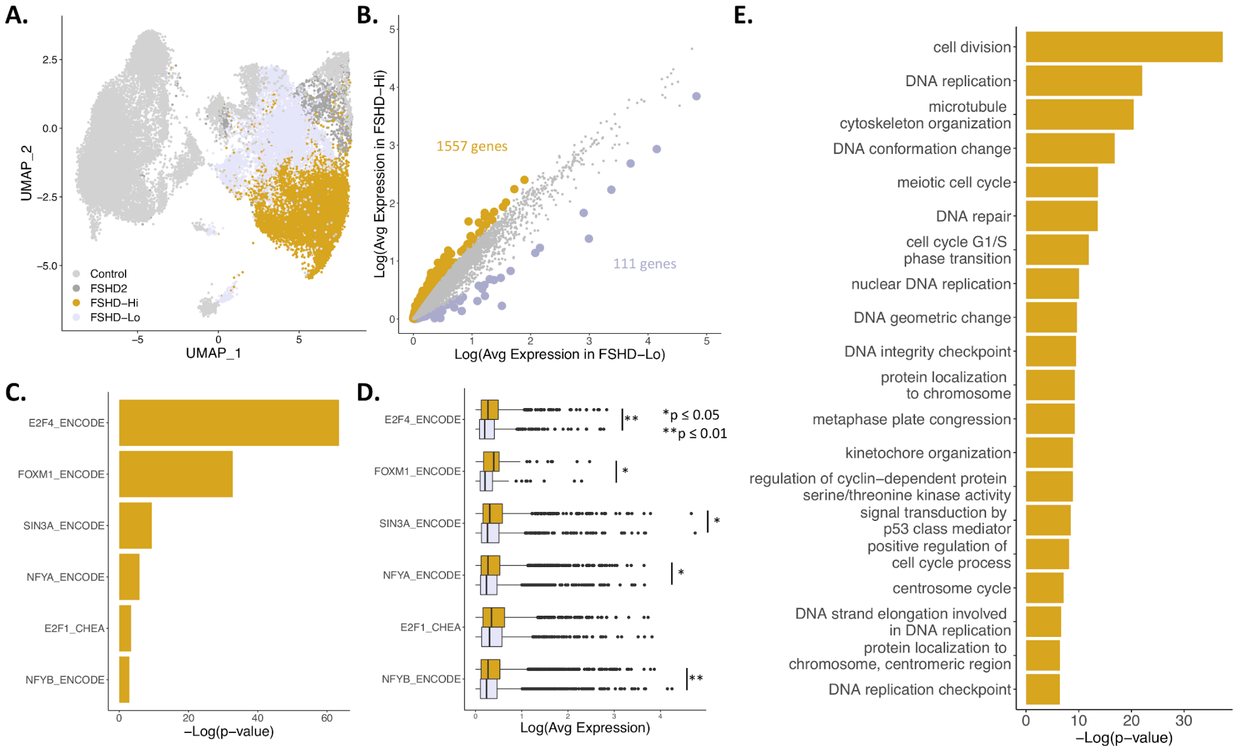


Figure 2.5: FSHD-Hi nuclei upregulate cell cycle transcription factors and genes (A) UMAP from Figure 2.3A colored by designation of FSHD-Lo or FSHD-Hi from ddSeq nuclei. (B) Scatterplot of average expression of 19,615 genes in the Low and FSHD-Hi populations. Highlighted are genes with FDR <0.05 and abs(log2FC) >1. In gold are genes with higher average expression in FSHD-Hi nuclei. In lavender are genes with higher average expression in FSHD-Lo nuclei. (C) Transcription factors and DNA binding proteins with enrichment for binding, as identified from ENCODE and ChEA ChIP-seq datasets, genes significantly higher in the FSHD-Hi population than the FSHD-Lo population. (D) Boxplot of average expression of target genes of indicated transcription factors or DNA binding proteins. In gold is the average expression of the targets in the FSHD-Hi nuclei. In lavender is the average expression of the same targets in the FSHD-Lo nuclei. Significance calculated with t-test. All significant differences are marked by asterisks, and p-value is adjusted with FDR. (E) Gene ontology terms associated with the 1,557 genes more highly expressed in FSHD-Hi.

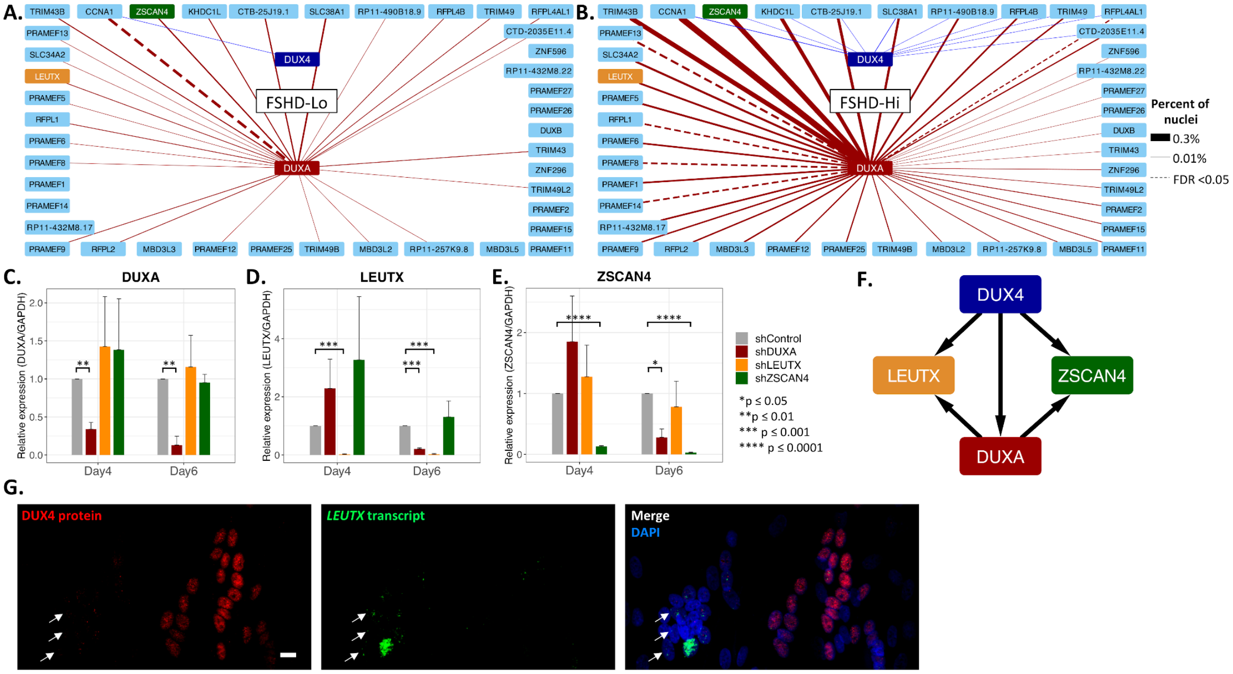


Figure 2.6: DUXA regulates FSHD-induced genes Coexpression network of 41 FSHD-induced genes which are coexpressed with *DUX4* and/or *DUXA* in (A) the ddSeq FSHD-Lo population and (B) the ddSeq FSHD-Hi population. Line thickness is the percent of nuclei coexpressing the two genes. Red lines represent coexpression with *DUXA*. Blue lines represent coexpression with *DUX4*. Dashed lines indicates FDR <0.05 using Fisher's exact test. (C-E) RT-qPCR analyses of the effects of lentiviral shRNA depletion of three DUX4 target transcription factors. Relative RNA expression of (C) *DUXA*, (D) *LEUTX*, and (E) *ZSCAN4* on days 4 and 6 of differentiation in FSHD2-2 cells following depletion of each gene product as indicated is shown. Expression measured by qPCR, and values are normalized to *GAPDH* expression and the non-targeting shControl. Significance calculated with t-test, and n=3 for each condition. All significant differences are marked by asterisks. Color indicates the shRNA used as listed on the right. (F) Proposed model for DUXA regulating FSHD-induced genes in addition to DUX4. (G) Expression of DUX4 protein and its downstream target gene are not always concordant. Immunofluorescence detection of DUX4 protein (red) and RNAScope for *LEUTX* transcript (green) in FSHD2-2 day 7 myotubes. Examples of a *LEUTX transcript*-positive myotube with no significant DUX4 protein (left) and a DUX4 protein-positive myotube with very little LEUTX transcript signal (right) are shown. Bar, 10 µm. DAPI is in blue. Nuclei with *LEUTX* transcripts with no DUX4 protein are indicated by white arrows.

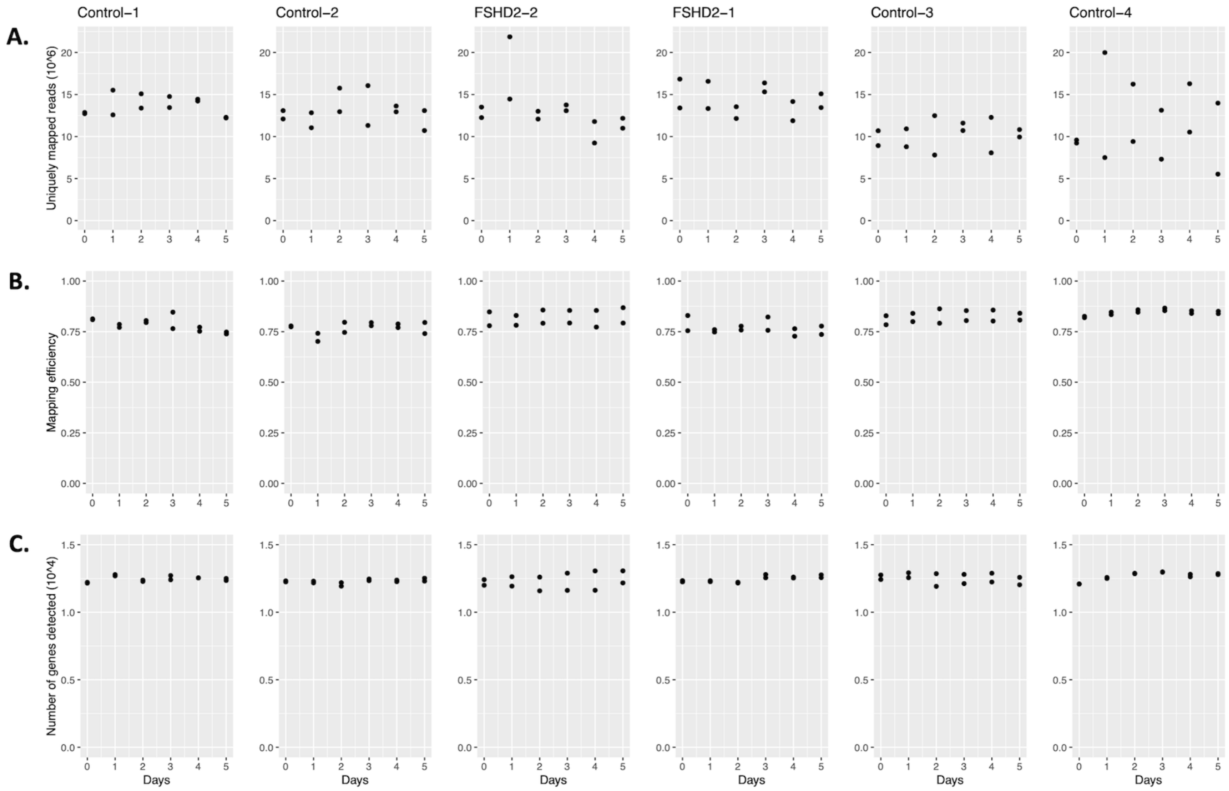


Figure S2.1: Quality metrics of RNA-seq time-course data Control and FSHD2 time-course quality metrics for (A) the number of uniquely mapped reads, (B) mapping efficiency, (C) the number of genes detected (TPM>=1).

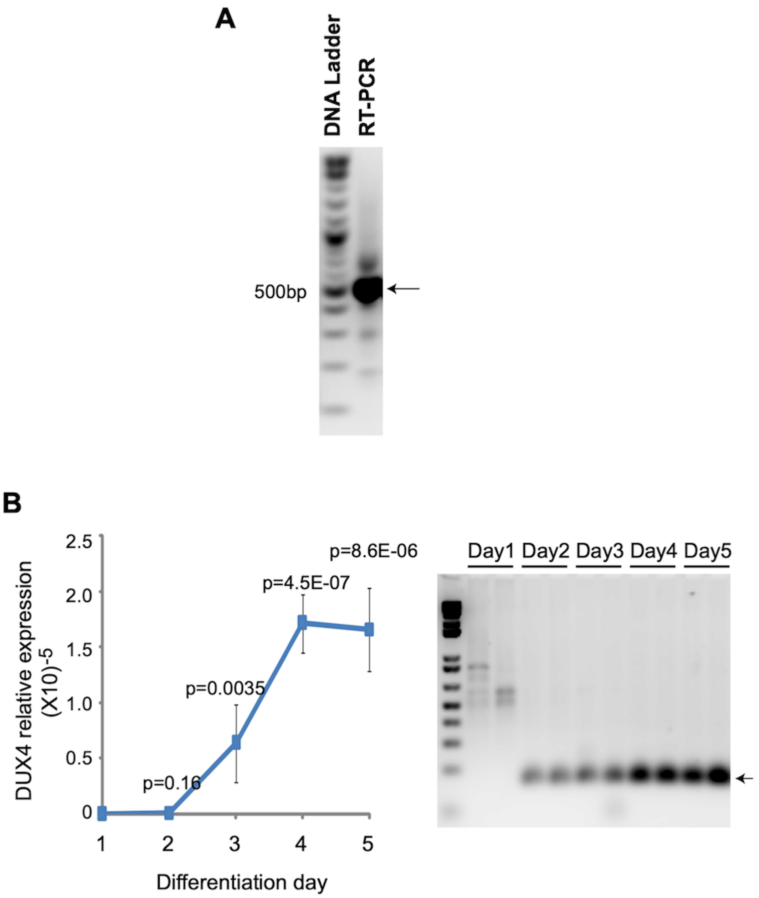


Figure S2.2: Expression of *DUX4-fl* in FSHD2-2 cells (A) Nested RT-PCR analysis of *DUX4-fl* expression in differentiated FSHD2-2 cells at day 3. The PCR product was sequenced to confirm its identity. The nested PCR was done using the primer sets (182–183 and 1A–184) previously published [2]. (B) FSHD2-2 cells were incubated in differentiation medium for the indicated days, and RT-qPCR was used to assess DUX4 mRNA expression during differentiation. Left, RT-qPCR data are normalized to GAPDH and the graph shows the relative abundance of *DUX4* mRNA at indicated time points. Error bars are standard deviation. P values comparing to Day 1 were shown. At Day 1, the *DUX4* mRNA is so low that nonspecific PCR product was amplified. Other PCR product was verified by sequencing. The qPCR primers are 5'-CCCAGGTACCAGCAGACC-3' and 5'-TCCAGGAGATGTAACTCTAATCCA-3' [9]. Right: the qPCR products were run on the gel and their identity was confirmed by sequencing (data not show).

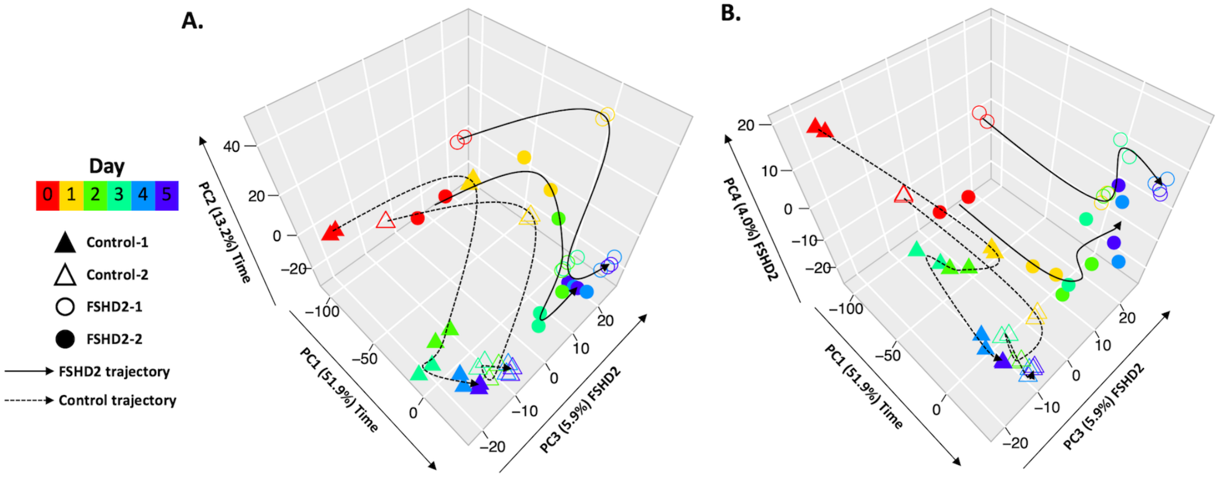


Figure S2.3: Principal component analysis (PCA) on control and FSHD2 myoblast differentiation time-course (A) PCA with PC1, PC2 and PC3 for FSHD2 and control myoblasts from tibialis anterior. PC2 further explains the expression variance across differentiation. (B) PCA with PC1, PC2, and PC3 for controls from tibialis anterior (TA) and controls from quadricep (quad). PC2 and PC3 combined explain the expression variance for muscle source and sex. Gene expression level was measured each day for duplicates by using RNA-seq. Cell types are labeled by shape, and time-points are labeled by color.

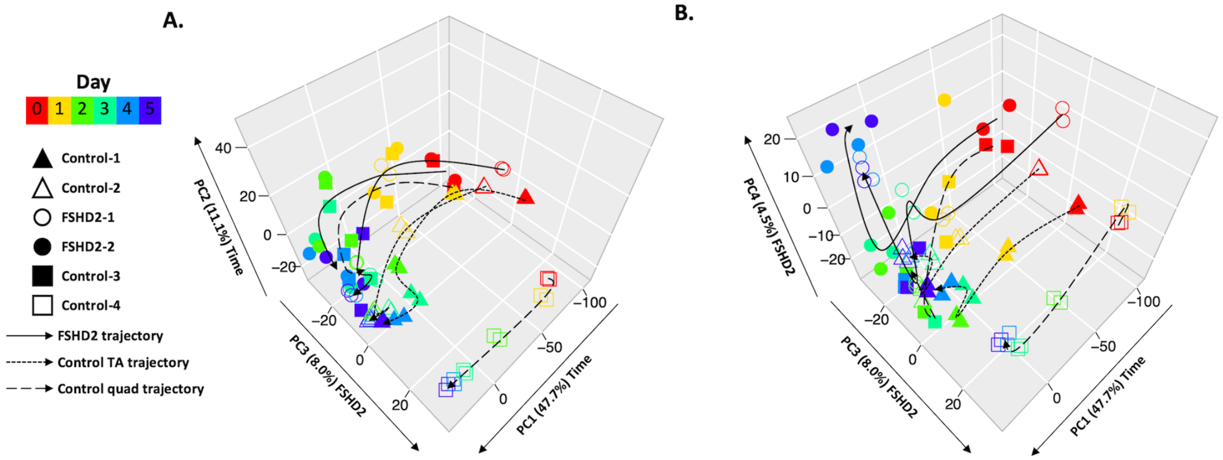


Figure S2.4: Principal component analysis (PCA) on control and FSHD2 myoblast differentiation time-course (A) PCA with PC1, PC2, and PC3 for FSHD2, controls from tibialis anterior (TA) and controls from quadricep (quad). PC2 further explains the expression variance across differentiation. (B) PCA with PC1, PC3, and PC4 for FSHD2, controls from tibialis anterior (TA) and controls from quadricep (quad). PC3 and PC4 account for variation in gene expression between FSHD2 and control samples. Gene expression level was measured each day for duplicates by using RNA-seq. Cell types are labeled by shape, and time-points are labeled by color.

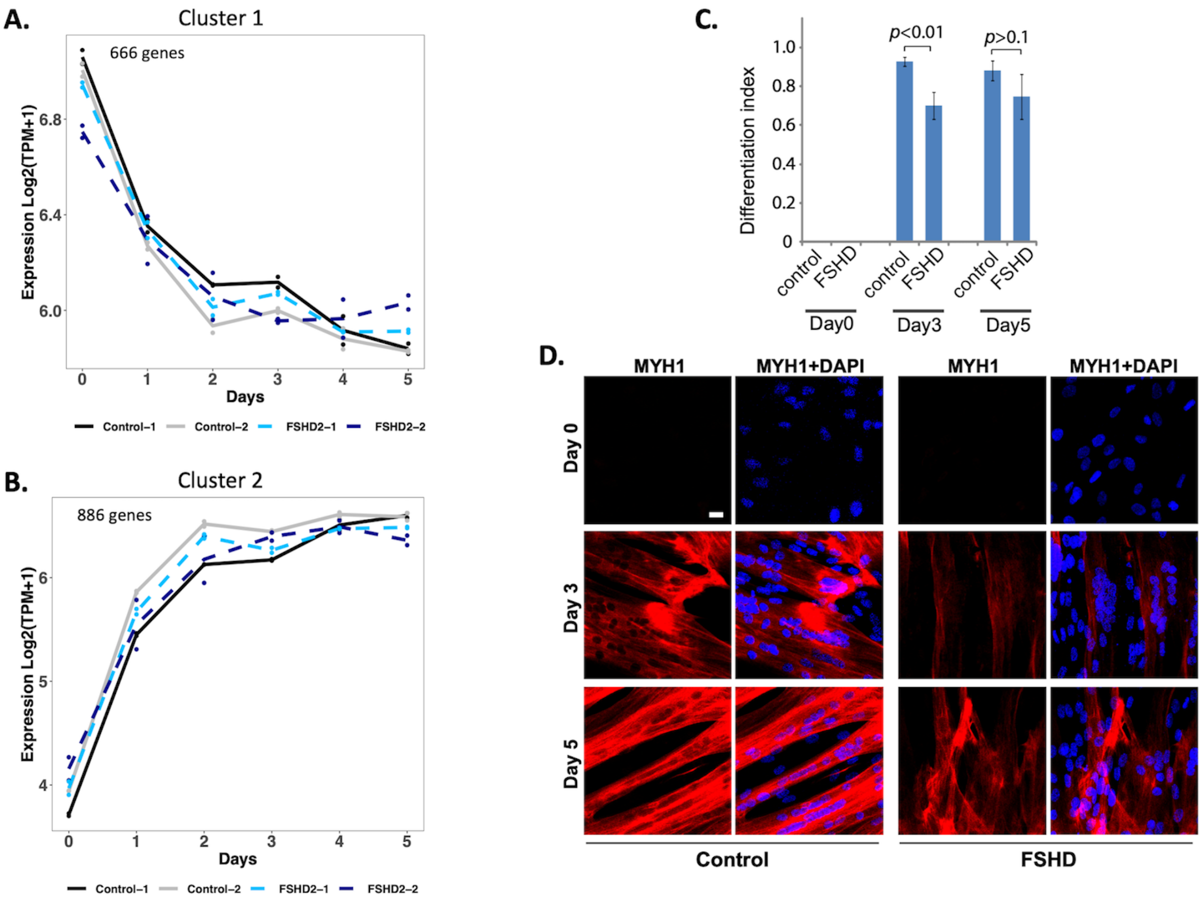


Figure S2.5: Genes variable across time but not between FSHD and control form two clusters (A) Cluster 1 gene decrease during differentiation. (B) Cluster 2 gene increase during differentiation. (C) Quantification of differentiation index in myosin heavy chain1(MYH1) stained control-2 and FSHD2-2 myoblast cell lines for days 0, 3 and 5 of differentiation. Differentiation index is defined as the number of nuclei in myotubes expressing MYH1 divided by the total number of nuclei in a field. We determined the differentiation index by counting at least 600 nuclei from 3 random fields on each coverslip which was fixed at indicated days after differentiation. Myotubes with any detectable MYH1 signal are considered positive, and the signal strength of MYH1 staining is not taken into consideration. Statistically significant delay of differentiation was observed in FSHD myocytes compared to the control used on day 3 (~70% as opposed to 90%). On day 5, differentiation index is still lower in FSHD than control but the difference is no longer statistically significant. (D) Representative images of differentiation marker MYH1 (red) staining of days 0, 3 and 5 of differentiation in control-2 and FSHD2-2 cells. Bar, 10 µm. DAPI is in blue.

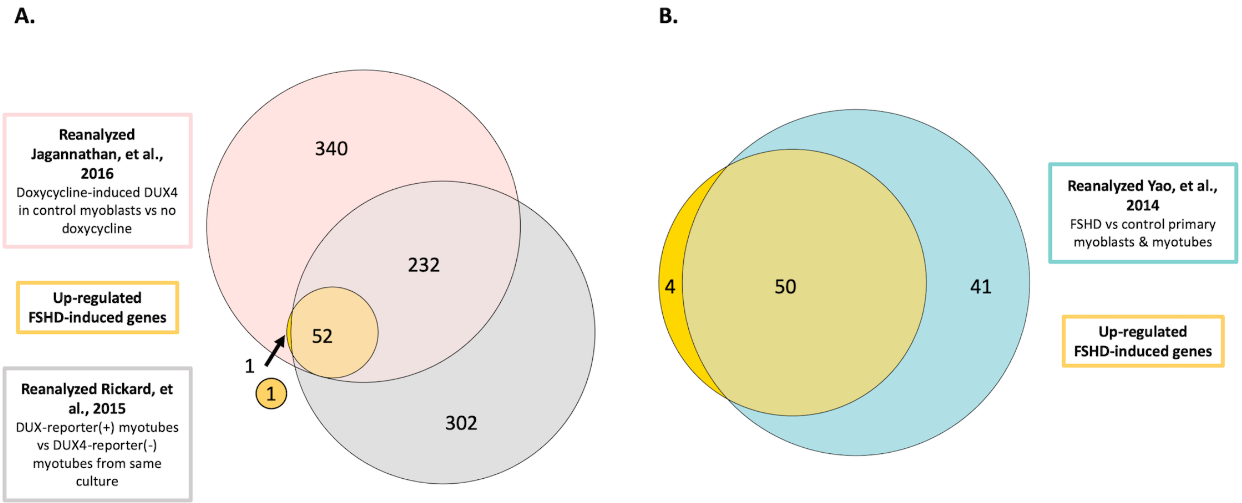


Figure S2.6: Venn diagram of FSHD-induced genes from this study and published FSHD and DUX4 induced genes (A) Overlap of 53 of the 54 genes upregulated during FSHD2 differentiation time-course from myoblasts to myotubes compared to 625 genes upregulated in myoblasts with doxycycline induced *DUX4* expression [25] and to 587 genes upregulated in *DUX4* expressing myotubes over non-expressing myotubes [22]. Published data was reanalyzed using the same analysis pipeline (Methods). (B) Overlap of 54 genes upregulated during FSHD2 differentiation time-course from myoblasts to myotubes compared to 91 genes upregulated in FSHD primary myoblasts and myotubes compared to control [20]. Published data was reanalyzed using the same analysis pipeline (Methods).

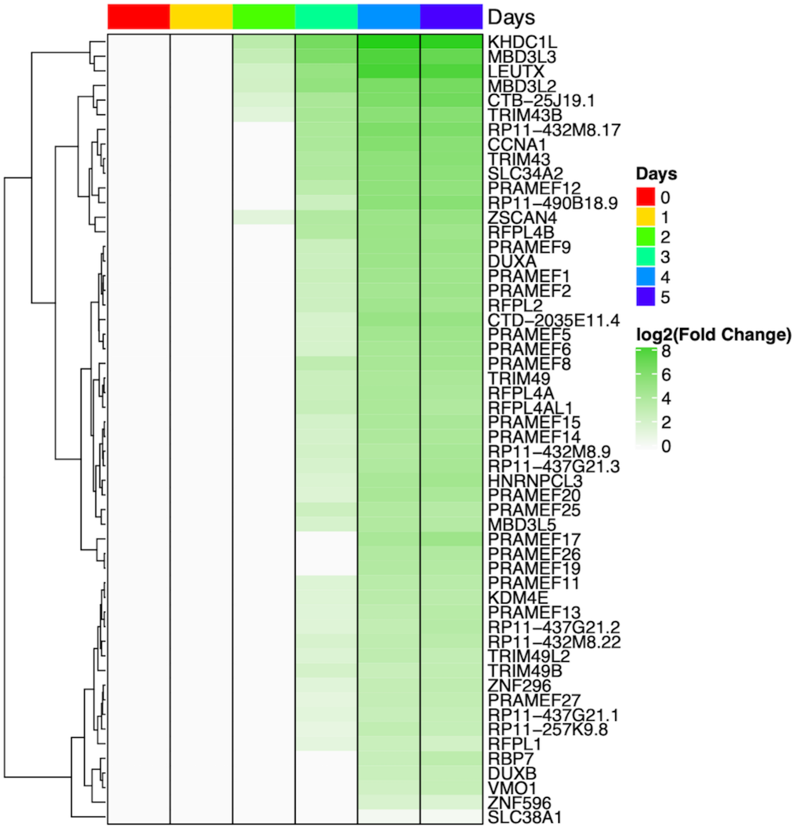
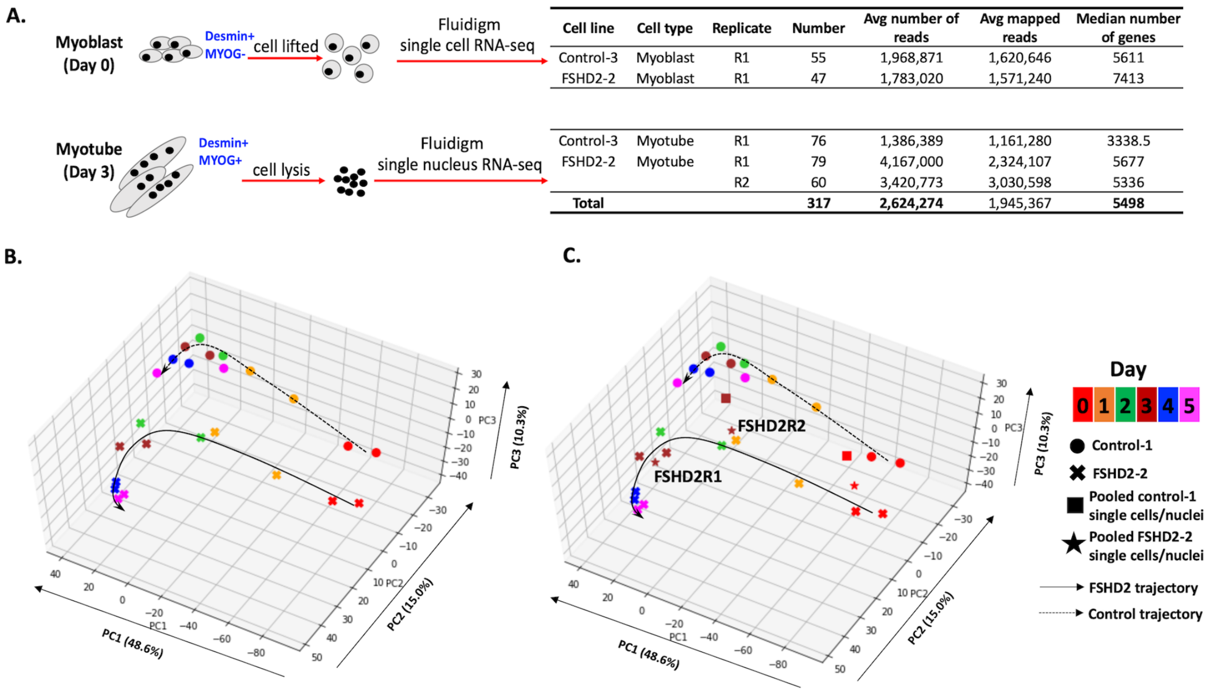


Figure S2.7: Fold change heatmap of FSHD-induced genes for FSHD2-1 and FSHD2-2 vs control-1 and control-2 All logFC with p <0.05 are shown for comparisons of FSHD2 to control for each day of differentiation.



Cell line	Cell type	Replicate	Number	Avg number of reads	Avg mapped reads	Median number of genes
Control-3	Myoblast	R1	55	1,968,871	1,620,646	5611
FSHD2-2	Myoblast	R1	47	1,783,020	1,571,240	7413
Control-3	Myotube	R1	76	1,386,389	1,161,280	3338.5
FSHD2-2	Myotube	R1	79	4,167,000	2,324,107	5677
		R2	60	3,420,773	3,030,598	5336
Total			**317**	**2,624,274**	1,945,367	**5498**

Figure S2.8: Overview of single-cell and single-nucleus samples from Fluidigm and comparison with time-course (A) Summary of single cells and single nuclei collected for sequencing. Single cells from myoblasts were selected to be *desmin*(+) *MYOG*(-) cells and retained for downstream analysis. Single nuclei from myotubes were selected to be *desmin*(+) *MYOG*(+) nuclei and retained for downstream analysis. Average number of reads, average number of mapped reads, and median number of genes detected are given per cell or nucleus for each sample. (B) Principal component analysis (PCA) of Control-1 and FSHD2-2 myoblast differentiation time-course. Gene expression level was measured each day for duplicates by using RNA-seq. Cell types are labeled by shape, and time-points are labeled by color. (C) Incremental PCA on pooled Control-1 single cells and pooled FSHD2-2 single nuclei as well as bulk Control-1 and FSHD2-2 differentiation time-courses with the same dimensions as the PCA in (B).

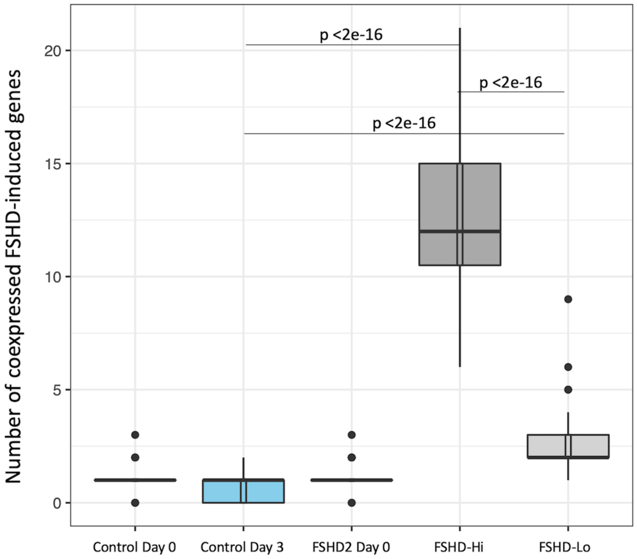


Figure S2.9: Differences in the number of FSHD-induced genes from the time-course which are detected across sample types Comparison of the number of FSHD-induced genes detected (TPM >1) from time-course analysis across different cell types. P-values are calculated with Wilcoxon and adjusted to FDR. Not all significant p-values are shown.

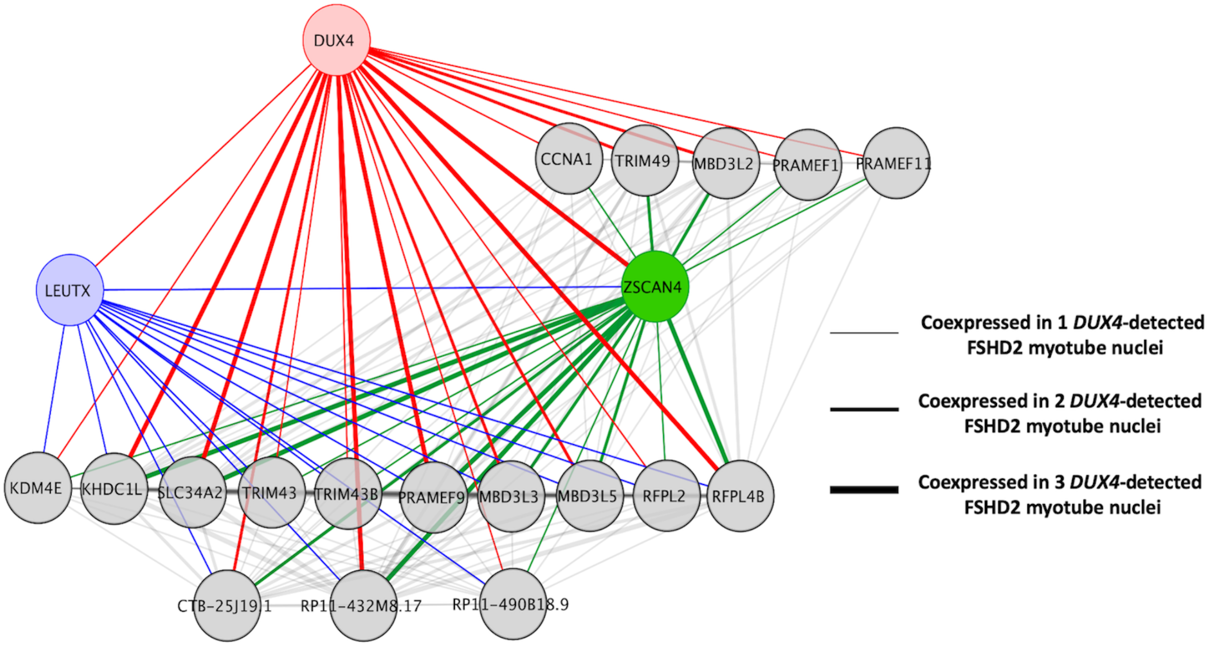

DUX4
LEUTX
ZSCAN4
CCNA1
TRIM49
MBD3L2
PRAMEF1
PRAMEF11
KDM4E
KHDC1L
SLC34A2
TRIM43
TRIM43B
PRAMEF9
MBD3L3
MBD3L5
RFPL2
RFPL4B
CTB-25J19.1
RP11-432M8.17
RP11-490B18.9
Coexpressed in 1 DUX4-detected FSHD2 myotube nuclei
Coexpressed in 2 DUX4-detected FSHD2 myotube nuclei
Coexpressed in 3 DUX4-detected FSHD2 myotube nuclei

Figure S2.10: Coexpression network of genes in the three *DUX4*-detected nuclei Twenty-three FSHD-induced genes are coexpressed (TPM >0) with *DUX4*, two of which are transcription factors, *LEUTX* and *ZSCAN4*.

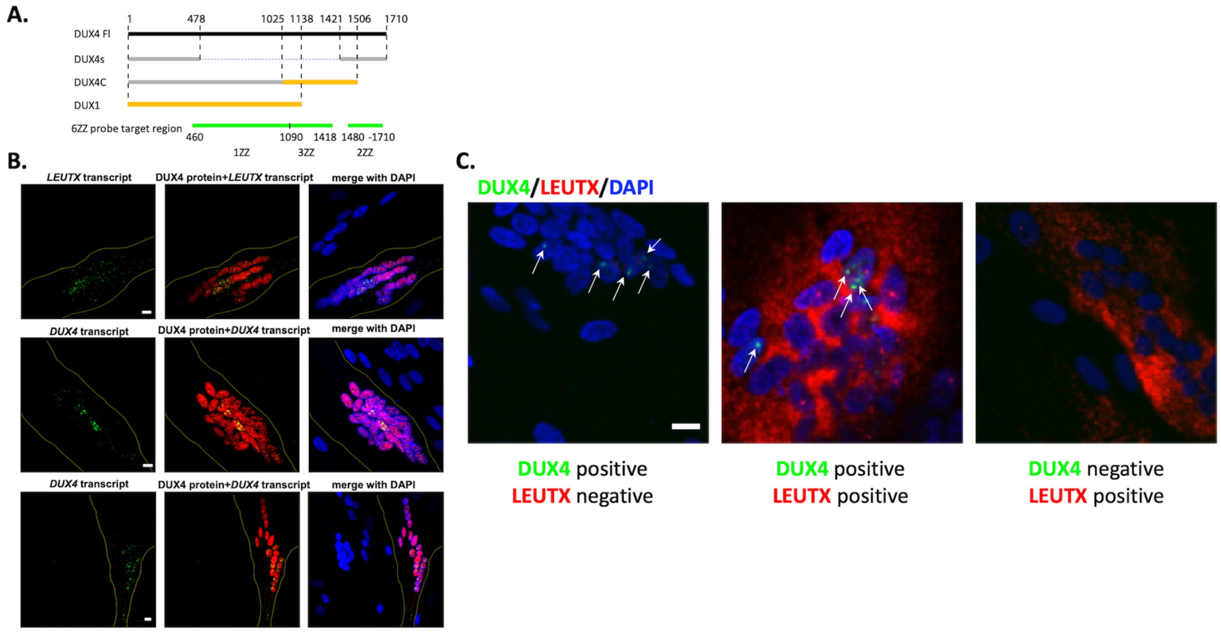
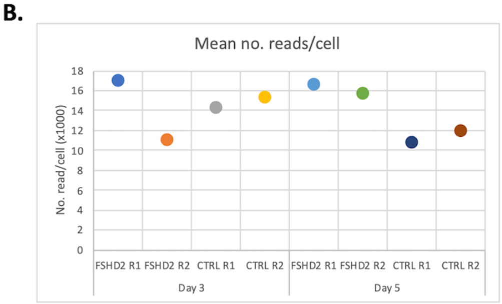
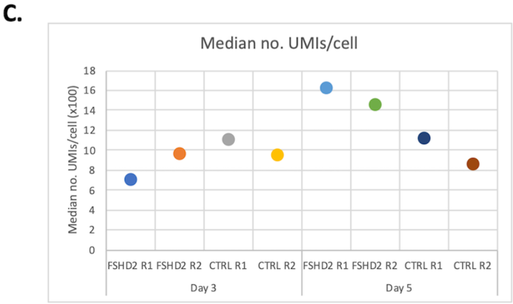


Figure S2.11: RNA FISH and IF of *DUX4* and *LEUTX* in FSHD2 myotubes at days 3 and 7 of differentiation (A) DUX4 RNAScope probe design. Schematic diagrams of *DUX4fl* mRNA (NM_001306068.2) and its isoform DUX4s and homologs (*DUX4C* and *DUX1)*. The "gray" sequence: almost 100% homology to *DUX4* mRNA. The "Orange" homologous sequences are different enough and would not be recognized by our *DUX4* probes. To minimize the crossdetection of *DUX4*s and *DUX4C*, we designed 6 ZZ probes (1 ZZ is a pair of RNAScope target probes): 1 ZZ falls in the region 460-1090 (common with *DUX4C*, but not in *DUX4s*), 3 ZZ in the region 1090-1418 (unique to *DUX4fl*, missing in *DUX4s* or *DUX4C*), and 2 ZZ in the region 1480-1710 (shared with *DUX4s* but missing in *DUX4C*) as indicated. Minimum 3 ZZ pairs are required for fluorescent RNAScope detection. (B) *LEUTX* (top) or *DUX4* (middle and bottom rows) RNAScopes are combined with immunofluorescence staining using antibody against DUX4 protein in FSHD2 myotubes at day 7 of differentiation. Myotubes containing positive *LEUTX* or *DUX4* RNA transcript signals are also positive for DUX4 protein staining. *LEUTX* or *DUX4* RNAScope signal, green; DUX4 antibody staining, red; DAPI, Blue. Yellow lines indicate the boundaries of DUX4 protein-positive myotubes. Scale bar, 10 µm. (C) *DUX4* (green) and *LEUTX* (red) RNAScope costaining in FSHD2-2 myotubes. DAPI is in blue. *DUX4* transcripts appear as nuclear foci (indicated with white arrows) while *LEUTX* transcripts are mostly diffuse in the cytoplasm with some additional nuclear foci. Scale bar, 10 µm

A.

Sample	No. cells after demultiplex	No. nuclei with ≥500 UMIs	No. cells with >150 genes detected	No. nuclei with <20% mito reads
Day 3 Control-2 R1	304,011	4404	4403	4365
Day 3 Control-2 R2	287,564	3445	3444	3390
Day 3 FSHD2-2 R1	328,961	4139	2328	2169
Day 3 FSHD2-2 R2	299,261	4078	4078	3938
Day 5 Control-2 R1	322,655	4231	4231	4200
Day 5 Control-2 R2	311,571	4804	4804	4770
Day 5 FSHD2-2 R1	425,562	4837	4837	4785
Day 5 FSHD2-2 R2	409,833	4656	4655	4611





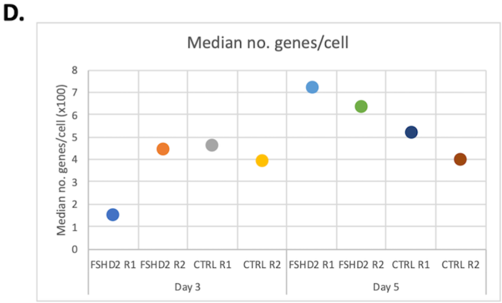


Figure S2.12: ddSeq 3' end RNA-seq quality metrics (A) Table of number of nuclei passing each quality filter. (B) Mean number of reads per cell for each ddSeq replicate. (C) Median number of UMIs per cell for each ddSeq replicate. (D) Median number of genes per cell for each ddSeq replicate.

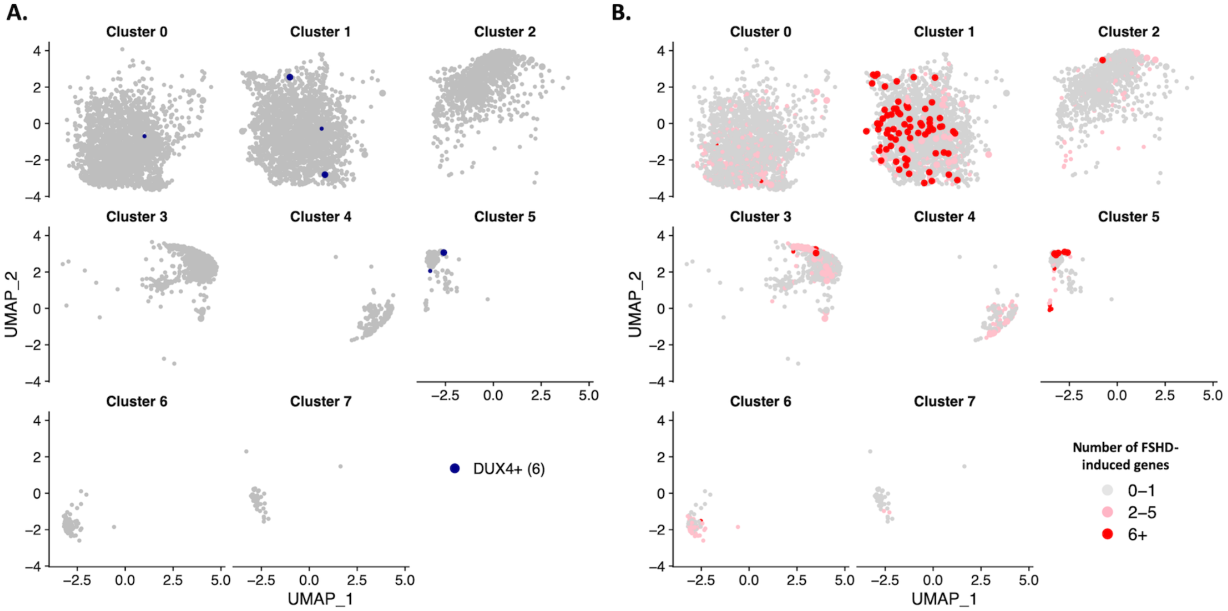


Figure S2.13: *DUX4*-detected nuclei do not exclusively cluster with nuclei with high number of FSHD-induced genes detected (A) UMAP from Figure 2.4A split by cluster. In blue are nuclei with DUX4 detected (counts >0). Larger points indicated nuclei data from the Fluidigm. (B) Same as A but colored by the number of FSHD-induced genes detected (counts >0).

A.

	Sample	Total cells count (after QC)	Number of cells expressing DUX4	Percentage of cells expressing DUX4	95% CI (Lower, Upper) for percentage of cells expressing DUX4	Number of DUX4-affected* cells	Percentage of DUX4-affected* cells	95% CI (Lower, Upper) for percentage of DUX4-affected* cells
van den Heuvel, et al., 2018	FSHD1.1 myocyte	2184	19	0.9	(0.54, 1.38)	12	0.5	(0.298, 0.987)
	FSHD1.2 myocyte	838	2	0.2	(0.0413, 0.958)	2	0.2	(0.0413, 0.958)
	FSHD2.1 myocyte	1225	5	0.4	(0.15, 1.01)	11	0.9	(0.473, 1.65)
	FSHD2.2 myocyte	655	1	0.2	(0.00797, 0.986)	2	0.3	(0.0529, 1.22)
	CTRL.1 myocyte	763	0	0	(0, 0.625)	0	0.0	(0, 0.625)
	CTRL.2 myocyte	1052	0	0	(0, 0.454)	0	0.0	(0, 0.454)
Smart-Seq	FSHD-Hi myotube	79	3	3.8	(0.986, 11.5)	79	100	(94.2, 100)
	FSHD-Lo myotube	60	0	0	(0, 7.5)	2	3.3	(0.579, 12.5)
	Control myotube	76	0	0	(0, 6)	0	0.0	(0, 6)
ddSeq	FSHD-Hi myotube	8135	9	0.1	(0.054, 0.218)	425	5.2	(4.76, 5.74)
	FSHD-Lo myotube	6210	3	0.05	(0.0125, 0.154)	59	1.0	(0.73, 1.23)
	Remaining FSHD2 myotube	1203	1	0.1	(0.00434, 0.538)	18	1.5	(0.916, 2.4)
	Control myotube	16725	0	0	(0, 0.0286)	0	0.0	(0, 0.0286)

*DUX4-affected cells were selected based on the criteria of expressing ≥5 genes of the DUX4-67 gene set [20] defined by [27].

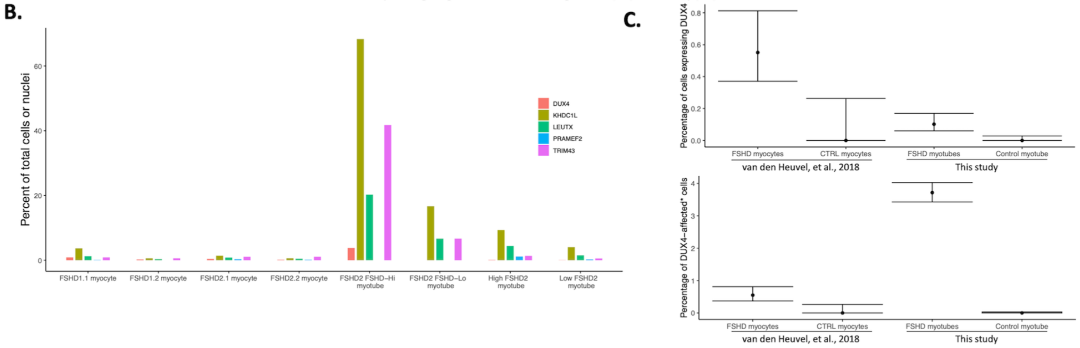


Figure S2.14: Comparison between published single-cell FSHD myocyte RNA-seq data [37] and single-nucleus FSHD myotube RNA-seq data in this study (A) Number and percentage of *DUX4* expressing and affected myocyte single cells in published study (Supplemental table 4 of [27]) and myotube single nuclei in this study. For this study, detected is considered TPM or counts >0. (B) Percentage of total cells/nuclei expressing *DUX4* and 4 FSHD markers in myocyte single cells [27] and myotube single nuclei. 4 FHSD markers were selected from the published study [27] as a quality check. (C) Percentage of cells expressing *DUX4* (top) and percentage of DUX4-affected cells (bottom) for all FSHD or control cells for [27] and this study with 95% confidence intervals.

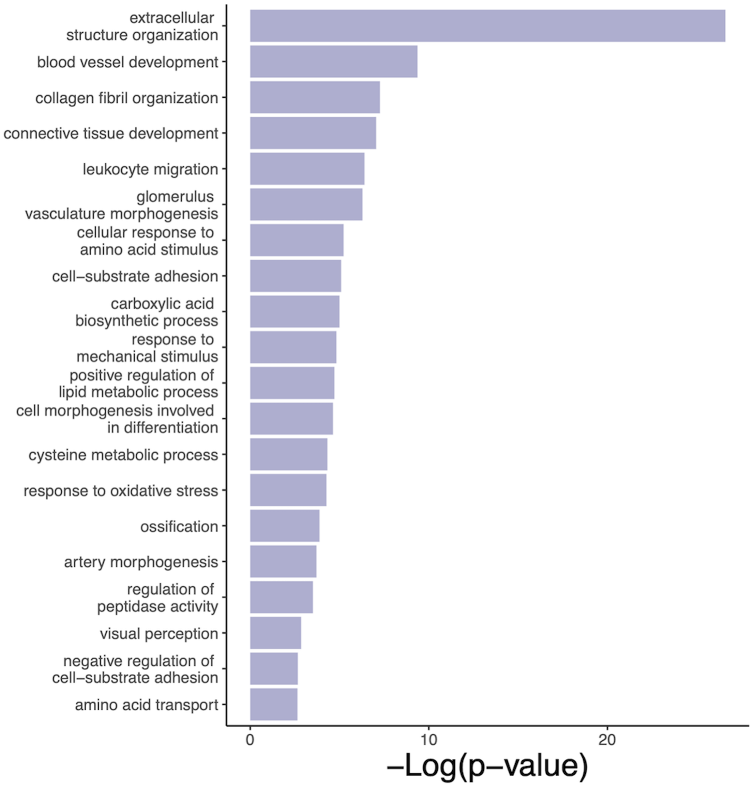
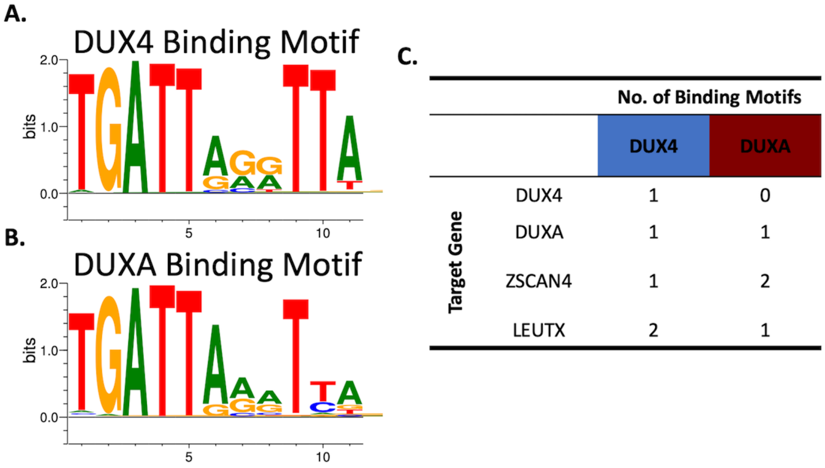


Figure S2.15: Gene ontology terms associated with genes upregulated in FSHD-Lo nuclei



		No. of Binding Motifs	
		DUX4	DUXA
Target Gene	DUX4	1	0
	DUXA	1	1
	ZSCAN4	1	2
	LEUTX	2	1

Figure S2.16: DUX4 and DUXA binding motifs in promoters of FSHD-induced genes (A) DUX4 and (B) DUXA binding motifs from HOCOMOCO v11. (C) Table of number of binding motifs for DUX4 and DUXA in the promoters of DUX4, DUXA, ZSCAN4 and LEUTX found using HOMER (Methods).

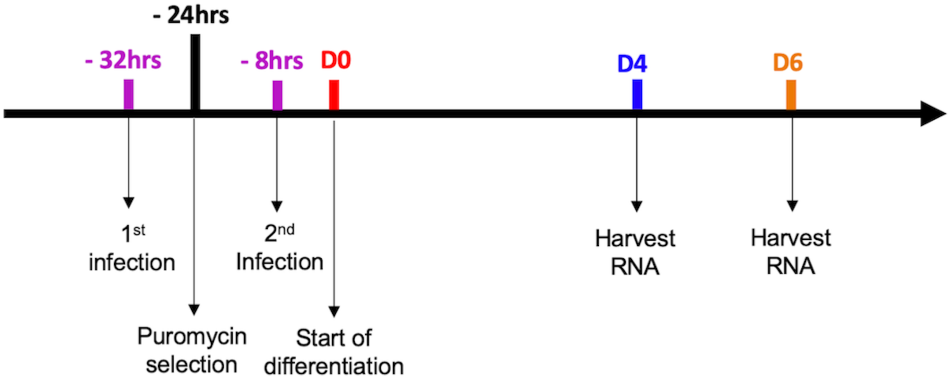


Figure S2.17: Schematic of shRNA knockdown and differentiation procedure in FSHD2-2 cells

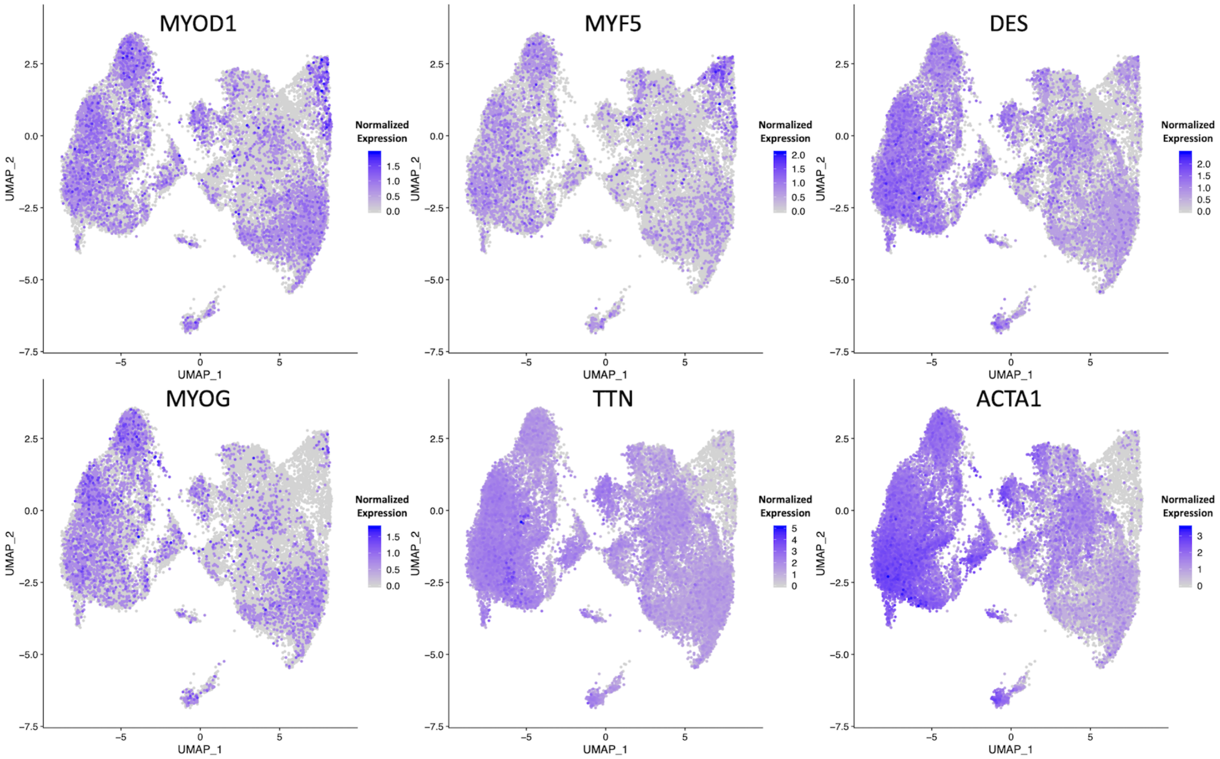


Figure S2.18: UMAPS of ddSeq nuclei colored by expression of myogenic markers

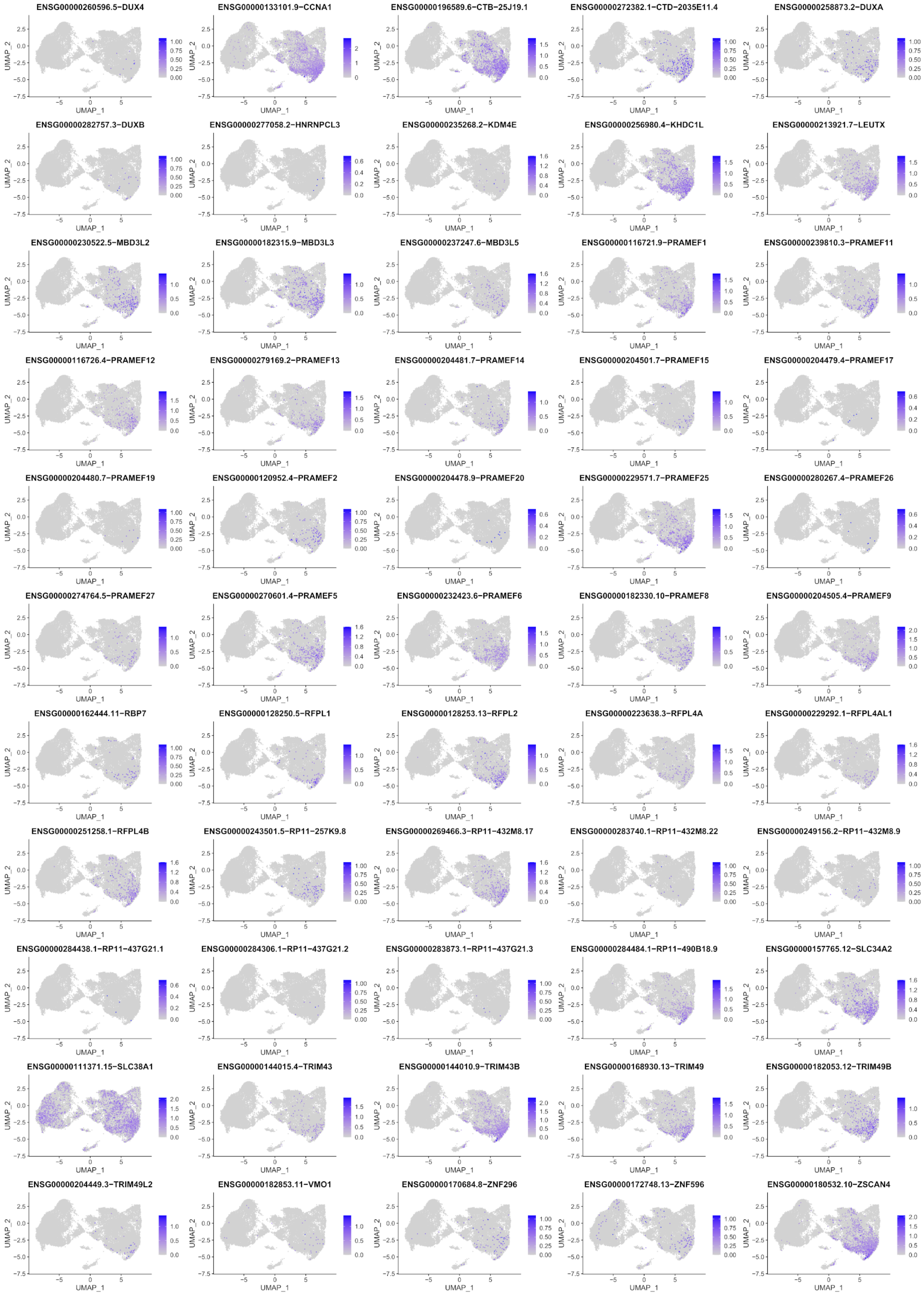


Figure S2.19: UMAPs of ddSeq nuclei colored by expression of indicated FSHD-induced gene ENSEMBL ID is given as well as gene name.

2.7 Tables

Table 2.1: Primer sequences used for qPCR

Primers	**Sequence**
ZSCAN4 Fwd	5' – TGGAAATCAAGTGGCAAAAA – 3'
ZSCAN4 Rev	5' – CTGCATGTGGACGTGGAC – 3'
LEUTX Fwd	5' – GGGAAACTGGCTTCAAAGCTA – 3'
LEUTX Rev	5' – TGATGGCCGTGTCTGCATTT – 3'
DUXA Fwd	5' – GCCTTACCCAGGTTATGCTACC – 3'
DUXA Rev	5' – TGGAATCCGTGCCTAGCTCTT – 3'
GAPDH Fwd	5' – TCGACAGTCAGCCGCATCT – 3'
GAPDH Rev	5' – CTAGCCTCCCGGGTTTCTCT – 3'

Table 2.2: Accession numbers for published datasets used in this paper

Reference	Sample Name	SRA
[24]	Sample_1-MB135_HDUX4CA_nodox_rep1	SRR4019004
[24]	Sample_2-MB135_HDUX4CA_WITHdox_rep1	SRR4019005
[24]	Sample_3-MB135_HDUX4CA_nodox_rep2	SRR4019006
[24]	Sample_4-MB135_HDUX4CA_WITHdox_rep2	SRR4019007
[24]	Sample_5-MB135_HDUX4CA_nodox_rep3	SRR4019008
[24]	Sample_6-MB135_HDUX4CA_WITHdox_rep3	SRR4019009
[22]	FSHD_1_1_neg	SRR2020583
[22]	FSHD_1_2_neg	SRR2020584
[22]	FSHD_2_2_BFP	SRR2020585
[22]	FSHD_2_3_BFP	SRR2020586
[22]	FSHD_1_3_neg	SRR2020587
[22]	FSHD_1_1_BFP	SRR2020588
[22]	FSHD_1_2_BFP	SRR2020589
[22]	FSHD_1_3_BFP	SRR2020590
[22]	FSHD_2_1_neg	SRR2020591
[22]	FSHD_2_2_neg	SRR2020592
[22]	FSHD_2_3_neg	SRR2020593
[22]	FSHD_2_1_BFP	SRR2020594
[20]	Control_20_Mt	SRR1398556
[20]	Control_21_Mb	SRR1398557
[20]	Control_21_Mt	SRR1398558
[20]	Control_22_Mb	SRR1398559
[20]	Control_22_Mt	SRR1398560
[20]	FSHD2_12_Mt	SRR1398561
[20]	FSHD2_14_Mb	SRR1398562
[20]	FSHD2_14_Mt	SRR1398563
[20]	FSHD2_20_Mb	SRR1398564
[20]	FSHD2_20_Mt	SRR1398565
[20]	FSHD1_4_Mb	SRR1398566
[20]	FSHD1_4_Mt	SRR1398567
[20]	FSHD1_6_Mb	SRR1398568
[20]	FSHD1_6_Mt	SRR1398569

Table S2.1: Cell line information Muscles biopsies were from either the tibialis anterior (TA) or the quadricep (quad). Percent methylation in D4Z4 region measured by FseI digestion.

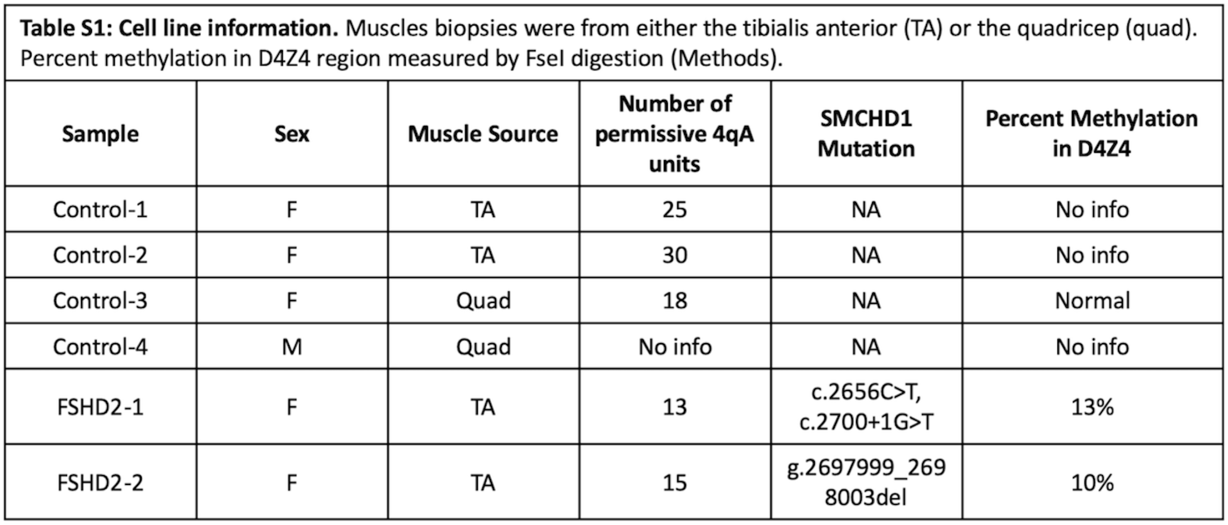


Table S1: Cell line information. Muscles biopsies were from either the tibialis anterior (TA) or the quadricep (quad). Percent methylation in D4Z4 region measured by FseI digestion (Methods).

Sample	Sex	Muscle Source	Number of permissive 4qA units	SMCHD1 Mutation	Percent Methylation in D4Z4
Control-1	F	TA	25	NA	No info
Control-2	F	TA	30	NA	No info
Control-3	F	Quad	18	NA	Normal
Control-4	M	Quad	No info	NA	No info
FSHD2-1	F	TA	13	c.2656C>T, c.2700+1G>T	13%
FSHD2-2	F	TA	15	g.2697999_2698003del	10%

Table S2.2: Gene ontology for clusters from maSigPro (Figure 2.1B and S2.5) Gene ontology enrichment performed with metascape, keeping only summary terms with FDR <0.05 for GO biological processes.

Cluster	Term	Description	LogP	Log(q-value)	InTerm_InList	Example 5 Genes
1	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	-11.5	-7.6	38/379	FUS,HNRNPA1,HNRNPF,MAGOH,HNRNPM
1	GO:1901137	carbohydrate derivative biosynthetic process	-8.4	-4.9	51/776	ADSS2,ACAN,DCN,DCTD,DDOST
1	GO:0015931	nucleobase-containing compound transport	-7.4	-4.0	24/241	SLC25A6,ZFP36L1,CETN3,HNRNPA1,EIF6
1	GO:0022613	ribonucleoprotein complex biogenesis	-5.6	-2.6	33/502	ADAR,FBL,GLUL,EIF6,NPM1
1	GO:0018205	peptidyl-lysine modification	-5.4	-2.4	28/400	ACTL6A,HDAC2,PER1,PLOD1,MAPK3
1	GO:0019439	aromatic compound catabolic process	-5.3	-2.4	41/717	ZFP36L1,DFFA,GALK1,HIF1A,HMOX1
1	GO:0006338	chromatin remodeling	-5.1	-2.3	17/182	ACTL6A,H3-3B,HDAC1,HDAC2,HMGB3
1	GO:0001890	placenta development	-4.8	-2.1	15/153	ANG,BIRC2,ZFP36L1,BSG,CTSV
1	GO:0090501	RNA phosphodiester bond hydrolysis	-4.7	-2.1	15/155	ANG,MOV10,EXOSC9,TSN,TSNAX
1	GO:0071396	cellular response to lipid	-4.6	-2.0	35/610	ATP2B1,ZFP36L1,TSPO,CDK4,CDK7
1	GO:0044387	negative regulation of protein kinase activity by regulation of protein phosphorylation	-4.5	-1.9	4/8	ADAR,NPM1,DUSP10,CORO1C
1	GO:0002446	neutrophil mediated immunity	-4.4	-1.9	30/500	ANXA2,CAPN1,CD59,COPB1,DDOST
1	GO:0001933	negative regulation of protein phosphorylation	-4.4	-1.9	27/429	ADAR,CALM2,FOXO1,GSTP1,SMAD7
1	GO:0006412	translation	-4.3	-1.8	38/715	ANG,ZFP36L1,MRPL49,CDK4,EGFR
1	GO:0010506	regulation of autophagy	-4.0	-1.7	22/330	TSPO,CAPN1,DCN,FOXO1,FOXO3
1	GO:0031647	regulation of protein stability	-4.0	-1.7	20/286	CCNH,CDK7,HSPD1,SMAD7,PPIB
1	GO:0006979	response to oxidative stress	-4.0	-1.6	27/454	ATOX1,TOR1A,EGFR,FOXO1,FOXO3

1	GO:0042273	ribosomal large subunit biogenesis	-3.9	-1.6	9/71	EIF6,NPM1,NVL,RPL7,RPL23A
1	GO:0046467	membrane lipid biosynthetic process	-3.9	-1.6	13/142	PIGF,TNFRSF1A,PLPP3,DPM2,PIGK
2	GO:0003012	muscle system process	-67.0	-62.8	116/465	ACTA1,ACTA2,ACTC1,ACTN2,ACTN3
2	GO:0061061	muscle structure development	-47.1	-43.3	114/674	ACTA1,ACTC1,ACTN2,ACTN3,AGT
2	GO:0048747	muscle fiber development	-18.1	-15.4	24/67	ACTA1,DMD,MYBPC1,MYBPC2,MYH6
2	GO:0014888	striated muscle adaptation	-16.2	-13.6	21/57	ACTA1,ACTN3,ATP2A2,CAMK2B,CAMK2G
2	GO:0055008	cardiac muscle tissue morphogenesis	-13.2	-10.7	20/69	ACTC1,MYBPC1,MYBPC2,MYH6,MYH7
2	GO:0070296	sarcoplasmic reticulum calcium ion transport	-12.6	-10.1	16/43	ATP1A2,ATP2A2,CACNG1,CALM1,CAMK2D
2	GO:0043462	regulation of ATPase activity	-9.8	-7.4	18/81	MYH6,MYL3,MYL4,RAB3A,SLN
2	GO:0014902	myotube differentiation	-9.7	-7.3	21/114	ACTA1,KLF5,DMPK,MEF2C,MYOG
2	GO:0014819	regulation of skeletal muscle contraction	-8.4	-6.2	8/14	ACTN3,CASQ1,DMD,DMPK,KCNJ2
2	GO:0071417	cellular response to organonitrogen compound	-7.8	-5.6	48/589	ACTA1,ACTN2,AGT,ATP6V1B2,ATP6V1C1
2	GO:0009150	purine ribonucleotide metabolic process	-6.3	-4.1	42/543	ACTN3,AK1,ALDOA,AMPD1,ATP1A2
2	GO:0010649	regulation of cell communication by electrical coupling	-6.0	-3.8	6/12	ANK3,CALM1,CAMK2D,CASQ2,HRC
2	GO:1902903	regulation of supramolecular fiber organization	-5.9	-3.8	31/352	ACTN2,BIN1,APOE,BBS4,CAPZA2
2	GO:0022411	cellular component disassembly	-5.7	-3.6	41/551	A2M,ACTN2,APEH,CAPZA2,CASP3
2	GO:0071415	cellular response to purine-containing compound	-5.7	-3.6	6/13	CACNA1S,CASQ2,P2RY6,RYR1,TMEM38B
2	GO:0033292	T-tubule organization	-5.7	-3.6	5/8	BIN1,ATP2A2,CSRP3,DYSF,SYPL2
2	GO:0070509	calcium ion import	-5.6	-3.5	13/80	ATP2A2,CACNA1S,CACNA2D1,PSEN2,CCL2

2	GO:0099641	anterograde axonal protein transport	-5.3	-3.3	5/9	DLG2,HSPB1,KIF5B,MAP1A,MAPK8IP3
2	GO:0042391	regulation of membrane potential	-5.3	-3.3	34/434	ACTN2,BIN1,ANK3,ATP1A2,ATP2A2
2	GO:0014874	response to stimulus involved in regulation of muscle adaptation	-5.3	-3.2	6/15	ACTN3,AGT,CASQ1,MYOG,PRKAG3
3	GO:0045596	negative regulation of cell differentiation	-12.9	-8.7	15/757	PRAMEF1,PRAMEF2,PRAMEF5,PRAMEF9,PRAMEF12
3	GO:0006346	methylation-dependent chromatin silencing	-7.2	-3.3	4/26	MBD3L2,MBD3L5,MBD3L3,MBD3L2B,KDM4E

Table S2.3: Transcription factors and cofactors differentially expressed between FSHD-Hi and FSHD-Lo

Higher in FSHD-Hi or FSHD-Lo	TF or Cofactor	Symbol	Ensembl	Family
FSHD-Lo	Cof	BCL9L	ENSG00000186174	BCL
FSHD-Lo	TF	GLI2	ENSG00000074047	zf-C2H2
FSHD-Lo	Cof	NOTCH3	ENSG00000074181	Notch
FSHD-Lo	Cof	RNF25	ENSG00000163481	Ring finger protein
FSHD-Lo	Cof	TGFB1	ENSG00000105329	Others
FSHD-Lo	TF	ZBED1	ENSG00000214717	zf-BED
FSHD-Hi	Cof	ABT1	ENSG00000146109	Others
FSHD-Hi	TF	AC025287.4	ENSG00000284484	Homeobox
FSHD-Hi	TF	AL033529.1	ENSG00000254553	ZBTB
FSHD-Hi	Cof	ANP32E	ENSG00000143401	ANP
FSHD-Hi	TF	ARGFX	ENSG00000186103	Homeobox
FSHD-Hi	Cof	ATAD2	ENSG00000156802	Others
FSHD-Hi	TF	ATOH8	ENSG00000168874	bHLH
FSHD-Hi	Cof	AURKB	ENSG00000178999	Others
FSHD-Hi	Cof	BEND3	ENSG00000178409	Others
FSHD-Hi	Cof	BIRC5	ENSG00000089685	Others
FSHD-Hi	Cof	BLM	ENSG00000197299	Others
FSHD-Hi	TF	BNC1	ENSG00000169594	zf-C2H2
FSHD-Hi	Cof	BRCA1	ENSG00000012048	Others
FSHD-Hi	Cof	BRIP1	ENSG00000136492	Others
FSHD-Hi	Cof	CASK	ENSG00000147044	CAMK
FSHD-Hi	Cof	CBX5	ENSG00000094916	Chromobox
FSHD-Hi	Cof	CCNA1	ENSG00000133101	Cyclin
FSHD-Hi	Cof	CCNA2	ENSG00000145386	Cyclin
FSHD-Hi	Cof	CCNE1	ENSG00000105173	Cyclin
FSHD-Hi	Cof	CDK1	ENSG00000170312	Cyclin
FSHD-Hi	Cof	CDK2	ENSG00000123374	Cyclin
FSHD-Hi	Cof	CDKN1C	ENSG00000129757	Cyclin
FSHD-Hi	TF	CENPA	ENSG00000115163	Others
FSHD-Hi	Cof	CENPF	ENSG00000117724	CENP
FSHD-Hi	Cof	CENPU	ENSG00000151725	CENP
FSHD-Hi	Cof	CHCHD3	ENSG00000106554	Coiled-coil
FSHD-Hi	Cof	CHD1	ENSG00000153922	CHD
FSHD-Hi	Cof	CHEK1	ENSG00000149554	Others
FSHD-Hi	Cof	CNOT7	ENSG00000198791	CCR4-NOT

FSHD-Hi	TF	CPHXL	ENSG00000283755	Homeobox
FSHD-Hi	TF	CREBRF	ENSG00000164463	Others
FSHD-Hi	Cof	CSRP3	ENSG00000129170	Others
FSHD-Hi	Cof	CTDP1	ENSG00000060069	Others
FSHD-Hi	Cof	CTH	ENSG00000116761	Others
FSHD-Hi	Cof	CTNNBIP1	ENSG00000178585	Casein
FSHD-Hi	Cof	DEK	ENSG00000124795	Other_CRF
FSHD-Hi	Cof	DEPDC1	ENSG00000024526	Others
FSHD-Hi	TF	DUXA	ENSG00000258873	Homeobox
FSHD-Hi	TF	DUXB	ENSG00000282757	Homeobox
FSHD-Hi	TF	E2F1	ENSG00000101412	E2F
FSHD-Hi	TF	E2F2	ENSG00000007968	E2F
FSHD-Hi	Cof	EAF1	ENSG00000144597	Others
FSHD-Hi	TF	EBF3	ENSG00000108001	COE
FSHD-Hi	Cof	EID3	ENSG00000255150	EID
FSHD-Hi	TF	EN2	ENSG00000164778	Homeobox
FSHD-Hi	Cof	ENY2	ENSG00000120533	Others
FSHD-Hi	TF	ETS2	ENSG00000157557	ETS
FSHD-Hi	Cof	EZH2	ENSG00000106462	Others
FSHD-Hi	TF	FOSL1	ENSG00000175592	TF_bZIP
FSHD-Hi	TF	FOXC1	ENSG00000054598	Fork_head
FSHD-Hi	TF	FOXM1	ENSG00000111206	Fork_head
FSHD-Hi	Cof	GMNN	ENSG00000112312	Others
FSHD-Hi	Cof	GPS2	ENSG00000132522	Others
FSHD-Hi	Cof	GTF2B	ENSG00000137947	GTF
FSHD-Hi	Cof	H2AFZ	ENSG00000164032	Others
FSHD-Hi	TF	HEYL	ENSG00000163909	bHLH
FSHD-Hi	TF	HIC2	ENSG00000169635	ZBTB
FSHD-Hi	Cof	HIST1H1E	ENSG00000168298	Histone cluster 1 H1
FSHD-Hi	TF	HKR1	ENSG00000181666	zf-C2H2
FSHD-Hi	TF	HMGB2	ENSG00000164104	HMG
FSHD-Hi	TF	HMGB3	ENSG00000029993	HMG
FSHD-Hi	TF	HMGXB4	ENSG00000100281	HMG
FSHD-Hi	TF	HOXA11	ENSG00000005073	Homeobox
FSHD-Hi	TF	HOXA7	ENSG00000122592	Homeobox
FSHD-Hi	TF	HOXB4	ENSG00000182742	Homeobox
FSHD-Hi	TF	HOXC13	ENSG00000123364	Homeobox
FSHD-Hi	Cof	ILF2	ENSG00000143621	Others

FSHD-Hi	TF	IRX5	ENSG00000176842	Homeobox
FSHD-Hi	Cof	ITGB3BP	ENSG00000142856	Others
FSHD-Hi	Cof	KDM3B	ENSG00000120733	Lysine demethylase
FSHD-Hi	Cof	KDM4D	ENSG00000186280	Lysine demethylase
FSHD-Hi	TF	KLF17	ENSG00000171872	zf-C2H2
FSHD-Hi	TF	KLF3	ENSG00000109787	zf-C2H2
FSHD-Hi	Cof	KMT2D	ENSG00000167548	Lysine methyltransferase
FSHD-Hi	Cof	LANCL2	ENSG00000132434	Others
FSHD-Hi	Cof	LBH	ENSG00000213626	Others
FSHD-Hi	TF	LEUTX	ENSG00000213921	Homeobox
FSHD-Hi	Cof	LIN54	ENSG00000189308	Others
FSHD-Hi	Cof	MAD2L2	ENSG00000116670	Others
FSHD-Hi	Cof	MAK	ENSG00000111837	Others
FSHD-Hi	Cof	MBD3L2	ENSG00000230522	MBD
FSHD-Hi	TF	MECOM	ENSG00000085276	zf-C2H2
FSHD-Hi	Cof	MED21	ENSG00000152944	Mediator complex
FSHD-Hi	Cof	MED26	ENSG00000105085	Mediator complex
FSHD-Hi	Cof	MED27	ENSG00000160563	Mediator complex
FSHD-Hi	Cof	MED29	ENSG00000063322	Mediator complex
FSHD-Hi	Cof	MED30	ENSG00000164758	Mediator complex
FSHD-Hi	Cof	MNAT1	ENSG00000020426	Others
FSHD-Hi	TF	MYBL2	ENSG00000101057	MYB
FSHD-Hi	Cof	MYCBP	ENSG00000214114	Others
FSHD-Hi	TF	MYCN	ENSG00000134323	bHLH
FSHD-Hi	TF	MYEF2	ENSG00000104177	Others
FSHD-Hi	Cof	MYOCD	ENSG00000141052	Others
FSHD-Hi	TF	MYOD1	ENSG00000129152	bHLH
FSHD-Hi	TF	NHLH1	ENSG00000171786	bHLH
FSHD-Hi	TF	NKX3-2	ENSG00000109705	Homeobox
FSHD-Hi	TF	NSD2	ENSG00000109685	HMG
FSHD-Hi	TF	OSR2	ENSG00000164920	zf-C2H2
FSHD-Hi	TF	PATZ1	ENSG00000100105	ZBTB
FSHD-Hi	TF	PAX9	ENSG00000198807	PAX
FSHD-Hi	TF	PBX3	ENSG00000167081	Homeobox
FSHD-Hi	Cof	PDLIM1	ENSG00000107438	Others
FSHD-Hi	Cof	PHF19	ENSG00000119403	PHF
FSHD-Hi	TF	PITX3	ENSG00000107859	Homeobox

FSHD-Hi	Cof	PLK1	ENSG00000166851	Others
FSHD-Hi	Cof	PNRC2	ENSG00000189266	Others
FSHD-Hi	Cof	POLR3F	ENSG00000132664	POLR
FSHD-Hi	TF	PRDM16	ENSG00000142611	zf-C2H2
FSHD-Hi	Cof	PTTG1	ENSG00000164611	Others
FSHD-Hi	TF	PURB	ENSG00000146676	Others
FSHD-Hi	Cof	RAD21	ENSG00000164754	Others
FSHD-Hi	TF	RARB	ENSG00000077092	THR-like
FSHD-Hi	TF	RBAK	ENSG00000146587	zf-C2H2
FSHD-Hi	Cof	RBL1	ENSG00000080839	Other_Co-activator/repressors
FSHD-Hi	TF	RLF	ENSG00000117000	zf-C2H2
FSHD-Hi	Cof	RNF111	ENSG00000157450	Ring finger protein
FSHD-Hi	Cof	RUVBL2	ENSG00000183207	Other_CRF
FSHD-Hi	Cof	SAFB2	ENSG00000130254	Others
FSHD-Hi	TF	SALL1	ENSG00000103449	zf-C2H2
FSHD-Hi	TF	SALL2	ENSG00000165821	zf-C2H2
FSHD-Hi	TF	SATB2	ENSG00000119042	CUT
FSHD-Hi	Cof	SFRP4	ENSG00000106483	Others
FSHD-Hi	Cof	SGF29	ENSG00000176476	Others
FSHD-Hi	Cof	SMARCD2	ENSG00000108604	SWI/SNF
FSHD-Hi	TF	SOX8	ENSG00000005513	HMG
FSHD-Hi	TF	SP100	ENSG00000067066	SAND
FSHD-Hi	Cof	SUMO1	ENSG00000116030	Others
FSHD-Hi	Cof	TADA1	ENSG00000152382	Transcriptional adaptor
FSHD-Hi	Cof	TAF1A	ENSG00000143498	TATA-box
FSHD-Hi	Cof	TAF4B	ENSG00000141384	TATA-box
FSHD-Hi	Cof	TAF5	ENSG00000148835	TATA-box
FSHD-Hi	Cof	TAF9B	ENSG00000187325	TATA-box
FSHD-Hi	TF	TCF12	ENSG00000140262	bHLH
FSHD-Hi	TF	TCF19	ENSG00000137310	Others
FSHD-Hi	TF	TCF3	ENSG00000071564	bHLH
FSHD-Hi	TF	TFAP2C	ENSG00000087510	AP-2
FSHD-Hi	TF	TFB1M	ENSG00000029639	Others
FSHD-Hi	TF	TFDP2	ENSG00000114126	E2F
FSHD-Hi	TF	THAP1	ENSG00000131931	THAP
FSHD-Hi	TF	THAP8	ENSG00000161277	THAP
FSHD-Hi	Cof	TIMELESS	ENSG00000111602	Others

FSHD-Hi	Cof	TOPORS	ENSG00000197579	Others
FSHD-Hi	TF	TOX	ENSG00000198846	HMG
FSHD-Hi	TF	TPRX1	ENSG00000178928	Homeobox
FSHD-Hi	Cof	TRAF6	ENSG00000175104	Others
FSHD-Hi	Cof	TRIM11	ENSG00000154370	Tripartite motif
FSHD-Hi	Cof	TRIP13	ENSG00000071539	Thyroid hormone receptor
FSHD-Hi	TF	USF1	ENSG00000158773	bHLH
FSHD-Hi	Cof	VGLL2	ENSG00000170162	Vestigial like
FSHD-Hi	Cof	WDTC1	ENSG00000142784	WD
FSHD-Hi	TF	YEATS4	ENSG00000127337	Others
FSHD-Hi	TF	ZBTB10	ENSG00000205189	ZBTB
FSHD-Hi	TF	ZBTB18	ENSG00000179456	ZBTB
FSHD-Hi	TF	ZBTB2	ENSG00000181472	ZBTB
FSHD-Hi	TF	ZBTB6	ENSG00000186130	ZBTB
FSHD-Hi	TF	ZFP3	ENSG00000180787	zf-C2H2
FSHD-Hi	TF	ZFP64	ENSG00000020256	zf-C2H2
FSHD-Hi	TF	ZFP82	ENSG00000181007	zf-C2H2
FSHD-Hi	TF	ZFP91-CNTF	ENSG00000255073	zf-C2H2
FSHD-Hi	TF	ZKSCAN2	ENSG00000155592	zf-C2H2
FSHD-Hi	TF	ZKSCAN4	ENSG00000187626	zf-C2H2
FSHD-Hi	TF	ZNF101	ENSG00000181896	zf-C2H2
FSHD-Hi	TF	ZNF132	ENSG00000131849	zf-C2H2
FSHD-Hi	TF	ZNF169	ENSG00000175787	zf-C2H2
FSHD-Hi	TF	ZNF174	ENSG00000103343	zf-C2H2
FSHD-Hi	TF	ZNF177	ENSG00000188629	zf-C2H2
FSHD-Hi	TF	ZNF20	ENSG00000132010	zf-C2H2
FSHD-Hi	TF	ZNF222	ENSG00000159885	zf-C2H2
FSHD-Hi	TF	ZNF223	ENSG00000178386	zf-C2H2
FSHD-Hi	TF	ZNF225	ENSG00000256294	zf-C2H2
FSHD-Hi	TF	ZNF227	ENSG00000131115	zf-C2H2
FSHD-Hi	TF	ZNF256	ENSG00000152454	zf-C2H2
FSHD-Hi	TF	ZNF274	ENSG00000171606	zf-C2H2
FSHD-Hi	TF	ZNF280A	ENSG00000169548	Others
FSHD-Hi	TF	ZNF280B	ENSG00000275004	Others
FSHD-Hi	TF	ZNF286A	ENSG00000187607	zf-C2H2
FSHD-Hi	TF	ZNF286B	ENSG00000249459	zf-C2H2
FSHD-Hi	TF	ZNF296	ENSG00000170684	zf-C2H2

FSHD-Hi	TF	ZNF34	ENSG00000196378	zf-C2H2
FSHD-Hi	TF	ZNF35	ENSG00000169981	zf-C2H2
FSHD-Hi	TF	ZNF350	ENSG00000256683	zf-C2H2
FSHD-Hi	TF	ZNF367	ENSG00000165244	zf-C2H2
FSHD-Hi	TF	ZNF416	ENSG00000083817	zf-C2H2
FSHD-Hi	TF	ZNF433	ENSG00000197647	zf-C2H2
FSHD-Hi	TF	ZNF439	ENSG00000171291	zf-C2H2
FSHD-Hi	TF	ZNF442	ENSG00000198342	zf-C2H2
FSHD-Hi	TF	ZNF461	ENSG00000197808	zf-C2H2
FSHD-Hi	TF	ZNF486	ENSG00000256229	zf-C2H2
FSHD-Hi	TF	ZNF534	ENSG00000198633	zf-C2H2
FSHD-Hi	TF	ZNF544	ENSG00000198131	zf-C2H2
FSHD-Hi	TF	ZNF548	ENSG00000188785	zf-C2H2
FSHD-Hi	TF	ZNF549	ENSG00000121406	zf-C2H2
FSHD-Hi	TF	ZNF551	ENSG00000204519	zf-C2H2
FSHD-Hi	TF	ZNF555	ENSG00000186300	zf-C2H2
FSHD-Hi	TF	ZNF559	ENSG00000188321	zf-C2H2
FSHD-Hi	TF	ZNF565	ENSG00000196357	zf-C2H2
FSHD-Hi	TF	ZNF586	ENSG00000083828	zf-C2H2
FSHD-Hi	TF	ZNF596	ENSG00000172748	zf-C2H2
FSHD-Hi	TF	ZNF599	ENSG00000153896	zf-C2H2
FSHD-Hi	TF	ZNF630	ENSG00000221994	zf-C2H2
FSHD-Hi	TF	ZNF639	ENSG00000121864	zf-C2H2
FSHD-Hi	TF	ZNF649	ENSG00000198093	zf-C2H2
FSHD-Hi	TF	ZNF658	ENSG00000274349	zf-C2H2
FSHD-Hi	TF	ZNF678	ENSG00000181450	zf-C2H2
FSHD-Hi	TF	ZNF684	ENSG00000117010	zf-C2H2
FSHD-Hi	TF	ZNF689	ENSG00000156853	zf-C2H2
FSHD-Hi	TF	ZNF70	ENSG00000187792	zf-C2H2
FSHD-Hi	TF	ZNF705A	ENSG00000196946	zf-C2H2
FSHD-Hi	TF	ZNF705E	ENSG00000214534	zf-C2H2
FSHD-Hi	TF	ZNF705G	ENSG00000215372	zf-C2H2
FSHD-Hi	TF	ZNF718	ENSG00000250312	zf-C2H2
FSHD-Hi	TF	ZNF730	ENSG00000183850	zf-C2H2
FSHD-Hi	TF	ZNF77	ENSG00000175691	zf-C2H2
FSHD-Hi	TF	ZNF774	ENSG00000196391	zf-C2H2
FSHD-Hi	TF	ZNF776	ENSG00000152443	zf-C2H2
FSHD-Hi	TF	ZNF788	ENSG00000214189	zf-C2H2

FSHD-Hi	TF	ZNF793	ENSG00000188227	zf-C2H2
FSHD-Hi	TF	ZNF8	ENSG00000278129	zf-C2H2
FSHD-Hi	TF	ZNF813	ENSG00000198346	zf-C2H2
FSHD-Hi	TF	ZNF829	ENSG00000185869	zf-C2H2
FSHD-Hi	TF	ZNF844	ENSG00000223547	zf-C2H2
FSHD-Hi	TF	ZNF852	ENSG00000178917	zf-C2H2
FSHD-Hi	TF	ZNF879	ENSG00000234284	zf-C2H2
FSHD-Hi	TF	ZNF880	ENSG00000221923	zf-C2H2
FSHD-Hi	TF	ZNF888	ENSG00000213793	zf-C2H2
FSHD-Hi	TF	ZNF92	ENSG00000146757	zf-C2H2
FSHD-Hi	TF	ZSCAN12	ENSG00000158691	zf-C2H2
FSHD-Hi	TF	ZSCAN2	ENSG00000176371	zf-C2H2
FSHD-Hi	TF	ZSCAN31	ENSG00000235109	zf-C2H2
FSHD-Hi	TF	ZSCAN4	ENSG00000180532	zf-C2H2
FSHD-Hi	TF	ZSCAN5B	ENSG00000197213	zf-C2H2
FSHD-Hi	TF	ZSCAN5C	ENSG00000204532	zf-C2H2
FSHD-Hi	TF	ZXDB	ENSG00000198455	zf-C2H2

2.8 References

1. Tawil R, Van Der Maarel SM. Facioscapulohumeral muscular dystrophy. Muscle Nerve. 2006 Jul 1;34(1):1–15.
2. Zeng W, Chen YY, Newkirk DA, Wu B, Balog J, Kong X, et al. Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs. Hum Mutat. 2014;35(8):998–1010.
3. Young JM, Whiddon JL, Yao Z, Kasinathan B, Snider L, Geng LN, et al. DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. PLoS Genet. 2013 Nov;9(11).
4. Geng LN, Yao Z, Snider L, Fong AP, Cech JN, Young JM, et al. DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy. Dev Cell. 2012 Jan 17;22(1):38–51.
5. Lemmers RJLF, Tawil R, Petek LM, Balog J, Block GJ, Santen GWE, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat Genet. 2012 Dec;44(12):1370–4.
6. Sacconi S, Lemmers RJLF, Balog J, Van Der Vliet PJ, Lahaut P, Van Nieuwenhuizen MP, et al. The FSHD2 gene SMCHD1 Is a modifier of disease severity in families affected by FSHD1. Am J Hum Genet. 2013 Oct 3;93(4):744–51.
7. Larsen M, Rost S, El Hajj N, Ferbert A, Deschauer M, Walter MC, et al. Diagnostic approach for FSHD revisited: SMCHD1 mutations cause FSHD2 and act as modifiers of disease severity in FSHD1. Eur J Hum Genet. 2015 Jun 5;23(6):808–16.
8. Snider L, Geng LN, Lemmers RJLF, Kyba M, Ware CB, Nelson AM, et al. Facioscapulohumeral Dystrophy: Incomplete Suppression of a Retrotransposed Gene. Pearson CE, editor. PLoS Genet. 2010 Oct 28;6(10):e1001181.
9. Lemmers RJLF, van der Vliet PJ, Klooster R, Sacconi S, Camaño P, Dauwerse JG, et al. A unifying genetic model for facioscapulohumeral muscular dystrophy. Science. 2010 Sep 24;329(5999):1650–3.
10. Himeda CL, Jones TI, Jones PL. Facioscapulohumeral muscular dystrophy as a model for epigenetic regulation and disease. Antioxid Redox Signal. 2015 Jun 1;22(16):1463–82.
11. De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D. DUX-family transcription factors regulate zygotic genome activation in placental mammals. Nat Genet. 2017 Jun 1;49(6):941–5.
12. Hendrickson PG, Doráis JA, Grow EJ, Whiddon JL, Lim JW, Wike CL, et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat Genet. 2017 Jun 1;49(6):925–34.
13. Whiddon JL, Langford AT, Wong CJ, Zhong JW, Tapscott SJ. Conservation and innovation in the DUX4-family gene network. Nat Genet. 2017 Jun 1;49(6):935–40.
14. Bosnakovski D, Xu Z, Gang EJ, Galindo CL, Liu M, Simsek T, et al. An isogenetic myoblast expression screen identifies DUX4-mediated FSHD-associated molecular pathologies. EMBO J. 2008 Oct 22;27(20):2766–79.
15. Vanderplanck C, Ansseau E, Charron S, Stricwant N, Tassin A, Laoudj-Chenivesse D, et al. The FSHD Atrophic Myotube Phenotype Is Caused by DUX4 Expression. Chadwick BP, editor. PLoS One. 2011 Oct 28;6(10):e26820.
16. Tassin A, Laoudj-Chenivesse D, Vanderplanck C, Barro M, Charron S, Ansseau E, et al. DUX4 expression in FSHD muscle cells: How could such a rare protein cause a myopathy? J Cell Mol Med. 2013 Jan;17(1):76–89.

17. Zeng W, De Greef JC, Chen YY, Chien R, Kong X, Gregson HC, et al. Specific loss of histone H3 lysine 9 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). PLoS Genet. 2009 Jul;5(7).

18. Van Overveld PGM, Lemmers RJFL, Sandkuijl LA, Enthoven L, Winokur ST, Bakels F, et al. Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. Nat Genet. 2003 Dec;35(4):315–7.

19. Jansz N, Chen K, Murphy JM, Blewitt ME. The Epigenetic Regulator SMCHD1 in Development and Disease. Vol. 33, Trends in Genetics. Elsevier Ltd; 2017. p. 233–43.

20. Yao Z, Snider L, Balog J, Lemmers RJLF, Van Der Maarel SM, Tawil R, et al. DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. Hum Mol Genet. 2014 Oct 15;23(20):5342–52.

21. Zeng W, Jiang S, Kong X, El-Ali N, Ball AR, Ma CIH, et al. Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. Nucleic Acids Res. 2016 Dec 1;44(21).

22. Rickard AM, Petek LM, Miller DG. Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. Hum Mol Genet. 2015 Jun 5;24(20):5901–14.

23. Conesa A, Nueda M. maSigPro: Significant Gene Expression Profile Differences in Time Course Gene Expression Data. 2017.

24. Jagannathan S, Shadle SC, Resnick R, Snider L, Tawil RN, van der Maarel SM, et al. Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. Hum Mol Genet. 2016 Aug 17;ddw271.

25. Leidenroth A, Hewitt JE. A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol Biol. 2010 Nov 26;10(1):364.

26. Banerji CRS, Panamarova M, Pruller J, Figeac N, Hebaishi H, Fidanis E, et al. Dynamic transcriptomic analysis reveals suppression of PGC1α/ERRα drives perturbed myogenesis in facioscapulohumeral muscular dystrophy. Hum Mol Genet. 2019;28(8).

27. Resnick R, Wong C-J, Hamm DC, Bennett SR, Skene PJ, Hake SB, et al. DUX4-Induced Histone Variants H3.X and H3.Y Mark DUX4 Target Genes for Expression. Cell Rep. 2019 Nov 12;29(7):1812-1820.e5.

28. Knopp P, Krom YD, Banerji CRS, Panamarova M, Moyle LA, den Hamer B, et al. DUX4 induces a transcriptome more characteristic of a less-differentiated cell state and inhibits myogenesis. J Cell Sci. 2016 Oct 15;129(20):3816–31.

29. van den Heuvel A, Mahfouz A, Kloet SL, Balog J, van Engelen BGM, Tawil R, et al. Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. Hum Mol Genet. 2018 Nov 16;

30. Wallace LM, Garwick SE, Mei W, Belayew A, Coppee F, Ladner KJ, et al. *DUX4* , a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo. Ann Neurol. 2011 Mar 1;69(3):540–52.

31. Sadasivam S, DeCaprio JA. The DREAM complex: master coordinator of cell cycle-dependent gene expression. Nat Rev Cancer. 2013 Aug 11;13(8):585–95.

32. Kulakovskiy I V, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018 Jan 4;46(D1):D252–9.

33. Saunders A, Huang X, Fidalgo M, Reimer MH, Faiola F, Ding J, et al. The SIN3A/HDAC Corepressor Complex Functionally Cooperates with NANOG to Promote Pluripotency. Cell Rep. 2017 Feb 14;18(7):1713–26.

34. Campbell AE, Shadle SC, Jagannathan S, Lim J-W, Resnick R, Tawil R, et al. NuRD and CAF-1-mediated silencing of the D4Z4 array is modulated by DUX4-induced MBD3L proteins. Elife. 2018 Mar 13;7.

35. Fleming JD, Pavesi G, Benatti P, Imbriano C, Mantovani R, Struhl K. NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. Genome Res. 2013 Aug;23(8):1195–209.

36. Dubrez L. Regulation of E2F1 transcription factor by ubiquitin conjugation. Int J Mol Sci. 2017;18(10):1–9.

37. Feng Q, Snider L, Jagannathan S, Tawil R, van der Maarel SM, Tapscott SJ, et al. A feedback loop between nonsense-mediated decay and the retrogene DUX4 in facioscapulohumeral muscular dystrophy. Elife. 2015 Jan 7;2015(4).

38. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171–81.

39. Library P, Data A. Illumina Bio-Rad SureCell ^{TM} WTA 3 ′ Library Prep Kit for the ddSEQ ^{TM} System. 2017;(Pub. No. 1070-2016-014-C):5–8.

40. Kong X, Mohanty SK, Stephens J, Heale JT, Gomez-Godinez V, Shi LZ, et al. Comparative analysis of different laser systems to study cellular responses to DNA damage in mammalian cells. Nucleic Acids Res. 2009 May 1;37(9):e68–e68.

41. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

42. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Dec 4;12(1):323.

43. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

44. Romagnoli D, Boccalini G, Bonechi M, Biagioni C, Fassan P, Bertorelli R, et al. ddSeeker: a tool for processing Bio-Rad ddSEQ single cell RNA-seq data. BMC Genomics. 2018 Dec 24;19(1):960.

45. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016 Sep 15;32(18):2847–9.

46. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 May 2;36(5):411–20.

47. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. bioRxiv. 2019 Mar 18;576827.

48. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019 Dec 3;10(1):1523.

49. Jawaid W. enrichR: Provides an R Interface to “Enrichr.” R package version 2.1. 2019.

50. Hu H, Miao Y-R, Jia L-H, Yu Q-Y, Zhang Q, Guo A-Y. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019 Jan 8;47(D1):D33–8.

51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov 1;13(11):2498–504.
52. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28;38(4):576–89.
53. Altschul SF, Wootton JC, Zaslavsky E, Yu Y-K. The Construction and Use of Log-Odds Substitution Scores for Multiple Sequence Alignment. Siepel A, editor. PLoS Comput Biol. 2010 Jul 15;6(7):e1000852.

CHAPTER 3

Muscle group specific transcriptomic and DNA methylation differences related to developmental patterning in FSHD

Notes: (1) Dr. Xiangduo Kong cultured cells and reviewed this work.
(2) Nam Viet Nguyen contributed to the review of this work.
(3) Cassandra McGill performed the sequencing for this work.
(4) Dr. Rabi Tawil provided the resources for this work.
(5) The Wellstone Center also provided resources for this work.
(6) Professor Kyoko Yokomori and Professor Ali Mortazavi contributed to the resources, supervision, writing and review of this work.

Chapter 3
Muscle group specific transcriptomic and DNA methylation differences related to developmental patterning in FSHD

3.1 Abstract

Muscle groups throughout the body are specialized in function and are specified during development by position specific gene regulatory networks. In developed tissue, myopathies affect muscle groups differently. Facioscapulohumeral muscular dystrophy, FSHD, affects upper body and tibialis anterior (TA) muscles earlier and more severely than others such as quadriceps. To investigate an epigenetic basis for susceptibility of certain muscle groups to disease, we perform DNA methylation and RNA sequencing on primary patient derived myoblasts from TA and quadricep for both control and FSHD2 as well as RNA-seq for myoblasts from FSHD1 deltoid, bicep and TA over a time course of differentiation. We find that TA and quadricep retain methylation and expression differences in transcription factors that are key to muscle group specification during embryogenesis. FSHD2 patients have differences in DNA methylation and expression related to SMCHD1 mutations and FGF signaling. Genes induced specifically in FSHD are more highly expressed in commonly affected muscle groups. We find a set of genes that distinguish more susceptible muscle groups including development-associated TFs and genes involved in WNT signaling. Adult muscle groups therefore retain transcriptional and DNA methylation differences associated with development, which may contribute to susceptibility in FSHD.

3.2 Background

The human body has over 650 named skeletal muscle groups [1,2]. These groups are heterogeneous in terms of compositions of fast and slow twitch fibers and regenerative capabilities [3]. Initial muscle cell specification through the activation of the key myogenic factors MRF4/MYF5, MYOD, and MYOG is regulated by upstream transcription factors that vary depending on the location of the muscle cells [4,5]. PITX2 plays a central role in regulating MYF5/MRF4 in extraocular muscles and can regulate MYOD in limb muscles [4–7]. SIX and EYA TFs activate PAX3 and MYF5/MRF4 in limb muscles [7].

The limbs are initially specified by HOX genes which activate TBX5 in the forelimb and PITX1 in the hindlimb [8]. PITX1 then activates TBX4 in the hindlimb [8]. TBX5 and TBX4 activate FGF10 which forms a gradient with FGF8 that is expressed at the apical ectodermal ridge to control limb outgrowth [8]. Limb outgrowth specifies the proximal/distal axis resulting in three major regions; the stylopod, zeugopod and autopod [8]. MEIS factors are expressed proximally in the stylopod but are repressed by SHOX2 distally into the zeugopod where HOXA11 is expressed [9,10]. Dorsal/ventral patterning is controlled by WNT and LMX1B expression on the dorsal side and EN1 on the ventral side which represses WNT [11].

Gene expression in embryonic development of different muscle groups is well studied, but few studies have surveyed gene expression in adult tissue [12,13]. An estimated 50% of transcripts are differentially expressed in adult muscle with some heterogeneity coming from fiber-type composition and expression of developmental related genes [14]. In addition to transcription, adult muscle groups retain DNA methylation differences [15]. The molecular differences between muscle groups may contribute to severity of affectedness in myopathies.

Many myopathies affect muscle groups of the body exclusively or more severely, such as facioscapulohumeral muscular dystrophy (FSHD), which is characterized by noticeable weakness in the most commonly affected muscle groups in the upper body including facial and humeral [16,17]. FSHD progression into muscle groups is sporadic, but some groups such as the tibialis anterior are more commonly affected or affected earlier [18]. Certain muscle groups are less frequently affected or affected later including the quadricep and the deltoid [19,20].

FSHD is caused by the misexpression of the embryonic transcription factor *DUX4* in skeletal muscle [17,19]. DUX4 activates expression of its target genes including a number of embryonic related transcription factors and chromatin remodelers as well as repeat elements including endogenous retroviruses such as ERVLs [21–24]. DUX4 target gene expression correlates with signs of active disease [25,26]. Comparison of the commonly affected bicep to the less affected deltoid found greater expression differences between FSHD and control in bicep than deltoid but did not assess differences in DUX4 target gene expression between the two muscle groups [27].

We survey transcription and DNA methylation to establish differences between muscle groups that may contribute to susceptibility in FSHD. We perform RNA-seq and probe enriched bisulfite sequencing to survey transcriptional and DNA methylation differences between tibialis anterior (TA) and quadricep in myoblasts from healthy patients and those with FSHD2 with SMCHD1 mutations. We examine differences in DNA methylation and expression between TA and quadricep, and find retained differences in TFs important for development. Next, we find a set of genes specifically upregulated in FSHD2 over time including DUX4 target genes. Notably, the promoters of these genes are highly methylated in both FSHD2 and control cells. We look at genome-wide DNA methylation and gene expression differences in FSHD2 and find hypomethylation and increased expression for genes in regions regulated by SMCHD1. We survey differences for TE loci and find repeat elements upregulated in response to DUX4 are especially upregulated in FSHD2 cells from TA but not quadricep. To determine possible differences in gene expression correlating with susceptibility, we performed RNA-seq for myoblasts from the TA, bicep and deltoid of FSHD1 patients. We identify muscle group specific gene expression for TA, quadricep, bicep and deltoid, and genes differentially expressed between muscle groups with different susceptibilities to FSHD.

3.3 Results

3.3.1 Muscle group specific DNA methylation and gene expression in developmental TFs

To identify epigenetic differences between muscle groups, we performed capture enrichment bisulfite sequencing (Methods) on two primary myoblast cell lines from quadricep and two from tibialis anterior (TA) for days 0, 3 and 12 of differentiation (Figure S3.1). We merged CpG sites within 200 bp into regions which were then filtered for coverage (Methods). We found very few differentially methylated regions (DMRs) (percent diff 25, q-value <0.01) between the differentiation days and decided to treat the different days as replicates for further comparisons (Figure S3.2A, S3.2B).

We compared TA and quadricep and found 1081 regions more highly methylated in quadricep as well as 686 regions with higher methylation in TA (Figure 3.1A, S3.2B). The regions with the highest percent methylation differences are associated with transcription factors that play a role in limb specific muscle specification. These include *PITX2,* which controls location specific gene networks to induce MRFs [4]. Also included is *MEIS1,* which is expressed in the stylopod and is more highly methylated in TA than quad (Figure 3.1A) [9]. Of the 1767 DMRs, 357 are associated with 265 TFs, and 83 of those are involved in pattern specification while 39 are involved in muscle structure development (Figure S3.2C). In support of this, gene ontology analysis for these DMRs revealed significant association with development and structure morphogenesis (Figure 3.1B). Quadricep and TA therefore retain DNA methylation patterns from their specification.

We looked for TF motifs enriched in the DMRs between TA and quadricep to see what could be regulating these regions. The DMRs are enriched for motifs of developmental transcription factors (Figure S3.2D). Importantly, we found motifs for key transcription factors involved in axis specification, skeletal muscle development and differentiation, and limb specification and morphogenesis, including motifs for 19 TFs controlling anterior/posterior, dorsal/ventral and/or proximal/distal pattern specification; six HOXA (HOXA4, 5, 6, 7, 9, 10), five HOXB (HOXB2, 3, 5, 6, 8), two HOXC (HOXC4, 10), two HOXD (HOXD8, 11), as well as CDX1, LMX1B, LHX1, and BARX1 (Figure 3.1C, S3.2D). Seven motifs are from TFs important for limb specification, patterning or development, and seven motifs are from TFs involved in muscle development or differentiation (Figure 3.1C, S3.2D). MEOX2 is involved in both muscle and limb development when it acts in concert with PAX3 and SIX1/4 to activate *MYF5* in myoblasts migrating into the limb [28].

Having identified epigenetic differences in developmental TFs, we looked to see if the TFs are differentially expressed between the tissues. We performed RNA-seq on days 0 to 5 and day 12 of differentiation for the two quadricep and two TA cell lines (Fig S1A). We found 867 genes differentially expressed (|logFC| >1, FDR <0.01) including 54 TFs (Figure 3.1D). Of these TFs, 26 are related to axis patterning, 16 are involved in appendage development, and 12 are involved in muscle organ development (Figure S3.2E). The hindlimb determining TF *PITX1* is more highly expressed in TA than quadricep (Figure 3.1D) [29]. PITX1 activates *TBX4*, which was differentially methylated, in the developing hindlimb [8]. *TBX5*, which is responsible for specifying the forelimb, is more highly expressed in the quadricep than the TA (Figure 3.1D) [8]. Five of the differentially expressed TFs also had motifs which were significantly enriched in the differentially methylated regions; CDX1, MEOX2, HOXD8, FOXC1 and FOXD2. CDX1 is expressed in the limb bud and is responsive to retinoic acid (RA), WNT and FGF signals [30,31]. HOXD8 is expressed during early patterning in the proximal limb [8].

We compared the differences in DNA methylation with changes in gene expression by associating the differentially methylated regions with genes using GREAT (Methods) [32]. Seventeen of the TFs were both differentially expressed and differentially methylated (Figure 3.1E). This included *EN1*, which represses *WNT7A* in dorsal/ventral patterning, and *SALL4*, which regulates hindlimb outgrowth with TBX4 by regulating and *FGF10* [11,33]. Differences in methylation around *EN1* included proximal regions and the gene body that overlap annotated candidate cis-regulatory elements (Figure 3.1F). We compared the differentially expressed genes in TA and quadricep from our primary myoblasts from control and FSHD patients to those from RNA-seq from TA and quadricep patient biopsy samples [25]. *PITX1* and *IRX5* were upregulated in TA in both cell lines and biopsy samples for FSHD and control (Figure S3.2F). IRX5 expression in the hindlimb bud is important for specifying proximal and anterior regions of the limb [32]. Thus, muscle groups retain DNA methylation and expression differences in adult tissue based on early patterning.

3.3.2 Genes induced upon myogenesis in FSHD are more highly expressed in more susceptible muscle

We previously identified a set of genes induced in FSHD2 cells upon differentiation up to day 5 [53]. To assess FSHD specific gene upregulation in different muscle groups, we performed

RNA-seq for both TA and quadricep from SMCHD1 mutated FSHD2. We found a set of 74 genes that are upregulated in both TA and quadricep FSHD2 cell lines starting around day 3 of differentiation (Figure 3.2A, 3.2B). This includes 47 of our original 54 FSHD-induced genes and an additional 27 genes (Figure 3.2B, S3.3A). Some of the new genes include previously identified DUX4 target genes such as *PRAMEF10* (Figure S3.2A) [21]. These additional genes appear to be more lowly expressed than the set we identified previously (Figure S3.2B). Interestingly, the quadricep, which is less susceptible to FSHD, has lower expression of these FSHD-induced genes (Figure 3.2A). However, when comparing FSHD biopsy samples from TA and quadricep, these genes are not significantly higher in the TA (Figure S3.3C, S3.3D).

We wanted to know whether differences in DNA methylation near these FSHD-induced genes could partially explain their strong upregulation. We summarized the promoter CpG sites for the 74 genes and recovered 53 gene promoters passing our coverage filters (Methods). The promoters of these genes are not significantly differentially methylated despite substantial increases in expression (Figure 3.2C). In fact, most of the promoters (33 out of 53) are highly methylated with greater than 50% methylation in both FSHD2 and control cell lines (Figure S3.3E). Hypermethylation in the promoter is generally associated with repressing gene expression, so we decided to determine whether DUX4 binds these promoter regions or could be regulating expression through binding of other regulatory regions. Eleven of the highly methylated FSHD-induced gene promoters overlap DUX4 binding sites determined by ChIP-seq (Figure S3.2D) [24].

To determine what TFs could be binding the methylated FSHD-induced gene promoters, we looked for enrichment of unmethylated and methylated TF motifs determined by SELEX [35]. While a methylated motif is not available for DUX4, the methylated motif for its paralog, DUXA, is highly enriched in 41 of the 74 promoters (Figure S3.3F, Table S3.1). Additionally, the methylated motif for OTX1, a PRD-like TF, is also enriched (Table S3.1). OTX1 has a similar motif to the DUX4 target TF LEUTX [36]. Performing enrichment using canonical motifs only, motifs for DUX4, DUXA, and OTX2, another PRD-like TF similar to LEUTX, are all enriched in the target promoters (Figure S3.3F, Table S3.1). The MEOX2 motif is also enriched, and *MEOX2* expression was significantly higher in quadricep than in TA (Figure 3.2A, Table S3.1).

3.3.3 SMCHD1 mutation associated differences in DNA methylation and gene expression

To assess global methylation differences between SMCHD1 mutated FSHD2 and control cell lines, we compared all FSHD2 samples to all control samples and found 4527 regions more highly methylated in control and 3542 in FSHD2 (percent diff >10, qvalue <0.01) (Figure 3.3A). The top most differentially methylated region is *DBET,* which is close to *FRG2* and *DUX4* in the D4Z4 region on chromosome 4 that is hypomethylated in FSHD patients (Figure 3.3A) [37]. *DBET* is also less methylated in TA than quadricep in FSHD2 (Figure S3.4A). The D4Z4 region on chromosome 10 near *FRG2B* and *SYCE1* contains two hypomethylated regions in FSHD2, and *SYCE1* expression is upregulated in FSHD2 (Figure 3.3A, 3.3B) [38]. Hypomethylation in the D4Z4 region on chromosome 10 has been shown to be FSHD2 specific, and the upregulation of *SYCE1* that we observe is not seen in FSHD1 (Figure S3.4B). SMCHD1 has been shown to regulate clusters of genes such as the PCDH cluster on chromosome 5 (chr5:140,759,009-141,523,383, hg38), the SNRPN cluster (chr15:23,548,232-23,697,319, hg38), the rRNA cluster (chr1:228,552,374-228,653,525, hg38), and the tRNA cluster (chr1:161,395,860-161,624,746, hg38) [41,42]. Out of 142 regions in the PCDH cluster and 6 in the rRNA cluster, we find 27 regions and 3 regions, respectively, with hypomethylation in FSHD2 compared to control (Figure 3.3A). The SNRPN and tRNA clusters have negligible methylation differences (Figure 3.3A). The hypomethylation observed in the PCDH and rRNA clusters is low (less than 25%) suggesting differences in methylation due to SMCHD1 heterozygosity are mild, especially when compared with methylation differences near D4Z4 on chromosomes 4 and 10.

We looked at expression of genes in the PCDH and SNRPN clusters and found slight upregulation in FSHD2 samples (Figure 3.3C). Despite not having significant methylation differences, *NDN* and *MAGEL2* from the SNRPN cluster both have slightly elevated expression in FSHD2 (Figure 3.3C). Interestingly, *NDN* expression is much higher in FSHD2 cells from the more susceptible muscle group TA (Figure 3.3C). Three genes from the PCDH cluster were upregulated in FSHD2 cells consistent with the hypomethylation in this region (Figure 3.3C). *SMCHD1* expression is slightly lower in FSHD2 samples than control, which is surprising since SMCHD1 mutations in these FSHD2 patients affect protein function, not expression (Figure 3.3C).

We assessed global expression differences between FSHD2 and control combining all differentiation days. The most highly upregulated genes in FSHD2 include the FSHD-induced genes previously identified (Figure 3.3B). The most significantly downregulated gene in FSHD2 is *FLVCR1-AS1*, a lncRNA, which can act as a sponge for miRNAs that inhibit proliferation (Figure 3.3B) [43]. The genes for signaling molecules *FGF7*, *FGF10* and *TGFA* are also down regulated in FSHD2 compared to control (Figure 3.3B). Using gene ontology analysis, the differentially methylated regions were identified as being associated with genes involved in histone modifications, including H3K4 methylation and H2A ubiquitination, and FGF signaling (Figure 3.3D).

3.3.4 LTR are loci highly expressed in muscle that is more susceptible to FSHD

DUX4 activates expression of transposable elements (TEs), and we therefore wanted to assess the extent of TE upregulation in our samples [24,44]. In FSHD2 cells, 580 repeat loci comprised of 175 different TE types are upregulated in a time specific manner starting at day 3 of differentiation (Figure 3.4A, 3.4B). Of these 175 types, 79 are LTRs including significant enrichment of ERVL-MaLR (p=4.9e-92), ERVL (p=2.6e-17) and ERV1 (p=2.7e-5) (Figure 3.4C). Specific repeat types, such as THE1D, MLT2A1 and THE1C, which are regulated by DUX4, are also significantly enriched and upregulated (Figure S3.5A) [24]. These loci are substantially upregulated in FSHD2 TA and show only slight upregulation in quadricep at day 12 of differentiation (Figure 3.4A).

These upregulated TEs significantly overlap (p=2.3e-70) DUX4 binding sites identified by ChIP-seq (Figure 3.4B) [24]. Out of 104 loci that overlap DUX4 binding sites, 99 are LTRs (p=2.1e-28) including 79 ERVL-MaLRs (p=5.7e-24) (Figure 3.4B, 3.4C). To determine whether the TEs are upregulated because they overlap FSHD-induced genes, we intersected the TE loci with the FSHD-induced genes. Thirty-six TE loci overlap the FSHD-induced genes, only 6 of which are LTRs (Figure 3.4B). A previous study found that LTRs bound by DUX4 could act as alternative promoters for protein coding and lncRNAs [24]. We find that 20 out of the 79 LTRs upregulated in FSHD2 overlap LTRs identified as alternative promoters, including 15 for protein coding genes such as *MLT1E1A-NT5C1B* and 5 for lncRNAs.

We summarized CpG sites over LTR loci to determine whether DUX4-regulated LTRs are differentially methylated (Methods). Between FSHD2 and control, 45 LTRs are more highly methylated (percent diff >25%, qvalue <0.01) in FSHD2, and 36 are more highly methylated in control (Figure 3.4D). These loci were not in the set identified as upregulated in FSHD2. Loci with more methylation in FSHD2 are enriched for ERVL (p=0.01) and Gypsy (p=0.04) elements (Figure 3.4C). FSHD2 hypomethylated TEs are enriched for ERVL-MaLR (p=0.03) which are enriched in the set of TEs upregulated in FSHD2 (Figure 3.4C). Additionally, two ERVL-MaLR loci overlap DUX4 binding sites (Figure 3.4D).

3.3.5 Muscle group specific gene expression in FSHD

To identify transcriptional differences between muscle groups that may cause or result from differential susceptibility to FSHD, we performed RNA-seq on FSHD1 patient matched deltoid and bicep derived cell lines for days 0 and 5 of differentiation as well as from TA for day 0 (Figure S3.1). We identified muscle specific expression patterns by performing pairwise comparisons between the muscle groups at day 0 and taking the intersect of genes specific for one muscle group (Methods). A total of 582 unique genes were specific to a muscle group with either especially high or low expression for the group (Figure 3.5A). Bicep specifically expresses *LHX8* which can regulate *SHH* expression in development of various tissues in the upper body (Figure 3.5B) [45]. Deltoid expresses *HOXD3* which is involved in limb outgrowth (Figure 3.5B) [46]. TA has higher expression of *EYA1* which is involved in activating PAX3 and MYF5 in limb muscles (Figure 3.5B) [5]. Quadricep has high expression of *TLL1* which is involved in dorsal ventral patterning and can indirectly regulate myostatin (Figure 3.5B) [47,48]. Interestingly, we see the hindlimb specific TF *TBX5* is high in the bicep, deltoid and quadricep but not the tibialis anterior (Figure S3.6A).

FSHD tends to affect upper musculature first, so we compared the two forelimb muscles (bicep and deltoid) to the two hindlimb ones (quadricep and TA). Genes in the 5' end of the

HOXC cluster, including *HOXC9* through *13* as well as *HOXC-AS1*, *HOXC-AS5* and *HOTAIR*, were all more highly expressed on average in the hindlimb than the forelimb (Figure S3.6B, S3.6C). Genes in the 3' end of the HOXD cluster are more highly expressed in the forelimb (Figure S3.6C).

We then compared expression between the commonly and less affected muscle groups in the forelimb and hindlimb separately to investigate differences between muscle groups more susceptible to FSHD. As noted previously, FSHD-induced genes are more highly expressed in TA than quadricep, 71 out of 74 at day 5 of differentiation (Figure 3.5C). FSHD-induced genes are also more highly expressed in bicep, which is more commonly affected, than deltoid (67 out of 74) (Figure 3.5D). Higher expression of DUX4 target genes has been shown previously to correlate with active signs of disease [25]. Here we show that the FSHD gene signature is also more highly expressed in more susceptible tissue.

To find genes specific to commonly affected groups, we took the intersect of the genes differentially expressed in bicep compared to deltoid and TA compared to quadricep in the same direction (Methods). This yields 28 genes more highly expressed in commonly affected groups, and 27 higher in the less affected groups including 7 TFs total (Figure S3.6D). The myogenic precursor TF *PAX3* and *HOXA11* are more highly expressed in more susceptible groups (Figure 3.5E). TFs more highly expressed in the less susceptible muscles include *SHOX2* and *MEIS1*. *MEIS1* and *SHOX2* are expressed in the proximal region of the limb [8,9,49]. *PAX3* marks myogenic precursor cells and may play a role in satellite cells postnatally [5]. *HOXA11* is expressed in the zeugopod in development [8,10]. We also see higher expression of genes involved in WNT signaling including *Noggin* (Figure 3.5E, S3.6C). Also higher in more affected tissues is *HEYL* which is involved in NOTCH signaling, represses MYOD expression, and is required for satellite cell proliferation in a model of hypertrophy (Figure S3.6C) [50,51]. In conclusion, we identified DNA methylation and transcriptional differences between muscle groups in cell lines derived from adult tissue.

3.4 Discussion

We identified developmental transcription factors with differences in DNA methylation and expression in cells derived from adult tissue. By examining cells from FSHD patients, we found differences between muscle groups that are related to disease, including *PITX1*, emphasizing the consideration of sample origin in transcriptome studies in FSHD. Genes induced specifically in FSHD are not differentially methylated, but genes regulated by SMCHD1 are hypomethylated and slightly overexpressed in SMCHD1 mutated FSHD2 patient cells. Importantly, we find a set of genes that possibly correlate with muscle group susceptibility to FSHD.

The differential DNA methylation and gene expression between muscle groups supports a potential role for developmental transcription factors in adult tissue. Previous work in mice showed transcriptional and methylation differences between extraocular muscles (EOM) and TA and that transplantation could alter the location specific transcriptional profiles [15]. Notably, EOM cells, which do not express HOX genes, upregulate TA specific HOX genes upon transplantation supporting a role for environmental stimuli in controlling location-associated TFs [15]. Notably, *PITX2* expression in EOM cells was resistant to change after transplantation to the TA, and PITX2 can increase the ability of satellite cells to regenerate [4,15]. We find substantial methylation differences in PITX2 between TA and quadricep. The role of some of these other TFs in adult tissue warrants further investigation.

The transcriptional differences that we observe between the bicep and TA compared with the deltoid and quadricep may be the result of inherent differences between these muscle groups or due specifically to disease. While the differences that we observed between TA and quadricep could be due to individual differences, our bicep and deltoid samples were taken from the same individuals. Comparison of these muscle groups from healthy individuals could identify whether differences between the susceptible groups are due to inherent differences, active disease or differences inherent to FSHD patients. The higher expression of FSHD-induced genes in the more commonly affected muscle groups may be due to either active disease or inherent differences between the groups. Indeed, FSHD-related gene expression has been shown to be higher in tissue with signs of active disease [25,26]. Differences in expression of homeobox transcription factors between commonly and less affected groups suggests a possible link to the governing of muscle groups affected.

A previous study of DNA methylation in FSHD and control patients found relatively little differential methylation compared to what we observe, which may be the result of increased sample size or to the SMCHD1 mutations in the FSHD2 patients [52]. The enrichment in hypomethylation of ERVL-MaLRs suggests a mechanism for their increased expression in FSHD. The differentially methylated ERVL-MaLRs do not significantly overlap DUX4 binding sites, which suggests that regulation of DNA methylation in those regions is independent of DUX4 binding but could possibly be attributed to SMCHD1 mutations. In addition, we noted differences in methylation related to FGF signaling and downregulation of two *FGFs* and *TGF*α in FSHD2 cells. FGF10 is an important contributor to limb outgrowth but is largely unexplored in FSHD. FGF1 and FGF2 were found in a biopsy of one patient with a severe phenotype [53].

We found *NOG* more highly expressed in more susceptible muscle groups in FSHD. Noggin inhibits multiple BMPs including BMP4 [54]. BMP4 can activate *FGF7* and *FGF10* expression [55].

The hypermethylation and lack of difference in the methylation of the FSHD-induced gene promoters could indicate several possibilities. First, DUX4 most likely binds other regulatory elements outside of these promoter regions, such as enhancers, such that highly methylated promoters are not prohibitive to activation by DUX4. Second, DUX4 and/or other binding partners may not be sensitive to DNA methylation. Indeed, homeobox transcription factors preferentially bind methylated DNA, and we find the methylated motif for DUXA enriched at the promoters of FSHD-induced genes [35]. Additionally, complexes that are indirectly called to methylated DNA, such as SIN3, regulate *DUX4* expression, and their targets are affected in DUX4-affected cells [56–59]. DUX4 also upregulates the *MBD3L* genes which are methyl binding domain proteins, and MBD3L2 is known to bind methylated DNA [60,61]. Third, demethylated promoters may only be present in the subset of nuclei that activate the DUX4 program. As affected FSHD nuclei make up a small percentage of total nuclei, differences in those nuclei would only be observable with single nucleus methylation assays [23,57].

In summary, we identified transcriptional and DNA methylation differences between muscle groups from adult tissue of healthy individuals and FSHD patients. We identified a set of genes potentially linked to susceptibility or progression of affected muscle groups in FSHD, including genes specifically upregulated in FSHD across myogenesis. Understanding the genomic basis of susceptibility is important for identifying key mechanisms of FSHD progression.

3.5 Methods

3.5.1 Human myoblast culture and differentiation

Human control, FSHD1 and FSHD2 myoblast cells from patient quadricep, tibialis anterior, bicep and deltoid biopsies were grown as previously described [57]. Cells were cultured on dishes coated with collagen in high glucose DMEM (Gibco) supplemented with 20% FBS (Omega Scientific, Inc.), 1% Pen-Strep (Gibco), and 2% Ultrasor G (Crescent Chemical Co.). Day 0 cells were kept at low confluency to prevent spontaneous differentiation. Upon reaching 80% confluence, differentiation was induced by using high glucose DMEM medium supplemented with 2% FBS and ITS supplement (insulin 0.1%, 0.000067% sodium selenite, 0.055% transferrin; Invitrogen). Fresh differentiation medium was changed every 24 hours.

3.5.2 DNA methylation library preparation

Genomic DNA was collected from a single 6 cm dish using the DNEasy Blood & Tissue kit (69504, Qiagen). For both reps of day 0, day 3 and rep 1 of day 12 for Control-4, gDNA from two plates were pooled and concentrated using DNA Clean & Concentrator-25 (D4033, Zymo) to obtain high enough input and concentration. All DNA methylation data was generated using the TruSeq Methyl Capture EPIC kit (FC-151-1003, Illumina) according to manufacturer's protocol [62]. Libraries were sequenced on the Illumina NextSeq500 with paired-end 75 bp reads to a depth of 20 to 82 million reads.

3.5.3 DNA methylation data processing

Raw reads from TruSeq Methyl Capture EPIC libraries were mapped to canonical chromosomes from hg38 and the patch region of D4Z4 (chr4_KQ983257v1_fix) using bismark (version 0.19.0) [63]. Sites were extracted from the bam file using bismark_methylation_extractor with paired end and no overlap specified. To remove bias from the ends of reads, 2 bp from the 5' end of read 2 and 1 bp from the 3' end of read 1 were ignored when extracting sites. All CpG sites were read into methylKit (version 1.16.0) [64]. Sites within 200 bp were merged into one region using bumphunter (version 1.32.0) [65]. Methylation over those regions was summarized using regionCounts and filtered for at least three CpG sites. Regions were then filtered for a coverage of at least 5. Regions were normalized using normalizeCoverage, and only regions with coverage in all samples were kept. Differential methylation was calculated using calculateDiffMeth with a chi-squared test with basic overdispersion correction. Differentially methylated regions (DMRs) were defined with a q-value <0.01 and a percent difference in methylation of at least 25% for TA versus quadricep and FSHD2 versus control for TE loci but 10% for FSHD2 versus control.

Regions were associated with genes and gene ontology using rGREAT (version 1.20.0) [32] with submitGreatJob with hg38 specified. GO term plots include the top 10 terms with greater than 10 genes associated with the background set for the term and a Benjamini-Hochberg corrected p-value of less than 0.05. Regions were overlapped with SMCHD1 regulated regions and DUX4 binding sites from [24] using findOverlaps from GenomicRanges (version 1.40.0) [66].

For the promoter analysis, CpG sites were summarized for 1.5 kb upstream and 0.5 kb downstream of the TSS using regionCounts and filtered for at least 1 CpG site. Promoters were filtered for a coverage of at least 5 then normalized. Promoters detected in all samples were kept for further analysis as described.

3.5.4 Motif analysis

Motif enrichment was performed on DMRs extended 24 bp in both directions since the regions end on CpG sites. Enrichment was performed using AME (meme, version 5.2.0) [67] on JASPAR 2020 core redundant motifs from human only [68]. Association of TFs with GO terms for figure 3.1C was done by associating the TFs with the following GO terms: limb morphogenesis, embryonic limb morphogenesis, embryonic forelimb morphogenesis, limb development, limb bud formation for limb; anterior/posterior pattern specification, proximal/distal pattern formation, dorsal/ventral pattern formation, dorsal/ventral pattern specification for axis patterning; skeletal muscle tissue development, skeletal system development, negative regulation of myoblast differentiation, muscle organ development, skeletal muscle cell differentiation, skeletal muscle tissue regeneration, myoblast development, and positive regulation of myoblast proliferation for muscle. Upset plots were created using UpsetR (version 1.4.0) [69].

3.5.5 TE methylation analysis

For the TE analysis, CpG sites were summarized over LTR loci from repeatmasker taken from UCSC and filtered for at least 1 CpG site. Loci were filtered for a coverage of at least 5 then normalized. Loci detected in all samples were kept for further analysis as described.

3.5.6 RNA sequencing library preparation

Total RNA was collected using the RNEasy kit (74106, Qiagen). RNA was converted to cDNA using the Smart-Seq2 protocol [70]. Libraries were constructed with the Nextera DNA Library Prep Kit (Illumina) for days 0 to 5 for Control-3, Control-4, FSHD2-2 and FSHD2-3.

The Nextera DNA Flex Library Prep Kit (Illumina) was used for all other RNA-seq samples. Libraries were sequenced on the Illumina NextSeq500 with paired-end 43 bp reads to a depth of 5 to 40 million reads. Data for days 0 through 5 for Control-1, Control-2, Control-3, Control-4, FSHD2-1 and FSHD2-2 were obtained from GEO (accession number GSE143493).

3.5.7 RNA sequencing data processing

Raw reads from bulk RNA-seq were mapped to hg38 by STAR (version 2.5.1b) [71] using defaults except with a maximum of 10 mismatches per pair, a ratio of mismatches to read length of 0.07, and a maximum of 10 multiple alignments. Quantitation was performed using RSEM (version 1.2.31) [72] with defaults with gene annotations for protein coding genes and lncRNAs from GENCODE v28. Counts were batch corrected for the two different library prep methods using ComBat-seq from sva (version 3.36.0) [73]. Genes were filtered for 10 counts in all samples of at least one condition (same FSHD status, muscle group and differentiation day) using filterByExpr from edgeR (version 3.30.3) [74]. Normalization for days 0 to 5 and 12 for control and FSHD2 for figures 3.1 through 3.3 was performed separately from days 0 and 5 for FSHD2 and FSHD1 for figure 3.5. TMM normalized counts from edgeR were used for differential expression analysis in edgeR and clustering genes into similar expression profiles using maSigPro (version 1.60.0) [75]. TMM normalized counts were TPM normalized for plotting using effective gene lengths for each sample calculated by RSEM. Heatmaps were created using ComplexHeatmap (version 2.4.3) [76]. Median expression of gene groups for each sample was calculated by taking the median TPM normalized TMM values for all genes, and lines represent the mean of the median values for the given samples. Muscle specific genes were identified by performing pairwise comparisons between muscle groups (i.e. TA vs quadricep, TA vs deltoid, TA vs bicep) as described with edgeR and taking the intersect of all genes specific to the same muscle group in all comparisons (i.e. up in all TA comparisons). Genes with higher expression in more or less susceptible tissue were identified by intersecting the genes higher in TA than quadricep and higher in bicep than deltoid and vice versa. The list of transcription factors was taken from AnimalTFDB (version 3.0) [77].

3.5.8 TE mapping and processing

To map to TEs, fastqs were aligned as described above to the full GENCODE v28 annotation with a maximum of 100 multiple alignments and maximum of 100 loci anchors. Reads mapping to TE loci from repeatmasker from UCSC were estimated using featureCounts (subread version 2.0.1) [78] with fractional counts for multimapped reads for paired end reads. Genes were filtered for 2 counts in all samples of at least one condition (same FSHD status, muscle group and differentiation day) using filterByExpr from edgeR (version 3.30.3) [74]. Counts were normalized as described above. Loci were clustered into similar expression profiles as described above. Loci were overlapped with FSHD-induced genes (TSS to TES) or DUX4 binding sites from [24] using findOverlaps from GenomicRanges (version 1.40.0) [66]. Fisher's exact test (stats version 4.0.2) was used for enrichment of TE classes and DUX4 binding site overlaps with the "greater" alternative hypothesis.

3.5.9 Processing of publicly available data

Fastqs for the biopsy RNA-seq data from [21,25,79,80] were obtained from GEO with the accessions in table 3.1. Fastqs were processed in the same way as bulk RNA-seq. Counts were normalized as described above but with no batch correction.

3.6 Figures

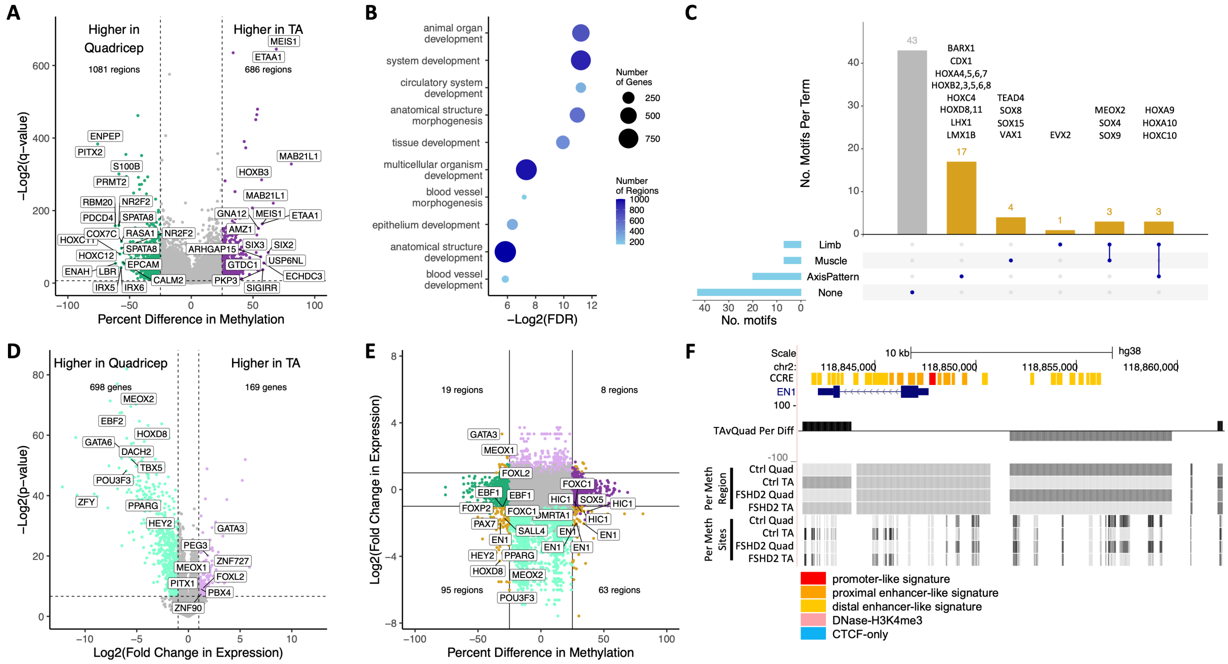


Figure 3.1: Adult muscle cells retain differences in DNA methylation and gene expression for transcription factors involved in developmental patterning (A) Volcano plot of DNA methylation differences for control cells from TA versus quadricep. Regions with a p-value of >0.01 or a percent methylation difference of <25 are taken as not significant and colored in grey. Regions with more methylation in TA are in purple. Regions with more methylation in quadricep are in green. The top 10 regions with the highest percent difference in the positive or negative directions are labelled. The numbers of regions higher in TA and quadricep are labelled at the top. Vertical lines intersect y-axis at -25 and 25 percent. Horizontal line intersects y-axis at -log2(0.01). (B) Gene ontology terms enriched in differentially methylated regions between TA and quadricep. The top 10 significant terms (p-value <0.05) with at least 10 genes associated to the term from the background set of regions detected are shown. Color indicates number of differentially methylated regions in the term. Size indicates the number of genes associated with the differentially methylated regions in the term. (C) Upset plot of gene ontology terms of transcription factors whose motifs are enriched in differentially methylated regions between TA and quadricep. Each transcription factor was annotated with GO terms related to limb, muscle or axis patterning (see Methods for details). The number of transcription factors found enriched in the DMRs that falls into the individual categories are shown in blue on the left. The number of transcription factors belonging to the interest of the categories indicated in dark blue are in gold and grey on the top. (D) Volcano plot of gene expression differences for control cells from TA versus quadricep. Genes with a p-value of >0.01 or an absolute log2(fold change in expression) <1 are in grey. Genes with higher expression in the TA are in light purple. Genes with higher expression in the quadricep are in light green. The numbers of genes with higher expression in TA or quadricep are labelled at the top. The 10 transcription factors with the lowest p-values for positive or negative fold change are labelled. Vertical lines intersect y-axis at -1 and 1 log2(fold change). Horizontal line intersects y-axis at -log2(0.01). (E) Scatterplot of differences in DNA

methylation and expression of nearby gene for TA versus quadricep. Regions are associated with genes through GREAT. Regions with significant differences in both DNA methylation and gene expression are in gold, and the numbers of these regions are labelled in the corresponding quadrant. All regions associated with transcription factors in the given quadrants are labelled. Regions associated with genes with significant differences in gene expression but not DNA methylation are colored in light purple and light green. Regions with significant differences in DNA methylation but not in expression of the associated gene are colored in green and purple. Vertical lines intersect x-axis at -25 and 25 percent. Horizontal lines intersect y-axis at -1 and 1 log2(fold change). (F) UCSC genome browser shot of DNA methylation near *EN1*. *EN1* gene model is from gencode v28. Percent methylation for each group of samples in the merged region are labelled with increase in methylation indicated with darker grey. The percent difference in methylation is indicated as a barplot with higher methylation in TA as positive and higher methylation in quadricep as negative. Percent methylation at individual sites for the given groups are at the bottom with darker color indicating higher methylation.

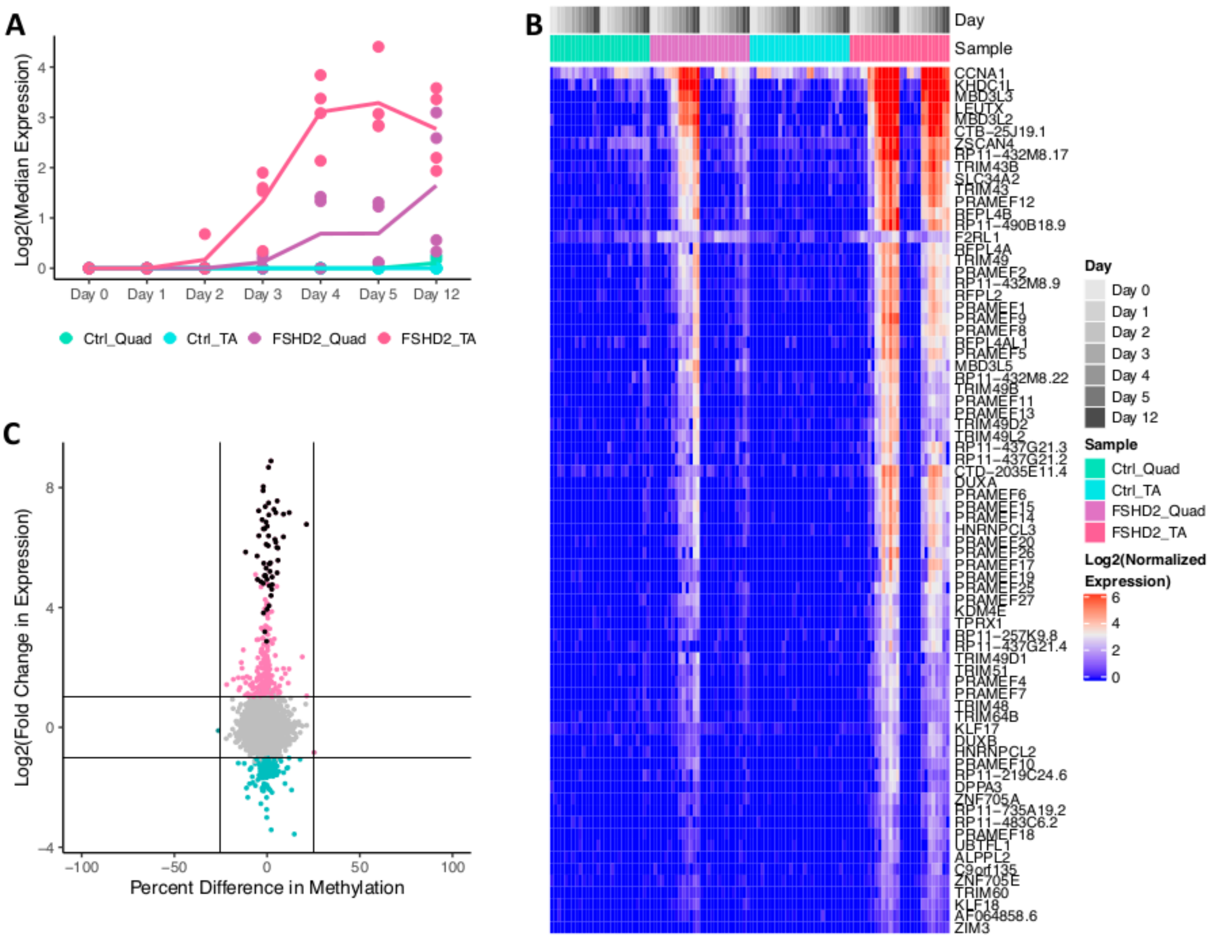


Figure 3.2: Promoters of FSHD-induced genes are not differentially methylated (A) Median expression for given groups of 74 FSHD-induced genes with increased expression in FSHD2 during myogenesis. Dots represent median expression for individual samples. Line represents the mean for the four samples in each group. (B) Heatmap of the 74 FSHD-induced genes ordered by hierarchical clustering. (C) Scatterplot of differences in expression and promoter DNA methylation for FSHD2 versus control. Promoters are defined as -1.5kb to +0.5kb around the TSS. FSHD-induced genes are colored in black. Regions associated with genes with significant differences in gene expression but not DNA methylation are colored in pink and teal. Regions with significant differences in DNA methylation but not in expression of the associated gene are colored in dark teal and dark pink. Vertical lines intersect x-axis at -25 and 25 percent. Horizontal lines intersect y-axis at -1 and 1 log2(fold change).

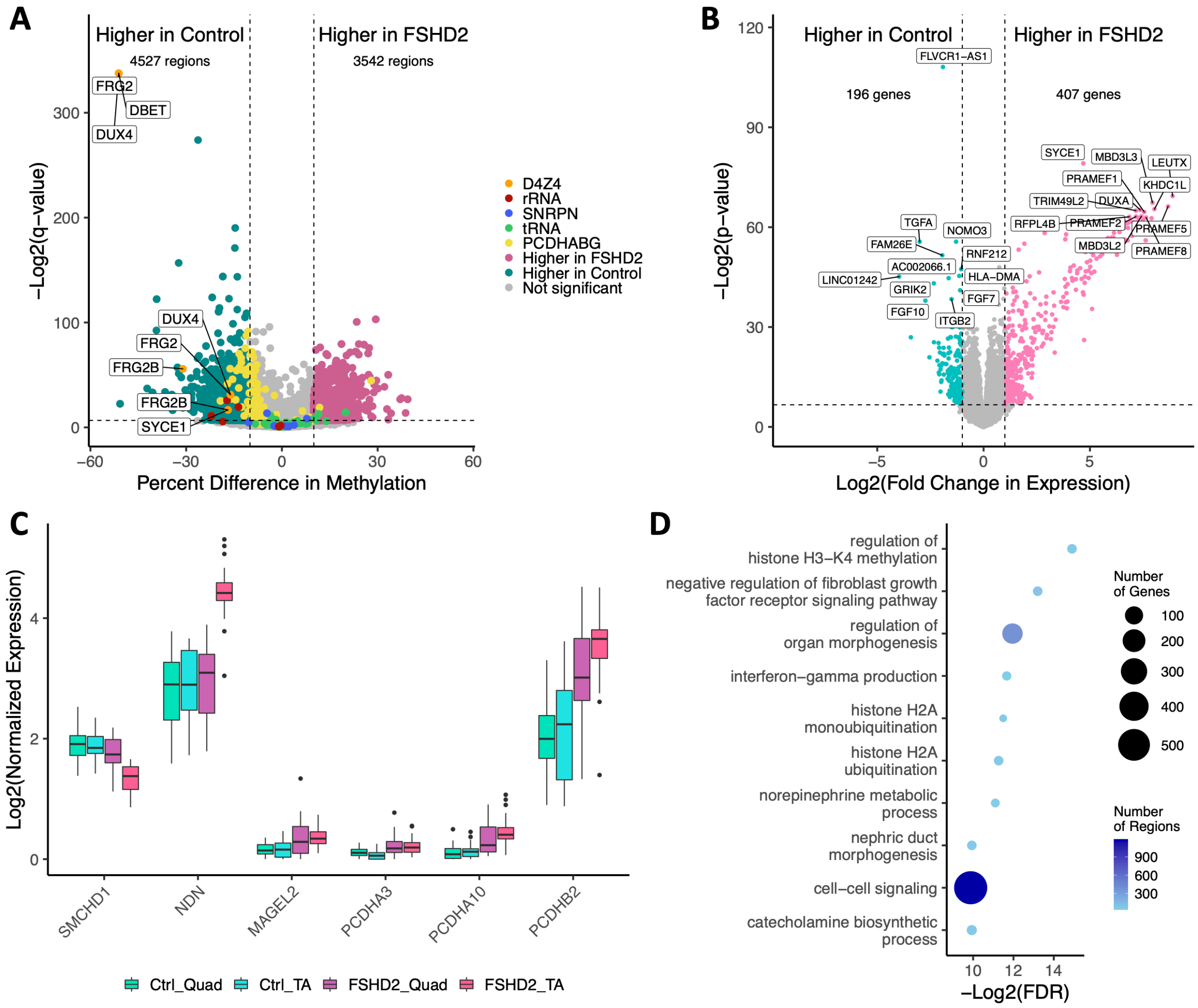


Figure 3.3: Global DNA methylation and gene expression differences in FSHD2 are associated with SMCHD1 mutations and signaling molecules (A) Volcano plot of DNA methylation differences for FSHD2 versus control. Regions with a p-value of >0.01 or a percent methylation difference of <10 are taken as not significant and colored in grey. Regions with more methylation in FSHD2 are in dark pink. Regions with more methylation in quadricep are in dark teal. Regions within SMCHD1 regulated regions are colored according to the legend. Regions in the D4Z4 region on chromosome 4 or 12 have associated genes labelled. The numbers of regions higher in FSHD2 and control are labelled at the top. Vertical lines intersect y-axis at -10 and 10 percent. Horizontal line intersects y-axis at -log2(0.01). (B) Volcano plot of gene expression differences for FSHD2 versus control. Genes with a p-value of >0.01 or an absolute log2(fold change in expression) <1 are in grey. Genes with higher expression in the TA are in pink. Genes with higher expression in the quadricep are in teal. The numbers of genes with higher expression in FSHD2 or control are labelled at the top. The 12 genes with the lowest p-values for positive or negative fold change are labelled. Vertical lines intersect y-axis at -1 and 1 log2(fold change). Horizontal line intersects y-axis at -log2(0.01). (C) Boxplot of expression for genes in SMCHD1 regulated clusters and *SMCHD1*. (D) Gene ontology terms enriched in differentially methylated regions between FSHD2 and control. The top 10 significant terms (p-

value <0.05) with at least 10 genes associated to the term from the background set of regions detected are shown. Color indicates number of differentially methylated regions in the term. Size indicates the number of genes associated with the differentially methylated regions in the term.

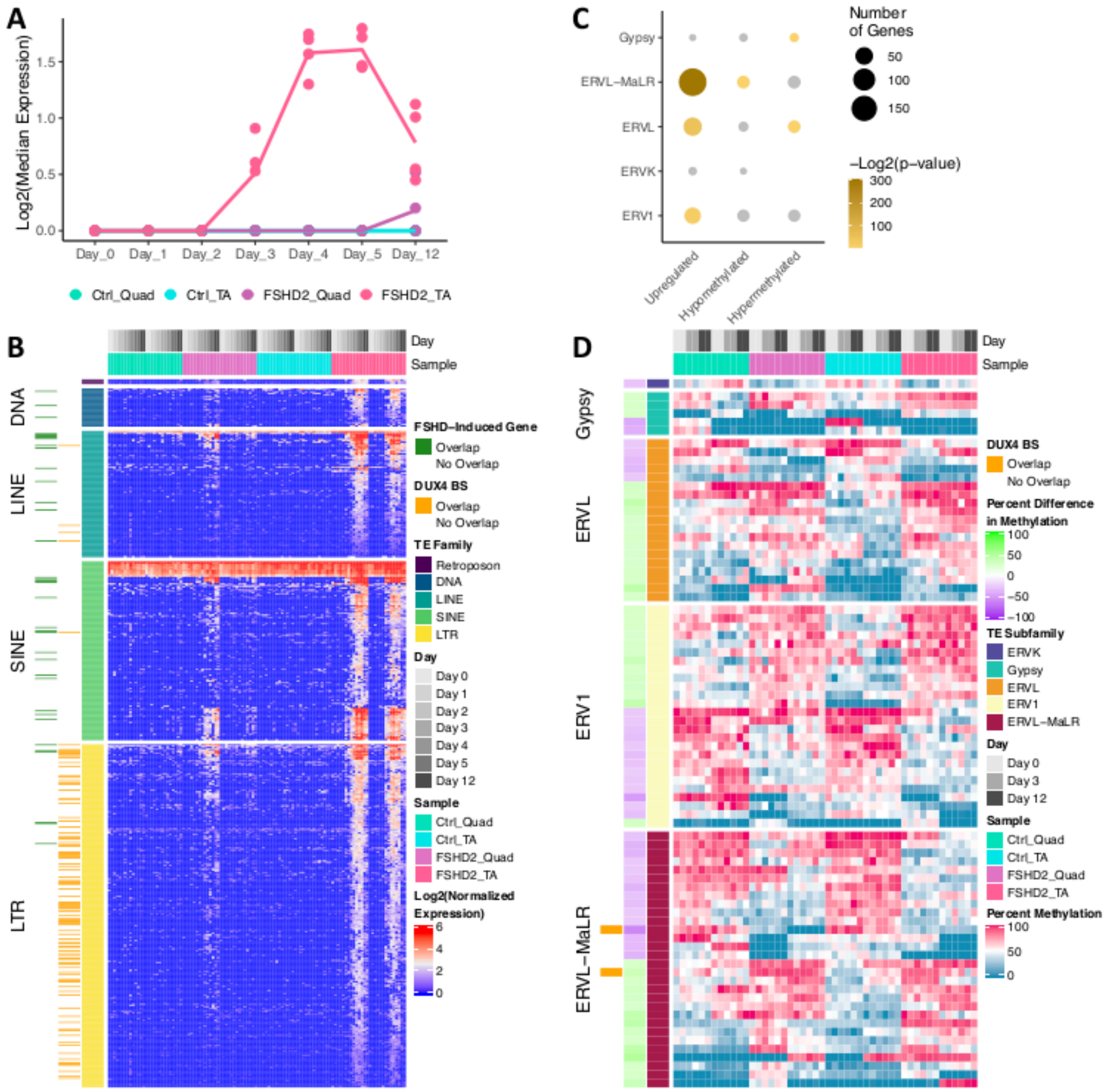


Figure 3.4: LTRs upregulated and hypomethylated in FSHD2 (A) Median expression for given groups of 580 TEs with increased expression in FSHD2 during myogenesis. Dots represent median expression for individual samples. Line represents the mean for the four samples in each group. (B) Heatmap of expression of TEs with increased expression in FSHD2 during myogenesis split by class of TE. Loci that overlap FSHD-induced genes are indicated in green on the left. Loci which overlap known DUX4 binding sites are labelled in orange. (C) Enrichment of subclasses of LTRs in TEs upregulated in FSHD2 or differentially methylated in FSHD2. Classes with p-value >0.05 are in grey. Size indicates number of loci in category. (D) Heatmap of percent methylation for LTR loci with significant methylation differences between FSHD2 and control split by LTR subclass. Loci which overlap a known DUX4 binding site are indicated in orange on the left.

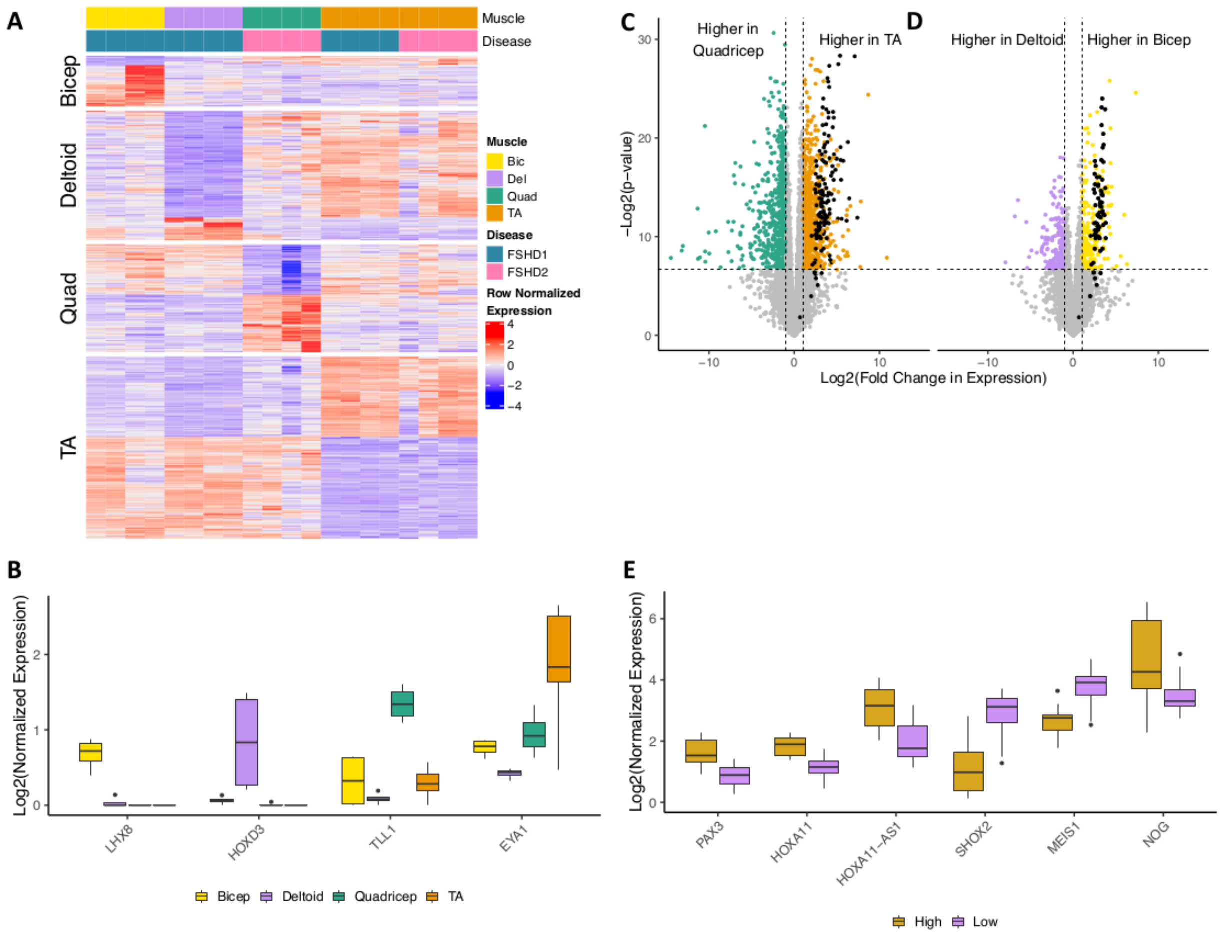


Figure 3.5: Muscle group specific expression in muscles more and less susceptible to FSHD (A) Heatmap of genes with muscle group specific expression. Top color bars indicate FSHD1 or FSHD2 and the muscle group of origin. (B) Boxplot of expression of four genes with muscle group specific expression. (C) Volcano plot of gene expression differences for FSHD2 TA versus FSHD2 quadricep at day 5 of differentiation. FSHD-induced genes are colored in black. Genes with a p-value of >0.01 or an absolute log2(fold change in expression) <1 are in grey. Genes with higher expression in the TA or quadricep are in orange or green, respectively. Vertical lines intersect y-axis at -1 and 1 log2(fold change). Horizontal line intersects y-axis at -log2(0.01). (D) Volcano plot of gene expression differences for FSHD1 bicep versus FSHD1 deltoid at day 5 of differentiation. FSHD-induced genes are colored in black. Genes with a p-value of >0.01 or an absolute log2(fold change in expression) <1 are in grey. Genes with higher expression in the bicep or deltoid are in yellow or purple, respectively. Vertical lines intersect y-axis at -1 and 1 log2(fold change). Horizontal line intersects y-axis at -log2(0.01). (E) Boxplot of genes with differences in expression in muscle groups with high or low susceptibility to FSHD.

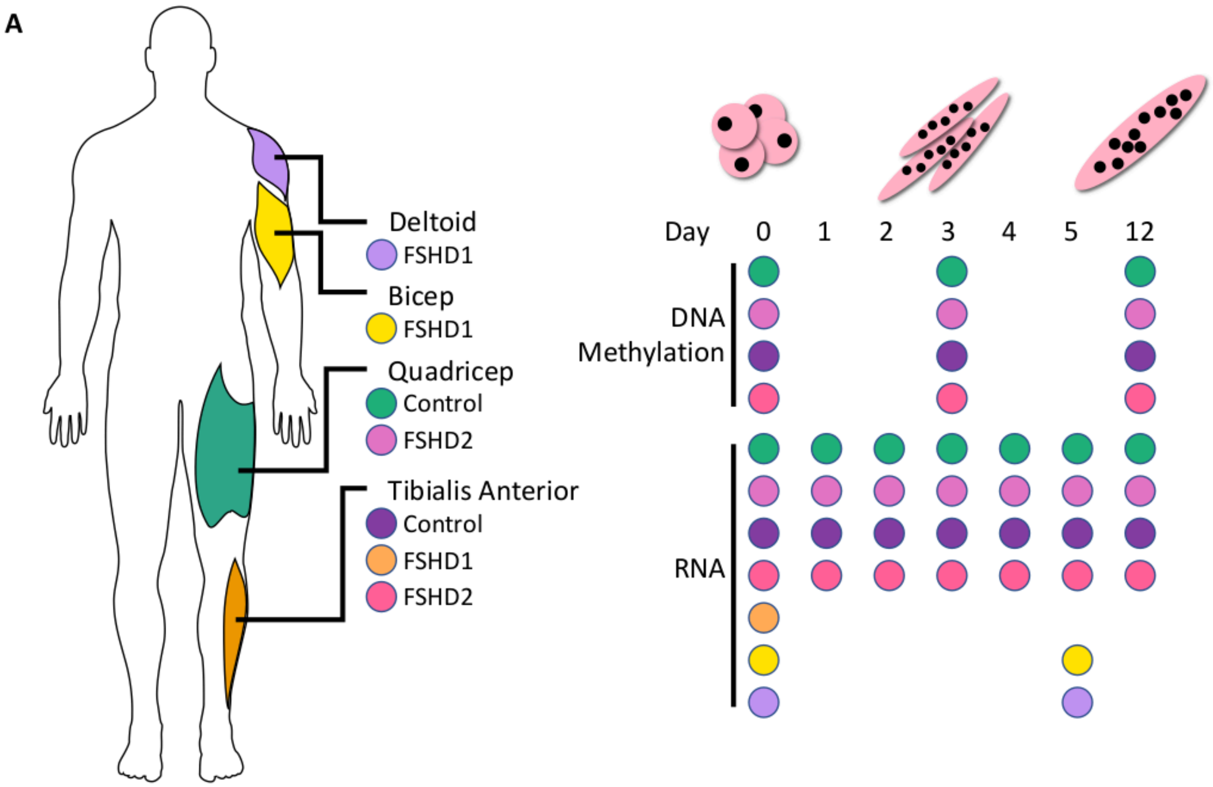


Figure S3.1: Experiment overview (A) Overview of samples used for bisulfite and RNA sequencing.

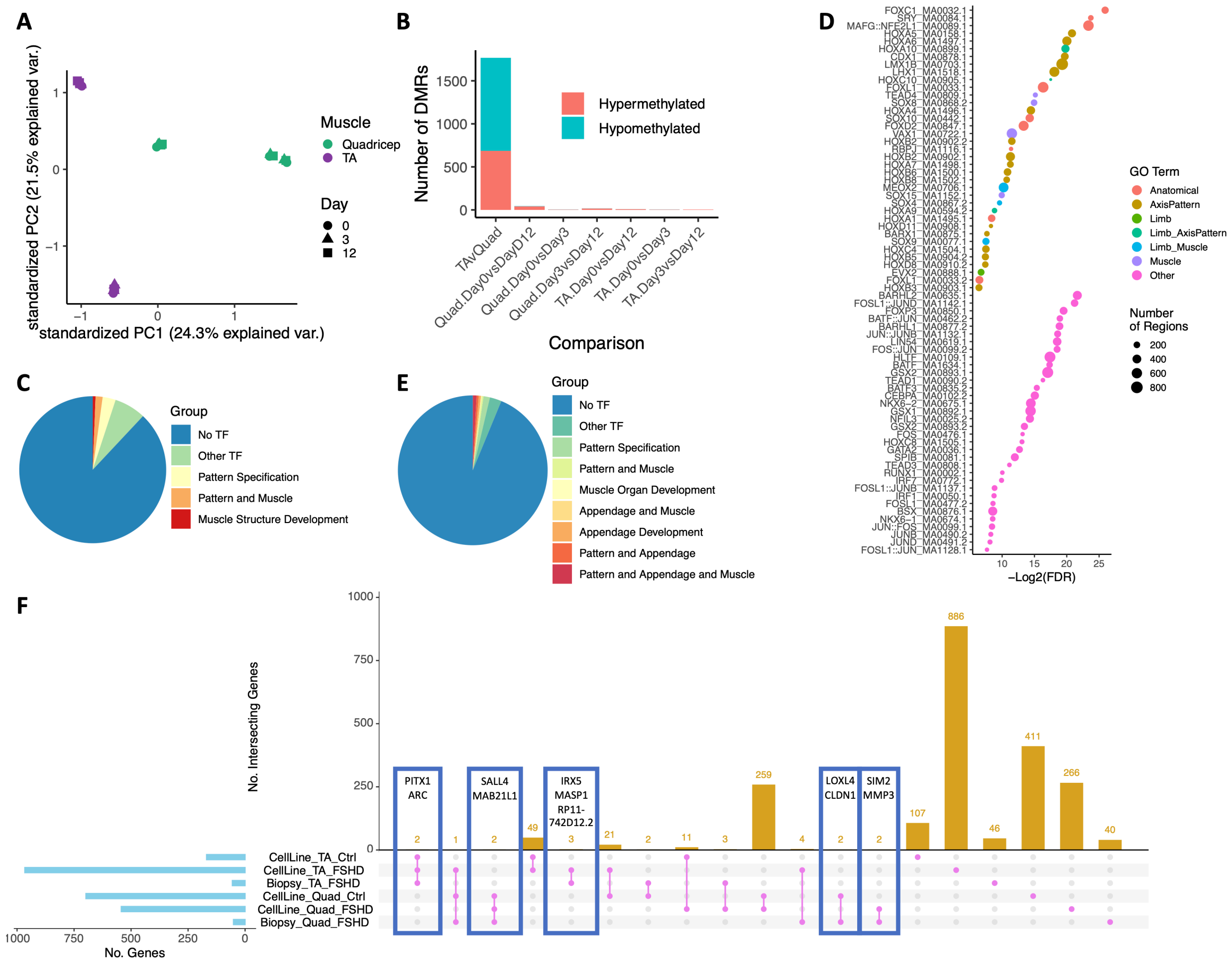


Figure S3.2: DNA methylation and gene expression differences for TA compared to quadricep (A) Principal component analysis (PCA) of control TA and quadricep DNA methylation. (B) Stacked barplot of the number of differentially methylated regions between TA and quadricep. (C) Pie chart of the number of differentially methylated regions associated with genes for transcription factors. Regions not associated with a transcription factor are in blue. Regions associated with transcription factors are colored by annotated gene ontology terms for pattern specification or muscle structure development. (D) Transcription factor motifs enriched in differentially methylated regions between TA and quadricep. Color indicates the gene ontology annotation associated with transcription factor, either anatomical development, axis patterning, limb development, or muscle development (see Methods for specific GO terms). Size indicates number of regions with given motif. FDR cutoff of 0.05. (E) Pie chart of the number of differentially expressed transcription factors for TA versus quadricep. Genes that are not transcription factors are in blue. Transcription factors are colored by annotated gene ontology terms for pattern specification, appendage development or muscle structure development. (F) Upset plot of genes differentially expressed between TA and quadricep in control cells, FSHD2 cells or in FSHD biopsies from [26]. Number of genes higher in the given muscle in the given comparison are indicated in the light blue barplot on the left. Number of genes found in the intersections indicated in magenta are given in the gold barplot at top. Genes in intersections outlined in blue are labelled.

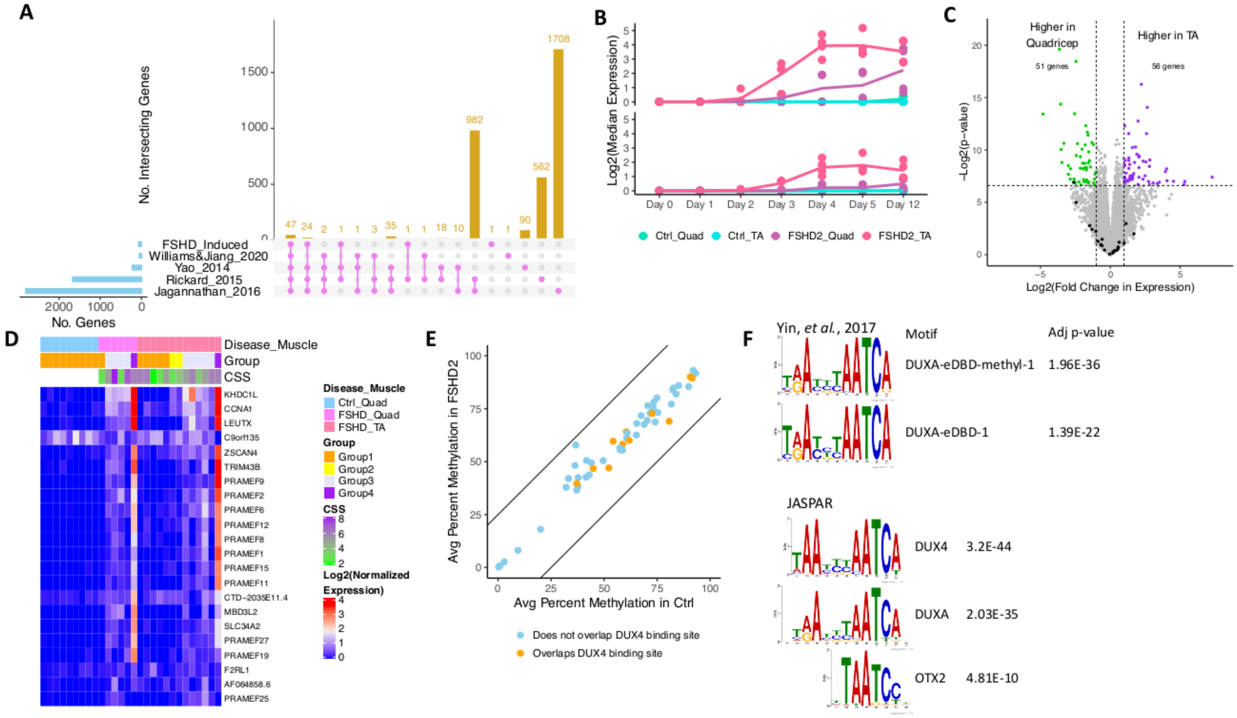


Figure S3.3: Assessment of FSHD-induced genes (A) Upset plot of FSHD-induced genes compared with the gene list from [53] and reanalyzed data from [21,75,76]. Yao 2014 is a comparison of differentiated FSHD and control myotubes [21]. Rickard 2015 is a comparison of DUX4 expressing myocytes identified by a reporter to those negative for the reporter [76]. Jagannathan 2016 is a comparison of induced expression of DUX4 to non-induced [75]. (B) Median expression for given groups of 74 FSHD-induced genes. Those identified in [53] are on top, and those new this analysis are on bottom. Dots represent median expression for individual samples. Line represents the mean for the four samples in each group. (C) Volcano plot of expression differences between TA and quadricep from FSHD biopsies from [26]. Genes with higher expression in TA or quadricep are in purple and green, respectively. FSHD-induced genes are colored in black. The number of genes higher in each muscle are labelled at the top. (D) Heatmap of FSHD-induced gene expression with detectable in control quadricep, FSHD quadricep and FSHD TA biopsy samples from [26]. Groups were identified in [26] with an increase in group number correlating with an increase in DUX4 target gene expression. Clinical severity score (CSS) is on a 10 point scale. (E) Scatterplot of percent methylation in the promoters of FSHD-induced genes for control and FSHD2. Lines indicate a 25% difference. Promoters which overlap known DUX4 binding sites are colored orange. (F) Motifs from [35] or [64] enriched in promoters of FSHD-induced genes.

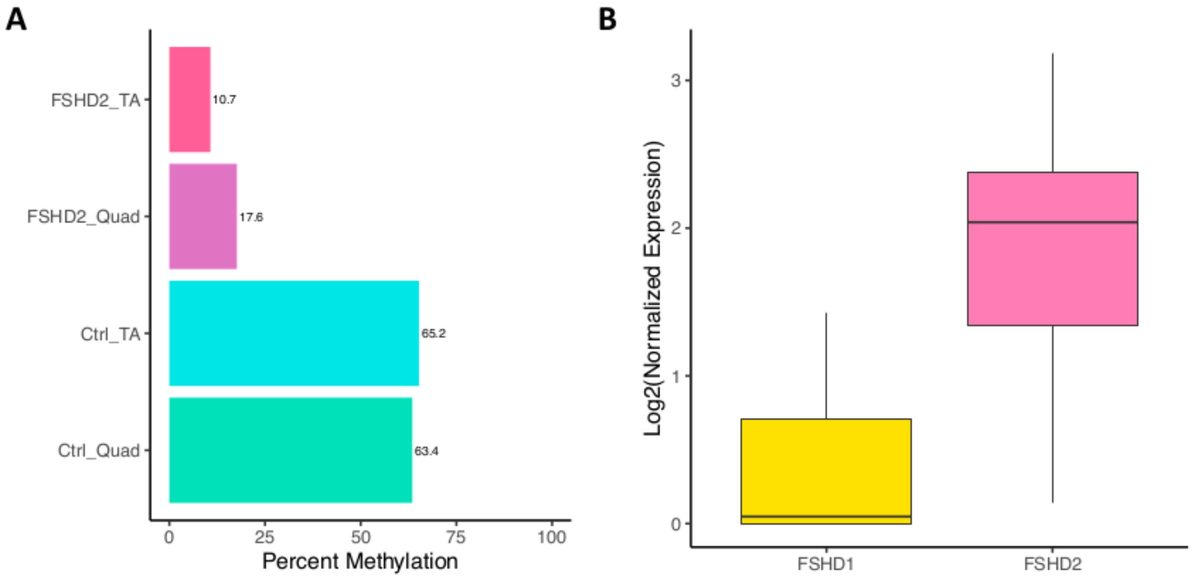


Figure S3.4: SMCHD1 related methylation and expression differences (A) Barplot of average percent methylation near *DBET* on chromosome 4. Numbers to the left of bars give the average percent methylation. (B) Boxplot of *SYCE1* expression in FSHD1 and FSHD2 samples at day 0 of differentiation.

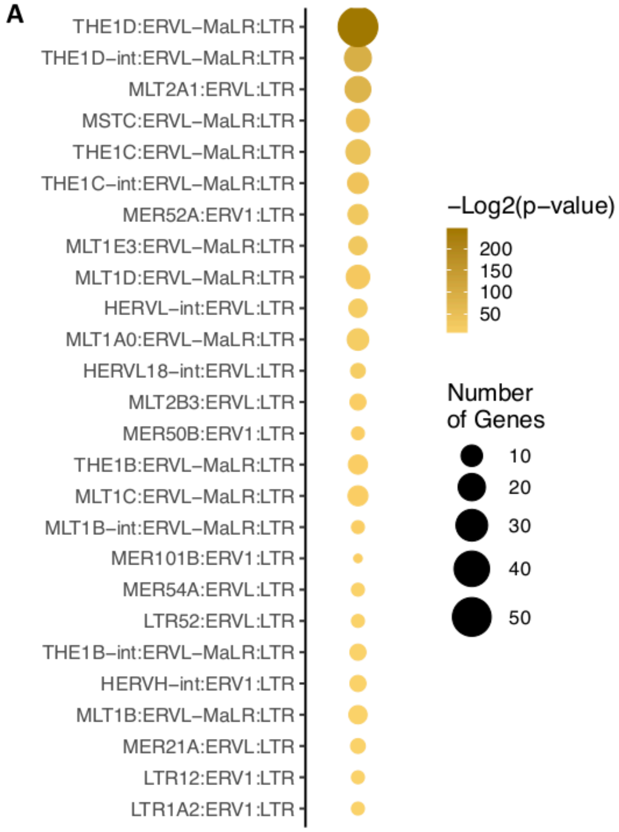


Figure S3.5: Enrichment of specific types of LTRs (A) Enrichment of specific types of LTRs in upregulated in FSHD2. Classes with p-value >0.05 are not shown. Size indicates number of loci in category.

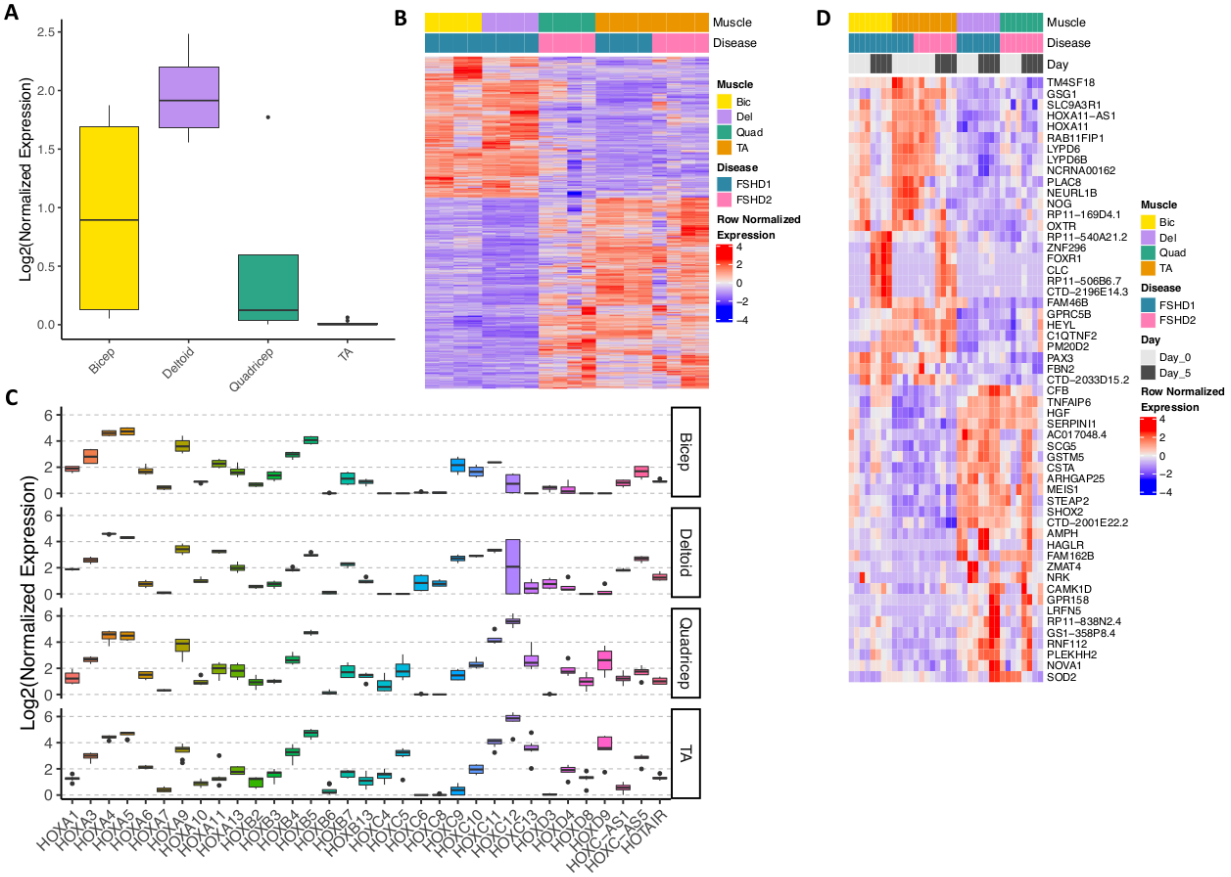


Figure S3.6: Muscle group specific expression (A) Boxplot of *TBX5* expression in the different muscle groups at day 0 of differentiation. (B) Heatmap of genes specific to the upper (bicep and deltoid) or lower (TA and quadricep) body muscles at day 0 of differentiation. (C) Boxplot of HOX gene expression at day 0 of differentiation split by muscle group. (D) Heatmap of genes with differential expression in highly (TA and bicep) and less (quadricep and deltoid) susceptible muscle groups at days 0 and 5 of differentiation combined.

3.7 Tables

Table 3.1: GEO Accession numbers for fastqs from previous studies

Disease	Muscle	SRA	Reference
Ctrl	Quad	SRR7293780	[25]
Ctrl	Quad	SRR7293781	[25]
Ctrl	Quad	SRR7293782	[25]
Ctrl	Quad	SRR7293783	[25]
Ctrl	Quad	SRR7293784	[25]
Ctrl	Quad	SRR7293785	[25]
Ctrl	Quad	SRR7293793	[25]
Ctrl	Quad	SRR7293794	[25]
Ctrl	Quad	SRR7293795	[25]
FSHD	Quad	SRR7293761	[25]
FSHD	Quad	SRR7293769	[25]
FSHD	Quad	SRR7293774	[25]
FSHD	Quad	SRR7293786	[25]
FSHD	Quad	SRR7293788	[25]
FSHD	Quad	SRR7293801	[25]
FSHD	TA	SRR7293759	[25]
FSHD	TA	SRR7293763	[25]
FSHD	TA	SRR7293766	[25]
FSHD	TA	SRR7293767	[25]
FSHD	TA	SRR7293770	[25]
FSHD	TA	SRR7293771	[25]
FSHD	TA	SRR7293787	[25]
FSHD	TA	SRR7293790	[25]
FSHD	TA	SRR7293791	[25]
FSHD	TA	SRR7293792	[25]
FSHD	TA	SRR7293796	[25]
FSHD	TA	SRR7293799	[25]
FSHD	TA	SRR7293800	[25]
MB135_HDUX4CA_nodox_rep1	NA	SRR4019004	[75]
MB135_HDUX4CA_WITHdox_rep1	NA	SRR4019005	[75]
MB135_HDUX4CA_nodox_rep2	NA	SRR4019006	[75]
MB135_HDUX4CA_WITHdox_rep2	NA	SRR4019007	[75]
MB135_HDUX4CA_nodox_rep3	NA	SRR4019008	[75]
MB135_HDUX4CA_WITHdox_rep3	NA	SRR4019009	[75]

FSHD_1_1_neg	NA	SRR2020583	[76]
FSHD_1_2_neg	NA	SRR2020584	[76]
FSHD_2_2_BFP	NA	SRR2020585	[76]
FSHD_2_3_BFP	NA	SRR2020586	[76]
FSHD_1_3_neg	NA	SRR2020587	[76]
FSHD_1_1_BFP	NA	SRR2020588	[76]
FSHD_1_2_BFP	NA	SRR2020589	[76]
FSHD_1_3_BFP	NA	SRR2020590	[76]
FSHD_2_1_neg	NA	SRR2020591	[76]
FSHD_2_2_neg	NA	SRR2020592	[76]
FSHD_2_3_neg	NA	SRR2020593	[76]
FSHD_2_1_BFP	NA	SRR2020594	[76]
Control_20_Mt	NA	SRR1398556	[21]
Control_21_Mb	NA	SRR1398557	[21]
Control_21_Mt	NA	SRR1398558	[21]
Control_22_Mb	NA	SRR1398559	[21]
Control_22_Mt	NA	SRR1398560	[21]
FSHD2_12_Mt	NA	SRR1398561	[21]
FSHD2_14_Mb	NA	SRR1398562	[21]
FSHD2_14_Mt	NA	SRR1398563	[21]
FSHD2_20_Mb	NA	SRR1398564	[21]
FSHD2_20_Mt	NA	SRR1398565	[21]
FSHD1_4_Mb	NA	SRR1398566	[21]
FSHD1_4_Mt	NA	SRR1398567	[21]
FSHD1_6_Mb	NA	SRR1398568	[21]
FSHD1_6_Mt	NA	SRR1398569	[21]

Table S3.1: Motifs enriched in promoters of FSHD-induced genes

rank	motif_DB	motif_ID	motif_alt_ID	adj_p-value
1	Yin2017	DUXA-eDBD-methyl-1	NA	1.96E-36
2	Yin2017	DUXA-eDBD-1	NA	1.39E-22
3	Yin2017	MIXL1-FL-1	NA	8.30E-13
4	Yin2017	MIXL1-FL-methyl-1	NA	1.12E-10
5	Yin2017	PROP1-eDBD-1	NA	7.17E-09
6	Yin2017	PHOX2B-FL-methyl-1	NA	2.96E-07
7	Yin2017	BACH2-eDBD-1	NA	7.61E-06
8	Yin2017	PHOX2B-FL-1	NA	8.41E-06
9	Yin2017	NR4A1-eDBD-methyl-1	NA	1.63E-05
10	Yin2017	PROP1-eDBD-methyl-1	NA	1.80E-05
11	Yin2017	SREBF2-eDBD-methyl-1	NA	4.03E-05
12	Yin2017	FOXI1-FL-methyl-1	NA	5.13E-05
13	Yin2017	SREBF1-eDBD-methyl-1	NA	5.49E-05
14	Yin2017	OTX1-FL-1	NA	1.18E-04
15	Yin2017	NR4A1-eDBD-1	NA	1.32E-04
16	Yin2017	PAX4-eDBD-3	NA	3.12E-04
17	Yin2017	NR2E1-FL-methyl-1	NA	3.33E-04
18	Yin2017	NR2E1-FL-1	NA	3.36E-04
19	Yin2017	PHOX2A-eDBD-1	NA	3.69E-04
20	Yin2017	DRGX-eDBD-methyl-1	NA	4.91E-04
21	Yin2017	NR4A1-eDBD-methyl-2	NA	6.88E-04
22	Yin2017	SPDEF-eDBD-methyl-1	NA	7.30E-04
23	Yin2017	BACH2-eDBD-methyl-1	NA	1.05E-03
24	Yin2017	OTX1-FL-methyl-1	NA	1.52E-03
25	Yin2017	LHX4-FL-methyl-1	NA	1.60E-03
26	Yin2017	POU2F2-eDBD-methyl-2	NA	1.71E-03
27	Yin2017	NR4A2-eDBD-methyl-1	NA	2.10E-03
28	Yin2017	IRF8-FL-methyl-1	NA	2.43E-03
29	Yin2017	NR4A1-eDBD-2	NA	2.56E-03
30	Yin2017	HNF4A-eDBD-methyl-2	NA	2.57E-03
31	Yin2017	PBX1-FL-methyl-1	NA	2.70E-03
32	Yin2017	NR4A2-eDBD-1	NA	3.55E-03
33	Yin2017	POU3F4-eDBD-5	NA	4.85E-03
34	Yin2017	NR1I3-FL-1	NA	5.39E-03
1	JASPAR2020_HumanOnly	MA0468.1	DUX4	3.20E-44
2	JASPAR2020_HumanOnly	MA0884.1	DUXA	2.03E-35

3	JASPAR2020_HumanOnly	MA0712.1	OTX2	4.81E-10
4	JASPAR2020_HumanOnly	MA0158.1	HOXA5	1.38E-07
5	JASPAR2020_HumanOnly	MA0829.2	SREBF1(var.2)	4.36E-07
6	JASPAR2020_HumanOnly	MA1124.1	ZNF24	6.82E-07
7	JASPAR2020_HumanOnly	MA1101.2	BACH2	7.61E-06
8	JASPAR2020_HumanOnly	MA0714.1	PITX3	2.09E-05
9	JASPAR2020_HumanOnly	MA1644.1	NFYC	3.88E-05
10	JASPAR2020_HumanOnly	MA0472.1	EGR2	1.87E-04
11	JASPAR2020_HumanOnly	MA0711.1	OTX1	2.68E-04
12	JASPAR2020_HumanOnly	MA1152.1	SOX15	2.94E-04
13	JASPAR2020_HumanOnly	MA0901.2	HOXB13	3.18E-04
14	JASPAR2020_HumanOnly	MA0713.1	PHOX2A	3.36E-04
15	JASPAR2020_HumanOnly	MA0782.2	PKNOX1	5.29E-04
16	JASPAR2020_HumanOnly	MA0477.1	FOSL1	7.58E-04
17	JASPAR2020_HumanOnly	MA0060.3	NFYA	1.12E-03
18	JASPAR2020_HumanOnly	MA0046.2	HNF1A	1.58E-03
19	JASPAR2020_HumanOnly	MA0841.1	NFE2	2.86E-03
20	JASPAR2020_HumanOnly	MA1112.2	NR4A1	2.98E-03
21	JASPAR2020_HumanOnly	MA0491.1	JUND	3.98E-03
22	JASPAR2020_HumanOnly	MA1534.1	NR1I3	5.39E-03
23	JASPAR2020_HumanOnly	MA0715.1	PROP1	5.71E-03
24	JASPAR2020_HumanOnly	MA0706.1	MEOX2	6.20E-03
25	JASPAR2020_HumanOnly	MA0491.2	JUND	8.78E-03
26	JASPAR2020_HumanOnly	MA0072.1	RORA(var.2)	9.68E-03
27	JASPAR2020_HumanOnly	MA0153.2	HNF1B	1.02E-02

3.8 References

1. Science Reference Section L of C. What is the strongest muscle in the human body? [Internet]. 2019. Available from: https://www.loc.gov/everyday-mysteries/item/what-is-the-strongest-muscle-in-the-human-body/
2. Chal J, Pourquié O. Making muscle: Skeletal myogenesis in vivo and in vitro. Dev. 2017;144(12):2104–22.
3. Edgerton VR, Smith JL, Simpson DR. Muscle fibre type populations of human leg muscles. Histochem J. 1975;7(3):259–66.
4. Hernandez-Torres F, Rodríguez-Outeiriño L, Franco D, Aranega AE. Pitx2 in embryonic and adult myogenesis. Vol. 5, Frontiers in Cell and Developmental Biology. Frontiers Media S.A.; 2017. p. 46.
5. Buckingham M, Rigby PWJ. Gene Regulatory Networks and Transcriptional Mechanisms that Control Myogenesis. Vol. 28, Developmental Cell. 2014. p. 225–38.
6. L'Honoré A, Ouimette JF, Lavertu-Jolin M, Drouin J. Pitx2 defines alternate pathways acting through MyoD during limb and somitic myogenesis. Development. 2010;137(22):3847–56.
7. Braun T, Gautel M. Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. Nat Rev Mol Cell Biol. 2011;12(6):349–61.
8. McQueen C, Towers M. Establishing the pattern of the vertebrate limb. Development. 2020 Sep 1;147(17):dev177956.
9. Ye W, Song Y, Huang Z, Osterwalder M, Ljubojevic A, Xu J, et al. A unique stylopod patterning mechanism by shox2-controlled osteogenesis. Dev. 2016;143(14):2548–60.
10. Delgado I, Torres M. Gradients, waves and timers, an overview of limb patterning models. Semin Cell Dev Biol. 2016;49:109–15.
11. Ahn K, Mishina Y, Hanks MC, Behringer RR, Bryan Crenshaw E. BMPR-IA signaling is required for the formation of the apical ectodermal ridge and dorsal-ventral patterning of the limb. Development. 2001;128(22):4449–61.
12. Kang PB, Kho AT, Sanoudou D, Haslett JN, Dow CP, Han M, et al. Variations in gene expression among different types of human skeletal muscle. Muscle Nerve. 2005 Oct 1;32(4):483–91.
13. Porter JD, Merriam AP, Leahy P, Gong B, Feuerman J, Cheng G, et al. Temporal gene expression profiling of dystrophin-deficient (mdx) mouse diaphragm identifies conserved and muscle group-specific mechanisms in the pathogenesis of muscular dystrophy. Hum Mol Genet. 2004;13(3):257–69.
14. Terry EE, Zhang X, Hoffmann C, Hughes LD, Lewis SA, Li J, et al. Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues. Elife. 2018 May 29;7.
15. Evano B, Gill D, Hernando-Herraez I, Comai G, Stubbs TM, Commere PH, et al. Transcriptome and epigenome diversity and plasticity of muscle stem cells following transplantation. PLoS Genet. 2020 Oct 30;16(10):e1009022.
16. Ciciliot S, Rossi AC, Dyar KA, Blaauw B, Schiaffino S. Muscle type and fiber type specificity in muscle wasting. Int J Biochem Cell Biol. 2013;45(10):2191–9.
17. Wagner KR. Facioscapulohumeral Muscular Dystrophies. Muscle Neuromuscul Junction Disord. 2019;25(6):1662–81.
18. Olsen DB, Gideon P, Jeppesen TD, Vissing J. Leg muscle involvement in facioscapulohumeral muscular dystrophy assessed by MRI. J Neurol.

2006;253(11):1437–41.

19. Statland JM, Tawil R. Facio-scapulo-humeral muscular dystrophy. Indian J Pediatr. 2008;34(5):186–8.

20. Lu J, Yao Z, Yang Y, Zhang C, Zhang J, Zhang Y. Management strategies in facioscapulohumeral muscular dystrophy. Intractable Rare Dis Res. 2019;8(1):9–13.

21. Yao Z, Snider L, Balog J, Lemmers RJLF, Van Der Maarel SM, Tawil R, et al. DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. Hum Mol Genet. 2014 Oct 15;23(20):5342–52.

22. Geng LN, Yao Z, Snider L, Fong AP, Cech JN, Young JM, et al. DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy. Dev Cell. 2012 Jan 17;22(1):38–51.

23. Tassin A, Laoudj-Chenivesse D, Vanderplanck C, Barro M, Charron S, Ansseau E, et al. DUX4 expression in FSHD muscle cells: How could such a rare protein cause a myopathy? J Cell Mol Med. 2013 Jan;17(1):76–89.

24. Young JM, Whiddon JL, Yao Z, Kasinathan B, Snider L, Geng LN, et al. DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. PLoS Genet. 2013 Nov;9(11).

25. Wang LH, Friedman SD, Shaw D, Snider L, Wong CJ, Budech CB, et al. MRI-informed muscle biopsies correlate MRI with pathology and DUX4 target gene expression in FSHD. Hum Mol Genet. 2019 Feb 1;28(3):476–86.

26. Wong CJ, Wang LH, Friedman SD, Shaw D, Campbell AE, Budech CB, et al. Longitudinal measures of RNA expression and disease activity in FSHD muscle biopsies. Hum Mol Genet. 2020;29(6):1030–44.

27. Rahimov F, King OD, Leung DG, Bibat GM, Emerson CP, Kunkel LM, et al. Transcriptional profiling in facioscapulohumeral muscular dystrophy to identify candidate biomarkers. Proc Natl Acad Sci. 2012 Sep 18;109(40):16234–9.

28. Daubas P, Duval N, Bajard L, Vives FL, Robert B, Mankoo BS, et al. Fine-tuning the onset of myogenesis by homeobox proteins that interact with the Myf5 limb enhancer. Biol Open. 2015;4(12):1614–24.

29. Ouimette JF, Jolin ML, L'honoré A, Gifuni A, Drouin J. Divergent transcriptional activities determine limb identity. Nat Commun. 2010 Jul 13;1(4):1–9.

30. Draut H, Liebenstein T, Begemann G. New insights into the control of cell fate choices and differentiation by retinoic acid in cranial, axial and caudal structures. Biomolecules. 2019;9(12).

31. Sturgeon K, Kaneko T, Biemann M, Gauthier A, Chawengsaksophak K, Cordes SP. Cdx1 refines positional identity of the vertebrate hindbrain by directly repressing Mafb expression. Development. 2011;138(1):65–74.

32. Mclean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions HHS Public Access Author manuscript. Nat Biotechnol. 2010;28(5):495–501.

33. Koshiba-Takeuchi K, Takeuchi JK, Arruda EP, Kathiriya IS, Mo R, Hui CC, et al. Cooperative and antagonistic interactions between Sall4 and Tbx5 pattern the mouse limb and heart. Nat Genet. 2006;38(2):175–83.

34. Li D, Sakuma R, Vakili NA, Mo R, Puviindran V, Deimling S, et al. Formation of proximal and anterior limb skeleton requires early function of Irx3 and Irx5 and is negatively regulated by shh signaling. Dev Cell. 2014;29(2):233–40.

35. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science (80-). 2017 May 5;356(6337).
36. Katayama S, Ranga V, Jouhilahti E-M, Airenne TT, Johnson MS, Mukherjee K, et al. Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. Sci Rep. 2018 Dec 27;8(1):17421.
37. Himeda CL, Jones TI, Jones PL. Facioscapulohumeral muscular dystrophy as a model for epigenetic regulation and disease. Antioxid Redox Signal. 2015 Jun 1;22(16):1463–82.
38. Greco A, Goossens R, van Engelen B, van der Maarel SM. Consequences of epigenetic derepression in facioscapulohumeral muscular dystrophy. Clin Genet. 2020;97(6):799–814.
39. de Greef JC, Lemmers RJLF, van Engelen BGM, Sacconi S, Venance SL, Frants RR, et al. Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. Hum Mutat. 2009 Oct;30(10):1449–59.
40. Lemmers RJLF, Tawil R, Petek LM, Balog J, Block GJ, Santen GWE, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat Genet. 2012 Dec;44(12):1370–4.
41. Mason AG, Slieker RC, Balog J, Lemmers RJLF, Wong CJ, Yao Z, et al. SMCHD1 regulates a limited set of gene clusters on autosomal chromosomes. Skelet Muscle. 2017 Jun 6;7(1).
42. Jansz N, Chen K, Murphy JM, Blewitt ME. The Epigenetic Regulator SMCHD1 in Development and Disease. Vol. 33, Trends in Genetics. Elsevier Ltd; 2017. p. 233–43.
43. Pan Z, Ding J, Yang Z, Li H, Ding H, Chen Q. LncRNA FLVCR1-AS1 promotes proliferation, migration and activates Wnt/β-catenin pathway through miR-381-3p/CTNNB1 axis in breast cancer. Cancer Cell Int. 2020;20(1):1–12.
44. Hendrickson PG, Doráis JA, Grow EJ, Whiddon JL, Lim JW, Wike CL, et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat Genet. 2017 Jun 1;49(6):925–34.
45. Zhou C, Yang G, Chen M, He L, Xiang L, Ricupero C, et al. Lhx6 and Lhx8: Cell fate regulators and beyond. FASEB J. 2015;29(10):4083–91.
46. Tarchini B, Duboule D. Control of Hoxd genes' collinearity during early limb development. Dev Cell. 2006;10(1):93–103.
47. Hopkins DR, Keles S, Greenspan DS. The Bone Morphogenetic Protein 1/Tolloid-like Metalloproteinases. Bone. 2007;26(7):508–23.
48. Allen DL, Greyback BJ, Hanson AM, Cleary AS, Lindsay SF. Skeletal muscle expression of bone morphogenetic protein-1 and tolloid-like-1 extracellular proteases in different fiber types and in response to unloading, food deprivation and differentiation. J Physiol Sci. 2010;60(5):343–52.
49. Petit F, Sears KE, Ahituv N. Limb development: A paradigm of gene regulation. Nat Rev Genet. 2017;18(4):245–58.
50. Fukuda S, Kaneshige A, Kaji T, Noguchi YT, Takemoto Y, Zhang L, et al. Sustained expression of HeyL is critical for the proliferation of muscle stem cells in overloaded muscle. Elife. 2019;8:1–21.
51. Noguchi YT, Nakamura M, Hino N, Nogami J, Tsuji S, Sato T, et al. Cell-autonomous and redundant roles of Hey1 and HeyL in muscle stem cells: HeyL requires HeS1 to bind diverse DNA sites. Dev. 2019;146(4):1–12.

52. Tsumagari K, Baribault C, Terragni J, Varley KE, Gertz J, Pradhan S, et al. Early de novo DNA methylation and prolonged demethylation in the muscle lineage. Epigenetics. 2013;8(3):317–32.

53. Saito A, Higuchi I, Nakagawa M, Saito M, Uchida Y, Inose M, et al. An overexpression of fibroblast growth factor (FGF) and FGF receptor 4 in a severe clinical phenotype of facioscapulohumeral muscular dystrophy. Muscle Nerve. 2000;4(23):490–7.

54. Zimmerman LB, De Jesús-Escobar JM, Harland RM. The Spemann organizer signal noggin binds and inactivates bone morphogenetic protein 4. Cell. 1996;86(4):599–606.

55. Gozo MC, Aspuria PJ, Cheon DJ, Walts AE, Berel D, Miura N, et al. Foxc2 induces Wnt4 and Bmp4 expression during muscle regeneration and osteogenesis. Cell Death Differ. 2013;20(8):1031–42.

56. van den Heuvel A, Mahfouz A, Kloet SL, Balog J, van Engelen BGM, Tawil R, et al. Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. Hum Mol Genet. 2018 Nov 16;

57. Williams K, Jiang S, Kong X, Zeng W, Nguyen NV, Ma X, et al. Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. PLoS Genet. 2020;16(5):1–26.

58. Cramer JM, Pohlmann D, Gomez F, Mark L, Kornegay B, Hall C, et al. Methylation specific targeting of a chromatin remodeling complex from sponges to humans. Sci Rep. 2017;7:1–15.

59. Grzenda A, Lomberk G, Zhang J-S, Urrutia R. Sin3: Master Scaffold and Transcriptional Corepressor Adrienne. Biochim Biophys Acta. 2009;0(1789):443–50.

60. Peng L, Li Y, Xi Y, Li W, Li J, Lv R, et al. MBD3L2 promotes Tet2 enzymatic activity for mediating 5-methylcytosine oxidation. J Cell Sci. 2016 Mar 1;129(5):1059–71.

61. Campbell AE, Shadle SC, Jagannathan S, Lim J-W, Resnick R, Tawil R, et al. NuRD and CAF-1-mediated silencing of the D4Z4 array is modulated by DUX4-induced MBD3L proteins. Elife. 2018 Mar 13;7.

62. Illumina. TruSeq Methyl Capture EPIC Library Prep Kit. 2017.

63. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011 Jun 1;27(11):1571–2.

64. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012 Oct 3;13(10):R87.

65. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41(1):200–9.

66. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013;9(8):1–10.

67. McLeay RC, Bailey TL. Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. BMC Bioinformatics. 2010;11.

68. Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48(D1):D87–92.

69. Conway JR, Lex A, Gehlenborg N. UpSetR: An R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33(18):2938–40.

70. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171–81.
71. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.
72. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Dec 4;12(1):323.
73. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma. 2020;2(3):1–10.
74. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.
75. Conesa A, Nueda M. maSigPro: Significant Gene Expression Profile Differences in Time Course Gene Expression Data. 2017.
76. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016 Sep 15;32(18):2847–9.
77. Hu H, Miao Y-R, Jia L-H, Yu Q-Y, Zhang Q, Guo A-Y. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019 Jan 8;47(D1):D33–8.
78. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
79. Jagannathan S, Shadle SC, Resnick R, Snider L, Tawil RN, van der Maarel SM, et al. Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. Hum Mol Genet. 2016 Aug 17;ddw271.
80. Rickard AM, Petek LM, Miller DG. Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. Hum Mol Genet. 2015 Jun 5;24(20):5901–14.

CHAPTER 4

Candidate regulators of DUX4 activated gene expression in FSHD2

Notes: (1) Nam Viet Nguyen cultured cells and performed depletions for this work.
(2) Dr. Xiangduo Kong cultured cells for this work.
(3) Elisabeth Rebboah contributed to the experimental work.
(4) Cassandra McGill and Heidi Liang performed the sequencing for this work.
(5) Dr. Rabi Tawil provided the resources for this work.
(6) Professor Kyoko Yokomori and Professor Ali Mortazavi contributed to the resources, supervision, writing and review of this work.

Chapter 4

Candidate regulators of DUX4 activated gene expression in FSHD2

4.1 Abstract

FSHD is caused by the misexpression of the transcription factor DUX4 which activates target gene expression including transcriptional regulators *DUXA*, *LEUTX* and *ZSCAN4*. *DUX4* expression is rare and burst like, but target gene expression persists after DUX4 is no longer present. To determine the role of DUX4-target transcription factors in perpetuating gene dysregulation, we deplete *DUXA*, *LEUTX* and *ZSCAN4* in differentiated patient derived myoblasts. Depletion of *DUXA* at day 6 of differentiation results in the downregulation of a large number of DUX4 target genes including long terminal repeats (LTRs). Depletion of *LEUTX* and *ZSCAN4* has moderate effects on aberrant gene expression. Regions regulated by SMCHD1 are more accessible in FSHD2 myoblasts, but regions near genes regulated by DUX4 are not. Motifs for PRD-like homeobox TFs, DUXA and DUX4 are enriched in regions less accessible in FSHD2 myoblasts. We find correlated differences in gene expression and accessibility in transcriptional regulators that may contribute to inherent differences between FSHD2 and control. We provide evidence of inherent transcriptional and chromatin differences in FSHD2 as well as further support for DUXA as a regulator of DUX4-induced FSHD dysregulated gene expression.

4.2 Introduction

Facioscapulohumeral muscular dystrophy (FSHD) is caused by the misexpression of the transcription factor DUX4 in skeletal muscle [1,2]. DUX4 is located at the end of a series of macrorepeats that are repressed with DNA methylation, H3K9me3 and PRC2 recruitment [3,4]. In FSHD1, which accounts for 95% of FSHD cases, these repeats are contracted and dereprepessed, thus leading to pathogenic expression of DUX4 [3]. In FSHD2, which makes up 5% of FSHD cases, about 80% of patients have a mutation in structural maintenance of chromosomes hinge domain 1 (SMCHD1), which regulates DNA methylation in X chromosome inactivation and in certain autosomal gene clusters [5–7]. SMCHD1 can repress expression in the D4Z4 clusters on chromosomes 4 and 10, the PCDH cluster on chromosome 5, as well as tRNA and rRNA clusters [7].

DUX4 is only expressed in 0.1% to 0.5% of patient muscle cells [8–10], but the protein is able to diffuse through the cytoplasm to activate target gene expression in multiple nuclei thereby amplifying its signal [11]. *DUX4* expression has also been proposed to be expressed in short bursts [10,11], but target gene expression is sustained after DUX4 is no longer present [9,11]. Two histone variants, H3.X and H3.Y, as well as DUXA, which are all DUX4 targets, have been proposed to play a role in regulating this continued expression [9,12]. DUX4 is normally expressed as part of zygotic genome activation when it activates expression of other transcriptional regulators such as *DUXA*, *LEUTX* and *ZSCAN4* [13]. DUXA and LEUTX are PRD-like homeobox transcription factors (TFs) that are similar to PRD TFs but lack the PRD domain [14]. ZSCAN4 can bind telomeres and microsatellite regions and prevent DNA damage [15,16]. These TFs are upregulated in FSHD but have not been shown to play a specific role in the disease.

To assess the contribution of the DUX4 targets DUXA, LEUTX and ZSCAN4 in perpetuating gene dysregulation in FSHD, we perform RNA-seq on primary patient-derived myoblast cell lines depleted of the respective TF. We also perform ATAC-seq on FSHD2 and control myoblasts to determine differences in the chromatin landscape due to disease. We find support for DUXA as a transcriptional regulator of FSHD-induced gene dysregulation, and we identify disease specific differences in chromatin accessibility and gene expression present before DUX4 activation.

4.3 Results

4.3.1 Depletion of *DUXA* but not *LEUTX* or *ZSCAN4* in FSHD2 myotubes results in less expression of DUX4 target genes

To determine the transcriptome-wide effects of transcription factors downstream of DUX4, we depleted *DUXA*, *LEUTX* and *ZSCAN4* in two FSHD2 cell lines at days 4 and 6 of differentiation followed by RNA-seq (Methods) (Figure S4.1A). We confirmed the successful knockdown of the respective targets and removed replicates that did not show significant depletion (Figure S4.1B). Depletion does not appear to have a significant wide-spread effect on gene expression (Figure S4.1C).

To identify genes regulated by the target TFs, we compared the depleted samples to those treated with control shRNA for the same days (Figure 4.1A, 4.1B, 4.1C, S4.2A, S4.2B, S4.2C). For all but *DUXA*, a higher number of genes are downregulated with depletion, which is consistent with the role of these TFs as transcriptional activators [17,18]. Interestingly, DUXA has been shown to act as a repressor, but only a handful of genes that were differentially expressed upon *DUXA* overexpression in human embryonic stem cells were differentially expressed in the myotubes upon *DUXA* depletion (Figure S4.1D) [19]. The TFs that we targeted are more highly expressed at day 6 than day 4 in FSHD cells. The higher number of affected genes in depletion at day 6 than day 4 suggests that these TFs may be regulating expression of genes upon their upregulation in these myotubes (Figure 4.1A, 4.1B, 4.1C, S4.2A, S4.2B, S4.2C). One set of *DUXA* depletions on day 6 induced strong upregulation of interferon signaling related genes such as *IFIT2* but this could be due to technical variation and not a result of the *DUXA* depletion (Figure S4.2D, Table S4.1). *MYOG* has lower expression in *DUXA* depleted cells at both day 4 and day 6 of differentiation, however other markers of myogenic differentiation are not significantly different (Figure 4.1A, S4.2A, S4.2E).

We had previously identified a set of 74 genes that are specifically upregulated in FSHD, most of which are DUX4 target genes (see Chapter 3). DUXA and LEUTX have been proposed to regulate a few of these genes in addition to DUX4 [9,20]. We find 40 of these genes downregulated upon *DUXA* depletion, including the previously identified *LEUTX* and *ZSCAN4* (Figure 4.1A) [9,14]. The promoters (-1.5 kb to +0.5 kb around TSS) of these genes are enriched for DUX4 and DUXA motifs (Figure 4.1D). All FSHD-induced genes that have a DUXA motif in the promoter also have either a DUX4 motif or a known DUX4 binding site identified by ChIP-seq (Figure 4.1E) [21]. An additional 8 genes downregulated upon DUXA depletion have either a DUX4 motif or DUX4 binding site in the promoter and thus may be regulated by DUX4. This includes genes previously identified as being upregulated in FSHD and/or in gene clusters with other FSHD-induced genes such as *HNRNPCL1* and *ZSCAN5B* (Figure 4.1E). DUXA thus appears to regulate genes upregulated by DUX4 in FSHD.

We do not observe significant changes in expression of FSHD-induced genes following depletion of *LEUTX* or *ZSCAN4* (Figure 4.1B, 4.1C, S4.2B, S4.2C). *LEUTX* depletion at day 4 had lower expression of *PRAMEF10* but not of the other PRAMEFs (Figure 4.1B). ZSCAN4 depletion at day 4 had higher expression of *TRIM48* and *TRIM49D1*, while at day 6 had higher expression of *TRIM49B*. *TRIM49B* and *TRIM48* are on either side of the centromere on chromosome 11, and Zscan4 regulates pericentromeric heterochromatin activation in mice [22]. ZSCAN4 may play a role in regulating genes in constitutive heterochromatic regions in FSHD.

4.3.2 FSHD2 myoblast have accessible chromatin in SMCHD1 regulated regions, but no differences for DUX4 target genes

We wanted to determine potential differences in chromatin accessibility in FSHD myoblasts compared to control to see whether inherent differences in the chromatin landscape could contribute to activation of target gene expression following *DUX4* expression. We find 2,978 regions more accessible in FSHD2, and 2,384 regions more accessible in control (Figure 4.2A). The D4Z4 region on chromosome 4 in FSHD has a loss of DNA methylation and repressive factors leading to its derepression [4]. SMCHD1 mutations in FSHD2 contribute to the derepression of D4Z4 repeats on both chromosomes 4 and 10 [7]. We find one region between *FRG2* and *DBET* in the D4Z4 region of chromosome 4 that is more accessible in FSHD2 than control (chr4:190049830-190049979) (Figure 4.2B, S4.3A). Another three regions near D4Z4 on chromosome 10 are also more accessible in FSHD2 (chr10:133649026-133649175, chr10:133616855-133617005, chr10:133614956-133615105) (Figure S4.3B). We see higher accessibility in other SMCHD1 regulated regions such as in the *SNRPN* cluster (chr15:23690436-23690585) and the *PCDHABG* cluster (chr5:141513662-141513899) (Figure

S4.3D, S4.3E). Interestingly, we also see less accessibility in the tRNA cluster (chr1:161612516-161612742) (Figure S4.3F).

DUX4 expression increases during myogenesis and therefore so does the expression of its target genes [9,11]. Derepression of the chromatin around *DUX4* has been shown in FSHD myoblasts preceding *DUX4* expression [23,24]. To understand whether FSHD-induced genes are accessible in myoblasts, we looked for differentially accessible regions associated with the genes. We detect 58 regions associated with 14 FSHD-induced genes. We do not see significant differences in accessibility for these regions with the exception of one region over 340 kb downstream of *RFPL4B* with higher accessibility in FSHD and one region upstream of *CCNA1* with less accessibility in FSHD2 (Figure 4.2A, 4.2C, S4.3C).

To determine what TFs could be binding to the differentially accessible regions (DARs), we performed motif enrichment analysis. Regions with more accessibility in control myoblasts were enriched for PRD-like homeodomain TF motifs (such as for OTX2, OTX1 and DPRX), DUXA and DUX4 motifs (Figure 4.3A). The day 0 myoblasts surveyed here do not express DUX4 or its target genes yet, so these regions should not be actively bound by DUX4, DUXA or any other PRD-like homeodomain TF. This may imply that while the D4Z4 region is more accessible in FSHD myoblasts, regions regulated by DUX4 and its target genes are not. To determine whether these regions change accessibility following DUX4 expression would require follow-up ATAC-seq experiments on later days of differentiation.

Interestingly, regions more accessible in FSHD than control are enriched for motifs for forkhead TFs such as FOXC2 and FOXL1 (Figure 4.3A). A large number of regions near *FOXC2* and *FOXL1* are more accessible in FSHD2 (Figure 4.3B, S4.3G). Also in this region is a lncRNA *FENDRR,* which acts at regulatory elements to increase binding of PRC2 [25]. PRC2 is responsible for maintaining repression of DUX4 [26]. *FENDRR* is also more highly expressed in FSHD2 at day 0 than control (Figure 4.3C). We see additional genes with higher expression and accessibility in FSHD2 including *EYA1* and *PRRX2* (Figure 4.3C). EYA1 is a TF involved in location specific muscle regulatory GRNs [27]. PRRX2 is a paired homeodomain TF, and its motif is also enriched in more accessible regions in FSHD2 (Figure 4.3A, 4.3C). We had previously found that a region near TGFA is hypomethylated in FSHD2 compared to control, and here we also see TGFA is less accessible and less expressed in FSHD2 (Figure 4.3C) (see Chapter 3). We also see decreased accessibility and expression of *FAT3* in FSHD2 (Figure 4.3C). Low *FAT1* expression was shown previously to mark muscles that are affected earlier in FSHD [28]. Part of the conclusions of that study were gathered through the use of an antibody which the authors point out has high sequence similarity with FAT3. Further studies would be needed to determine whether FAT3 has a similar trend as FAT1. By combining differential chromatin accessibility with differential gene expression, we have identified differences in FSHD2 myoblasts that precede *DUX4* expression.

4.3.3 Transposable elements regulated by DUX4 are downregulated upon DUXA depletion and are less accessible in FSHD2 myoblasts

DUX4, *DUXA*, *LEUTX* and *ZSCAN4* are all expressed during early embryogenesis along with a number of transposable elements (TEs) [13,18,29]. In mice, Zscan4 can bind to microsatellites and activate expression of ERVs in the early embryo [15,30]. DUX4 activates endogenous retroviruses (ERVs) during embryogenesis and in FSHD [13,31,32]. To determine whether DUXA, LEUTX or ZSCAN4 regulate ERVs or other TEs in FSHD myotubes, we looked at expression of TE loci following their respective depletions. Out of 80,271 TE loci detected, 1,008 were downregulated upon *DUXA* depletion, 1,024 upon *LEUTX* depletion, and 839 upon *ZSCAN4* depletion at day 6 of differentiation (Figure 4.4A).

ZSCAN4 depletion at day 6 results in significantly lower expression of 5 types of TEs including a type of ERV element called SAR and a satellite repeat (GAATG)n (Table S4.2). However, we do not observe significant regulation of ERVs beyond that with *ZSCAN4* depletion. ZSCAN4 regulation of ERVs may be specific either to mice or embryogenesis. Interestingly, 1,080 TEs are upregulated upon *LEUTX* depletion including significant enrichment of ERVL-MaLRs (p=2.19E-3) (Figure 4.4A, 4.4B). These loci are enriched for TF motifs including TP53 and PRD-like homeodomain TFs (Table S4.3, Figure S4.4A).

Notably, TEs with lower expression upon *DUXA* depletion are enriched for LTRs such as ERVL-MaLR (p=3.71E-36), ERVL (p=1.1E-5) and ERV1 (p=1.55E-3) (Figure 4.4B). The TEs are enriched for DUXA binding motifs with the majority (75 out of 86 TEs) in LTRs (Figure 4.4C). DUXA and DUX4 have similar motifs, and out of 86 loci with the DUXA motif, 45 overlap DUX4 binding sites identified by ChIP-seq [21]. Among these LTRs are specific classes shown to be upregulated by DUX4 including THE1D and MLT1D (Table S4.2) [21]. DUXA therefore may induce expression of TEs, especially LTRs, including some activated by DUX4.

To determine whether the differences in expression of TEs that we observe correlates with differences in chromatin accessibility, we overlapped the differentially accessible regions with TE loci. Regions more accessible in FSHD2 are enriched for DNA elements (3.13E-6), including TcMar-Tigger (p=5.62E-14), also L1 (5.66E-8) and telo (6.75E-3) elements (Table

S4.4). Surprisingly, regions less accessible in FSHD2 are enriched for LTR (p=7.65E-8), including ERVL-MaLR (p=1.29E-6) and Gypsy (1.17E-4) (Table S4.4).

4.3.4 Identification of a set of genes upregulated in FSHD2 during myogenesis

Weighted correlation network analysis (WGCNA) identifies clusters of genes with correlated expression that can be associated with traits [33]. We used WGCNA to find sets of genes that are expressed similarly either in FSHD2 or upon TF depletion. To identify a disease specific signature, we use RNA-seq from two control cell lines and two FSHD2 cell lines for days 0 to 5 and day 12 from chapters 2 and 3 in combination with the depleted samples. The genes cluster into 44 modules based on expression across all samples, and we correlate these modules with specificity for FSHD2 or depletion at day 6 (Methods). To identify a potential signature correlating with DUX4 expression, we use the number of FSHD-induced genes expressed in each sample, called FSHD gene score. Three clusters of interest positively correlate with the FSHD gene score, red, royalblue and darkmagenta, and two are negatively correlated, green and sienna3 (Figure 4.5A, 4.5B, S4.5A, S4.5B). The greenyellow module is associated with muscle differentiation including key markers of differentiation, such as *TTN* and *MYBPH*, but is not significantly associated with any of our observed traits confirming that variability between traits is not due to differences in differentiation (Figure 4.5A, 4.5B, 4.5C). Interestingly, the two modules negatively correlated with FSHD gene score include genes in clusters regulated by SMCHD1 including *PCDHGA5*, *PCDHGB3* and *NDN* (Figure 4.5C). The green cluster also includes many genes related to regulation of gene expression, such as regulators of DNA methylation *TET1* and *MBD1* (Figure 4.5C). The saddlebrown cluster is higher in FSHD2 overall but does not correlate with gene score and may therefore represent genes with inherently different expression in FSHD2 (Figure 4.5B). This includes genes related to signaling such as *LEF1* and *NOTCH3* (Figure 4.5C).

The modules with positive correlation with FSHD gene score, royalblue and darkmagenta, appear to represent pathways known to be dysregulated by DUX4 activation, such as apoptosis and nonsense-mediated decay (Figure 4.5C) [34]. Notably, the red module includes all 74 of the FSHD-induced genes and significantly overlaps DUX4 binding sites (p=1.75E-6) (Figure 4.5D). This module may represent an expanded set of genes regulated upon DUX4 activation. The red module is positively correlated with *LEUTX* and *ZSCAN4* depletion but not *DUXA* depletion suggesting that *DUXA* affects a wider set of genes regulated by DUX4 (Figure 4.5A). To identify the genes within this module that have similar expression profiles as the FSHD-induced genes, we use k-means clustering to get four clusters (Methods). The second cluster contains genes upregulated during myogenesis in FSHD2 including all the FSHD-induced genes (Figure 4.5D). This cluster, and the whole module, contains genes correlated with DUX4 expression in FSHD cells and upon overexpression of DUX4 that were not included in our previous gene set, such as *HNRNPCL1* and *ZSCAN5B* (Figure S4.5C, S4.5D) [11,35]. Within this cluster, we find additional genes that reside in seven location clustered genes upregulated in FSHD including two in the cluster of genes which contains the alternative histone H3.Y (AKA *RP11-432M8.17*) (chr5:17,603,903-17,656,380) and two in the PRAMEF cluster (chr1:12,765,200-13,429,387) (Table S4.5). The module therefore appears to represent an expanded set of genes misregulated in FSHD2.

4.4 Discussion

We have explored the gene regulatory networks inherent in FSHD2 myoblasts and those that are activated following DUX4 expression. We have provided additional support for DUXA as a regulator of FSHD specific gene expression and showed that depletion of *LEUTX* or *ZSCAN4* does not significantly affect expression of FSHD related genes. We find more accessibility in SMCHD1-mutated FSHD2 myoblasts in SMCHD1 regulated autosomal gene clusters. FSHD-induced genes were not more accessible in FSHD2, and regions with less accessibility in FSHD2 were enriched for motifs of disease related TFs such as DUX4, DUXA and the PRD-like homeobox TFs. Some differences in accessibility correlated with differences in gene expression, including a lncRNA that can regulate PRC2 binding. Finally, we find an expanded set of genes with similar expression profiles to those upregulated in FSHD2.

We and others have proposed that the sustained expression of DUX4 activated genes in FSHD is mediated through the transcriptional regulators that are downstream of DUX4 [9,12]. The alternative histones H3.X and H3.Y have already been shown to incorporate into DUX4 target genes to increase their sensitivity to further activation [12]. Zscan4 can activate genes in heterochromatic regions in mouse embryonic stem cells [22]. While we do not find many FSHD-induced genes affected by ZSCAN4 depletion, two of the three that we do find are near the centromere of chromosome 11. We speculate that ZSCAN4 mediated derepression of constitutive heterochromatin may enable DUX4 to activate target gene expression in otherwise inaccessible regions. In this case, the sustained *ZSCAN4* expression seen in FSHD may contribute to more widespread genome activation.

Despite its involvement in zygotic genome activation, *LEUTX* depletion does not alter expression of DUX4 induced genes in FSHD2 [18]. We had shown previously that *LEUTX* depletion resulted in lower expression of *KDM4E,* which we do not observe as significant here when combining the *LEUTX* knockdowns of two cell lines but is significant in *LEUTX* depletion of one cell line (data not shown) [20]. Our previous work used high resolution measurements of *KDM4E* RNA performed over individual myotubes. Our bulk RNA-seq may not pick up the decrease in *KDM4E* expression if for example not enough cells activate *DUX4* to begin with. Additional work involving overexpression or ChIP-seq would be useful in determining the role LEUTX plays in regulating DUX4 target genes.

Many studies have looked at DUX4 as one of the main drivers of gene expression in ZGA, especially for TEs. We find that TEs may also be controlled by DUXA, especially in FSHD. LEUTX and, contrary to previous findings in mice [30], ZSCAN4 depletions result in significantly more ERVL-MaLR expression. ZSCAN4 could have context specific regulation of TEs either specific to species or tissue/development. The lower accessibility of LTRs in FSHD2 myoblasts suggests that they may change accessibility during differentiation, but confirmation of this would require analysis from later differentiation days.

FENDRR, FOXC2 and FOXL1 had increased accessibility in FSHD2. FENDRR was shown to regulate expression of *FOXC2* and *FOXL1* by acting as an enhancer [36]. FOXC2 is important for proliferating satellite cells and repressing myogenesis [37]. In mouse development, FENDRR is also able to increase PRC2 at regulatory elements [25]. PRC2 has been proposed as the primary contributor to DUX4 repression [26]. In SMCHD1-mutated FSHD2, loss of SMCHD1 function results in derepression of D4Z4, but PRC2 partially compensates with

H3K27 trimethylation [38]. The increased accessibility and expression of FENDRR could be due to this compensatory activity, in which case we would not expect similar patterns in FSHD1.

The lower expression and accessibility of the signaling molecule *TGFA* and the cadherin *FAT3* may suggest a role for these in FSHD. TGFA plays a role in many tissues including regulating WNT signaling but is not well understood in skeletal muscle [39]. *FAT1* expression is higher in FSHD patients and in muscle groups more commonly affected in FSHD [28]. FAT1 and FAT3 have very similar sequences and similar expression patterns [40,41]. TGFA and FAT3 should be looked into further to determine whether they play a role in FSHD. In conclusion, we have partially elucidated the role of transcriptional regulators downstream of DUX4 and have identified inherent differences between FSHD2 and control that precede DUX4 expression.

4.5 Methods

4.5.1 Human myoblast culture and differentiation

Human control and FSHD2 myoblast cell lines were grown as previously described [9]. Cell were cultured on dishes coated with collagen in high glucose DMEM (Gibco) supplemented with 20% FBS (Omega Scientific, Inc.), 1% Pen-Strep (Gibco), and 2% Ultrasor G (Crescent Chemical Co.) at 10% CO2. Day 0 cells were kept at low confluency to prevent spontaneous differentiation. Upon reaching 80% confluence, differentiation was induced by using high glucose DMEM medium supplemented with 2% FBS and ITS supplement (insulin 0.1%, 0.000067% sodium selenite, 0.055% transferrin; Invitrogen). Fresh differentiation medium was changed every 24 hours.

4.5.2 shRNA transfection

Lentiviruses carrying the shRNA plasmids in Table 4.1 were made in 293T cells using Lipofectamine 3000. The cells were transfected with 2 ug of shRNA plasmids, 1.5 ug of pCMV plasmids, and 0.5 ug of pMP2G plasmids. The media was changed after 24 hours. The lentiviruses were harvested at 48 hours and 72 hours post-transfection. FSHD2 myoblasts were infected once at 32 hours and once at 8 hours prior to addition of differentiation media. The myoblasts were selected for plasmid integration using puromycin. RNA was extracted using RNeasy kit (74106, Qiagen) at days 4 and 6 of differentiation for transfected samples. shDUXA #3 was used for FSHD2-2 round 1. shDUXA #2 was used for FSHD2-2 round 2. shDUXA #3 was used for FSHD2-1 round 1.

4.5.3 RNA sequencing library preparation

RNA was converted to cDNA using the Smart-Seq2 protocol [42]. Libraries were constructed with the Nextera DNA Library Prep Kit (Illumina) for untreated FSHD2-2. The Nextera DNA Flex Library Prep Kit (Illumina) was used for all other RNA-seq samples. Libraries were sequenced on the Illumina NextSeq500 with paired-end 43 bp reads to a depth of 8 to 24 million reads per library.

4.5.4 RNA sequencing data processing

Raw reads from bulk RNA-seq were mapped to hg38 by STAR (version 2.5.1b) [43] using defaults except with a maximum of 10 mismatches per pair, a ratio of mismatches to read length of 0.07, and a maximum of 10 multiple alignments. Quantitation was performed using RSEM (version 1.2.31) [44] with defaults with gene annotations for protein coding genes and lncRNAs from GENCODE v28. Genes were filtered for 10 counts in all samples of the same differentiation day and shRNA target using filterByExpr from edgeR (version 3.30.3) [45]. TMM normalized counts from edgeR were used for differential expression analysis in edgeR. TMM normalized counts were TPM normalized for plotting using effective gene lengths for each sample calculated by RSEM. Heatmaps were created using ComplexHeatmap (version 2.4.3) [46]. Differentially expressed genes from day 0 were taken from processed data from (see Chapter 3).

4.5.5 TE mapping and processing

To map to TEs, fastqs were aligned as described above to the full GENCODE v28 annotation with a maximum of 100 multiple alignments and maximum of 100 loci anchors. Reads mapping to TE loci from repeatmasker from UCSC were estimated using featureCounts (subread version 2.0.1) [47] with fractional counts for multimapped reads for paired end reads. Genes were filtered for 2 counts in all samples of at least one condition (same shRNA target and differentiation day) using filterByExpr from edgeR (version 3.30.3) [45]. Counts were normalized as described above, and differential expression was performed in edgeR. Loci were overlapped with DUX4 binding sites from [21] using findOverlaps from GenomicRanges (version 1.40.0) [48]. Fisher's exact test (stats version 4.0.2) was used for enrichment of TE classes and DUX4 binding site overlaps with the "greater" alternative hypothesis.

4.5.6 Motif enrichment

Enrichment was performed using AME (meme, version 5.2.0) [49] on JASPAR 2020 core redundant motifs from human only [50]. Promoters were taken as -1.5kb to +0.5kb around the TSS for the gene model in the GENCODE gtf. Motif logos were generated with ceqlogo from meme-suite (version 5.2.0) [51].

4.5.7 ATAC sequencing library preparation

Control and FSHD2 cells from days 0 and 5 were collected for ATAC-seq (Cat. No. 53150, Active Motif). For day 0, 100,000 cells were used as input. For day 5, cells were lysed according to manufacturer's protocol [52] and brought up in 1X ice cold PBS. Nuclei were counted and normalized so 100,000 nuclei were used as input. Library preparation was performed according to manufacturer's protocol [52]. Libraries were sequenced on the Illumina NextSeq 500 with paired-end 40 bp reads to a depth of 14 to 33 million reads per library.

4.5.8 ATAC sequencing data processing

Fastqs were mapped to the mitochondrial genome from hg38 using bowtie2 [53] very-sensitive mode with up to 6 alignments reported, a maximum fragment length between reads of

3000 and no discordant alignments. Reads not aligned to the mitochondrial genome were mapped using the same parameters to hg38 canonical chromosomes and the patch region of D4Z4 (chr4_KQ983257v1_fix). Duplicate reads were removed using Picard (version 2.24.1) [54] MarkDuplicates. Reads were then shifted by +4 bps on the positive strand and -5 bps on the negative strand to adjust the 9 bp duplication artefact from Tn4 transposase [55]. Read coverage was calculated using bamCoverage from deeptools (version 3.5.0) [56] and visualized in the UCSC genome browser [57]. A tag directory was created for each sample using Homer (version 4.11.1) [58], and peaks were called using findPeaks with a size of 150 bp, a minimum distance of 50 and a local size of 50000. Homer-idr [59] was used to identify peaks which are reproducible across replicates for FSHD or control. Reproducible peaks were then overlapped with ENCODE blacklist regions for hg38 except for the D4Z4 regions on chromosomes 4 and 10 [60]. AnnotatePeaks from Homer was then used to summarize read counts in the peaks and annotate peaks with nearby genes.

Read counts were then filtered using edgeR for at least 10 reads in all FSHD or control samples. Counts were TMM normalized using edgeR, and differentially accessible regions were calculated with a p-value cutoff of <0.01 and a logFC threshold of 0.5. Overlaps with DUX4 binding sites from [21] and SMCHD1 regulated clusters were done using findOverlaps from GenomicRanges (version 1.40.0) [48] with the following coordinates for the given clusters; PCDHA/B/G chr5:140759009-141523383, SNRPN chr15:23548232-23697319, rRNA chr1:228552374-228653525, tRNA chr1:161395860-161624746. Overlaps with TEs was also done with findOverlaps with the TE loci from repeatmasker from UCSC [61]. Peak regions were correlated with gene expression using the gene IDs from differentially expressed genes from (see

Chapter 3) from day 0 FSHD2 tibialis anterior myoblasts compared with day 0 control tibialis anterior myoblasts. TE class enrichment was performed as described above.

4.5.9 WGCNA

The previously described samples were combined with untreated samples for days 0 to 5 and 12 of differentiation for two control cell lines and the two FSHD2 cell lines used for depletion taken from (see Chapter 3). Data was then batch corrected with days 0 to 5 for one of the FSHD2 cell lines as one batch and the remain samples as the other batch. Batch correction was done using CombatSeq from sva (version 3.36.0) [62]. Data normalization was then performed as described. TPM normalized counts were used as input to WGCNA (version 1.70-3) [33]. Outliers identified by hierarchical clustering were removed. These included one control day 0 sample, two control day 5 samples, a day 4 and a day 5 FSHD2 sample and the batch of DUXA depletion with upregulation of interferon related genes. The network was then constructed according to [63] with a power of 6. Enrichment analysis was performed using enrichR (version 2.1) [64] and filtered for a p-value cutoff of 0.05. FSHD gene score is the number of FSHD-induced genes with a TPM of greater than 1. Module and trait correlations and p-values were calculated using cor and corPvalueStudent, respectively, from WGCNA. K-means clustering of the red module was performed in ComplexHeatmap (version 2.4.3) [46].

4.5.10 Processing of publicly available data

Reanalyzed data from [11,35,65] was obtained from Chapter 3.

4.6 Figures

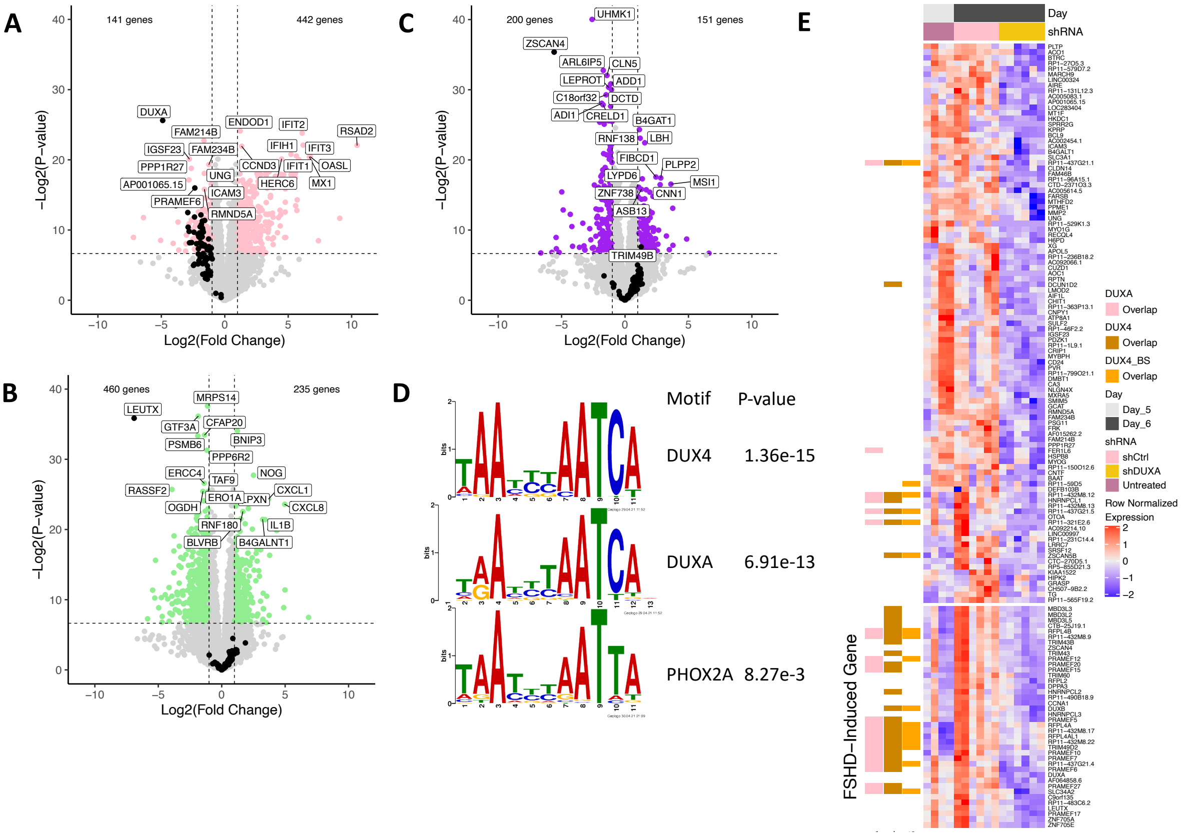


Figure 4.1: DUXA regulates similar genes as DUX4 (A) Volcano plot comparison of *DUXA* depletion to control shRNA at day 6 of differentiation. Genes with a p-value of <0.01 and absolute log2 fold change of greater than 2 are in pink. FSHD-induced genes identified from Chapter 3 are in black. The top 10 most significant genes are labelled. (B) Volcano plot comparison of *LEUTX* depletion to control shRNA at day 6 of differentiation. Genes with a p-value of <0.01 and absolute log2 fold change of greater than 2 are in green. FSHD-induced genes identified from Chapter 3 are in black. The top 10 most significant genes are labelled along with any significantly differentially expressed FSHD-induced genes. (C) Volcano plot comparison of *ZSCAN4* depletion to control shRNA at day 6 of differentiation. Genes with a p-value of <0.01 and absolute log2 fold change of greater than 2 are in purple. FSHD-induced genes identified from Chapter 3 are in black. The top 10 most significant genes are labelled along with any significantly differentially expressed FSHD-induced genes. (D) Motifs enriched in the promoters of genes with lower expression upon *DUXA* depletion at day 6 of differentiation. Logos generated using ceqlogo from meme-suite (version 5.2.0) [51]. (E) Heatmap of genes with lower expression upon *DUXA* depletion at day 6 of differentiation. The bottom 40 genes are also induced in FSHD during myogenesis as identified in Chapter 3. Genes marked in pink have the DUXA motif enriched in the promoter. Genes marked in gold have the DUX4 motif enriched in the promoter. Promoters of genes marked in orange overlap known DUX4 binding sites as identified by ChIP-seq.

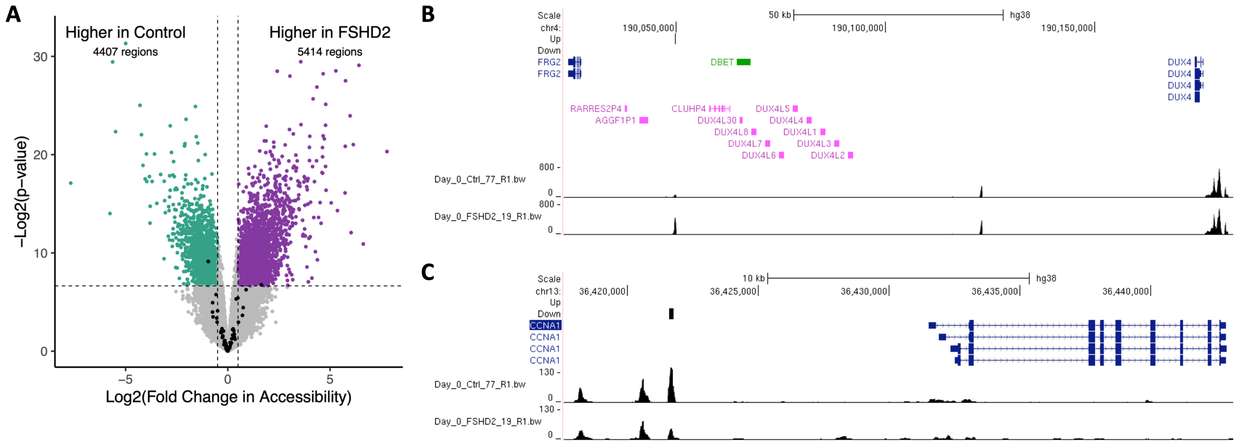


Figure 4.2: Lower chromatin accessibility in FSHD2 at D4Z4 repeats but not FSHD-induced genes (A) Volcano plot comparison chromatin accessibility in FSHD2 compared to control. In purple and green are regions with a p-value <0.01 and a log2 fold change greater than 0.5 or less than 0.5, respectively. Regions associated with FSHD-induced genes are in black. (B+C) UCSC genome browser shot of chromatin accessibility in the D4Z4 region of chromosome 4 (B) and near CCNA1 (C). Regions with significant accessibility changes are marked with a line in the rows labelled "Up" for higher accessibility in FSHD2 or "Down" for less accessibility in FSHD2. Two representative samples are shown. All samples are in figures S3A and S3B.

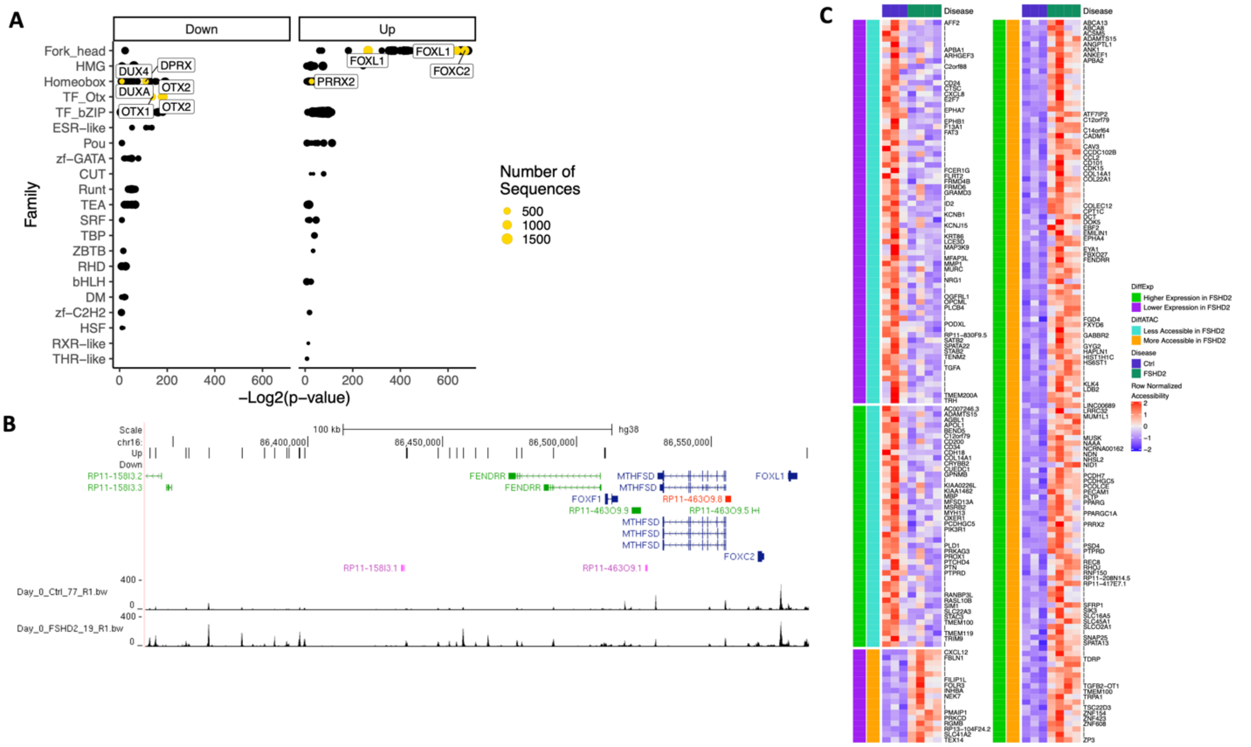


Figure 4.3: Higher chromatin accessibility in FSHD2 related to FOX TFs (A) Motifs enriched in regions more accessible (Up) or less accessible (Down) in FSHD2 compared with control. Motifs are grouped by family. Interesting motifs are labelled and colored in yellow. (B) UCSC genome browser shot of chromatin accessibility on chromosome 16 near three FOX genes and the lncRNA FENDRR. (C) Heatmap of chromatin accessibility for regions with significant differences in accessibility and corresponding gene expression. Regions for the same gene are labelled with a | following its first mention.

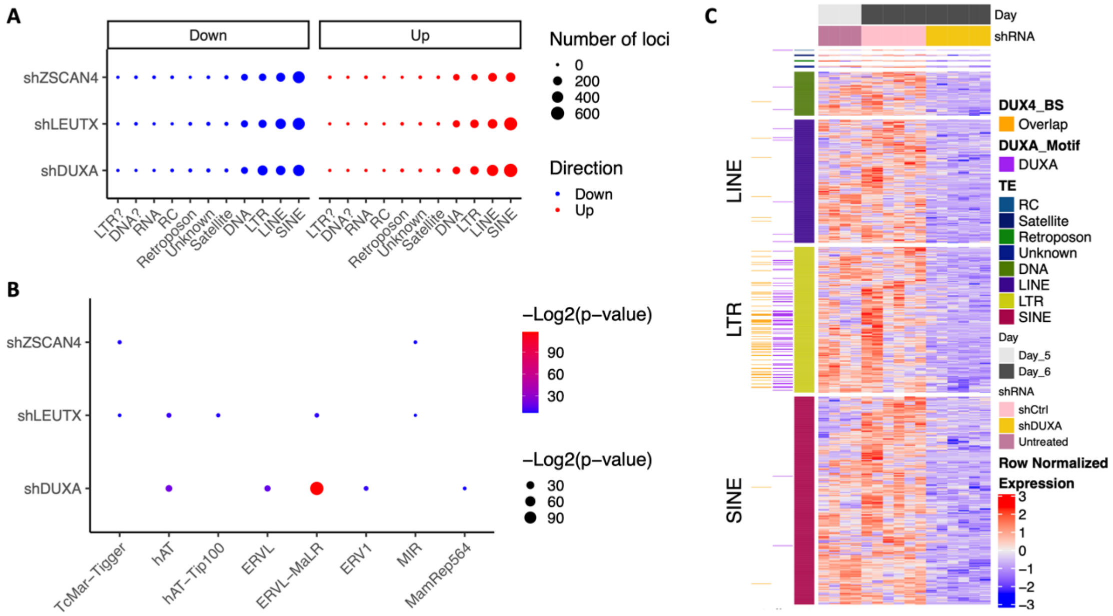
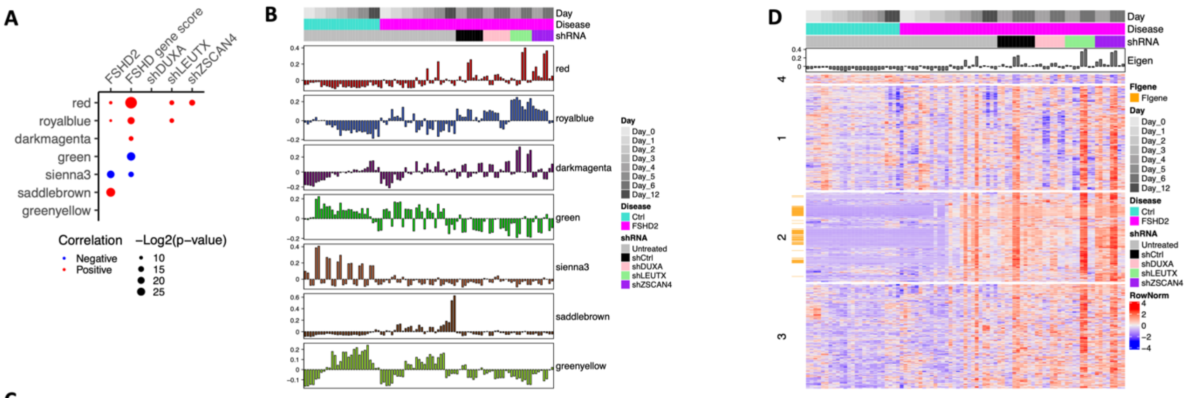


Figure 4.4: DUXA regulates LTRs regulated by DUX4 (A) Number of differentially expressed TE loci for the given depletions at day 6. Size represents number of loci. Loci with higher expression upon depletion are in red. Loci with lower expression upon depletion are in blue. (B) Enrichment of classes of TE loci differentially expressed upon depletion. P-values are from fisher's exact test. Correlations with a p-value of <0.01 are not shown. (C) Heatmap of expression of TE loci with lower expression upon DUXA depletion. Genes marked in purple have the DUXA motif enriched in the promoter. Promoters of genes marked in orange overlap known DUX4 binding sites as identified by ChIP-seq.



C

Module	Term	ID	DB	Adj P-value	Top genes
red	Gene Expression Homo sapiens	R-HSA-74160	Reactome_2016	9.13E-05	HNRNPF, ZNF256, DDX49, ZNF490, CCNA1
royalblue	Nonsense-Mediated Decay (NMD) Homo sapiens	R-HSA-927802	Reactome_2016	4.54E-71	RPL30, RPL32, RPLP1, RPL12, RPLP0
	Cellular Senescence Homo sapiens	R-HSA-2559583	Reactome_2016	4.27E-03	AGO4, H3F3A, HIST1H2BK, H2BFS, RPS27A
	positive regulation of myoblast differentiation	GO:0045663	GO_Biological_Process_2018	5.82E-03	MYF6, HIF1AN, MAPK14
	nucleosome organization	GO:0034728	GO_Biological_Process_2018	1.30E-02	SMARCC2, HIST1H2BK, H3F3A, H2BFS, NAP1L4
darkmagenta	Regulation of Apoptosis Homo sapiens	R-HSA-169911	Reactome_2016	4.33E-02	PSMB5, PSMC6
	p53-Independent DNA Damage Response Homo sapiens	R-HSA-69610	Reactome_2016	4.33E-02	PSMB5, PSMC6
green	Gene Expression Homo sapiens	R-HSA-74160	Reactome_2016	9.11E-05	RPL37, SRSF5, TET2, TET1, FOS
	regulation of transcription, DNA-templated	GO:0006355	GO_Biological_Process_2018	1.17E-02	PRDM2, CLOCK, WNT4, SIX4, MBD1
sienna3	chr5q31	NA	Chromosome_Location	1.56E-02	PCDHGA5, PCDHGB3, PCDHB16, PCDHB3
saddlebrown	negative regulation of signal transduction	GO:0009968	GO_Biological_Process_2018	2.85E-02	PDGFRA, NOTCH3, SFRP2, PEAR1, LGMN
	Binding of TCF/LEF:CTNNB1 to target gene promoters Homo sapiens	R-HSA-4411364	Reactome_2016	4.36E-02	LEF1, TCF7
greenyellow	muscle contraction	GO:0006936	GO_Biological_Process_2018	4.37E-24	TTN, MYBPH, MYH3, DES, DMD

Figure 4.5: Gene modules represent FSHD expression differences preceding and following DUX4 expression (A) Selected gene modules significantly correlated (p-value <0.01) with traits. Modules positively correlated with the given trait are marked in red, and negatively marked in blue. (B) Eigen values of the gene module for individual samples. Y-axis scale is unitless. (C) Gene ontology and pathway terms from GO and Reactome significantly associated (p-value <0.05) with the given gene module as determined by enrichR. Five representative genes are listed. (D) Heatmap of expression of genes in the red module. Eigen values for the module are represented as a barplot at the top. Genes previously identified as induced upon FSHD2 myogenesis are marked in orange.

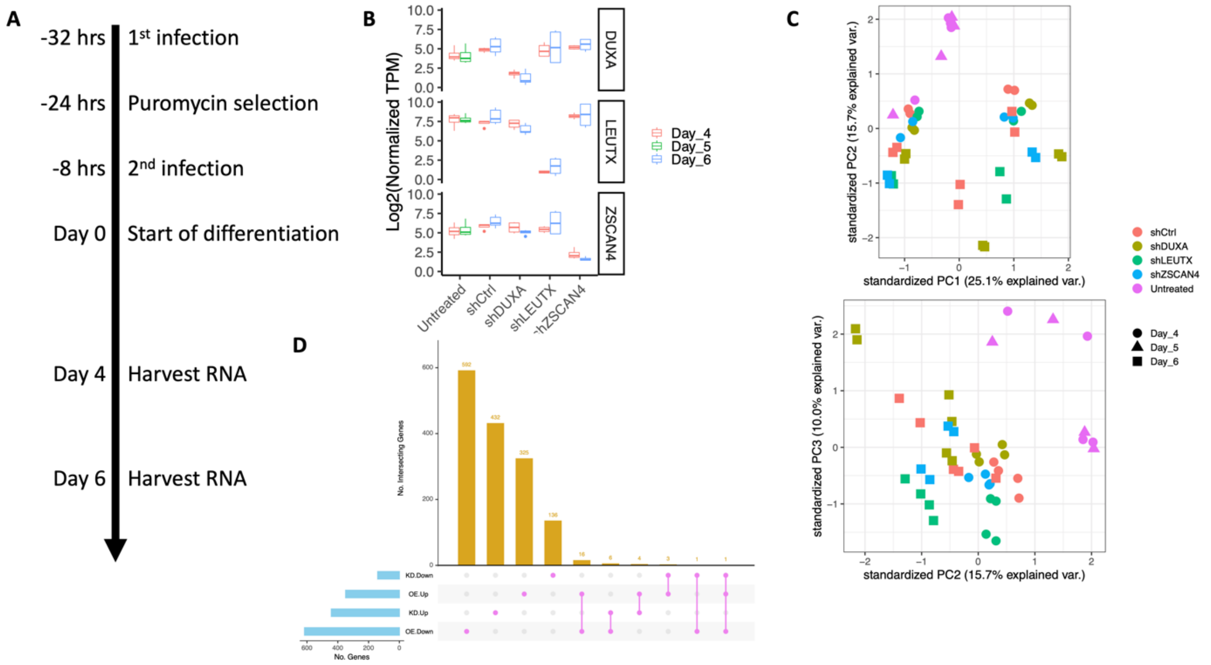


Figure S4.1: Data processing and exclusion (A) Timeline for transfection of shRNA into primary myoblasts. (B) Boxplot of expression of depleted genes. (C) Principal component analysis for components 1 and 2 (top) and 2 and 3 (bottom). Depletion targets are denoted by color, and differentiation day is denoted by shape. (D) Overlap of genes differentially expressed following DUXA depletion in myoblasts and DUXA overexpression in human embryonic stem cells [19]. Up denotes genes with higher expression upon depletion or over expression. Down denotes genes with lower expression upon depletion or over expression.

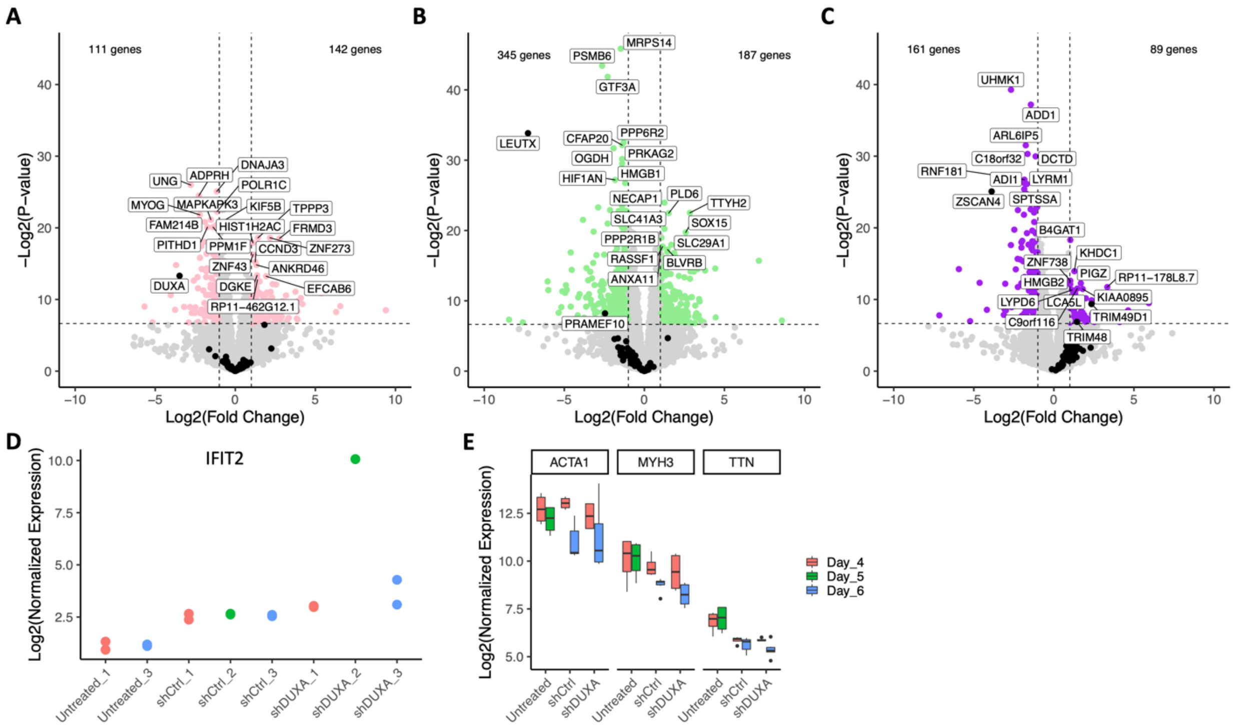


Figure S4.2: Depletion of *DUXA*, *LEUTX* and *ZSCAN4* (A) Volcano plot comparison of *DUXA* depletion to control shRNA at day 4 of differentiation. Genes with a p-value of <0.01 and absolute log2 fold change of greater than 2 are in pink. FSHD-induced genes identified from Chapter 3 are in black. The top 10 most significant genes are labelled along with any significantly differentially expressed FSHD-induced genes. (B) Volcano plot comparison of *LEUTX* depletion to control shRNA at day 4 of differentiation. Genes with a p-value of <0.01 and absolute log2 fold change of greater than 2 are in green. FSHD-induced genes identified from Chapter 3 are in black. The top 10 most significant genes are labelled along with any significantly differentially expressed FSHD-induced genes. (C) Volcano plot comparison of *ZSCAN4* depletion to control shRNA at day 4 of differentiation. Genes with a p-value of <0.01 and absolute log2 fold change of greater than 2 are in purple. FSHD-induced genes identified from Chapter 3 are in black. The top 10 most significant genes are labelled along with any significantly differentially expressed FSHD-induced genes. (D) IFIT2 expression in individual samples from batches of DUXA depletion (denoted as 1, 2, 3) from day 6 of differentiation. Untreated samples are from day 5 of differentiation. (E) Boxplot of myogenic marker genes for *DUXA* depletion and controls.

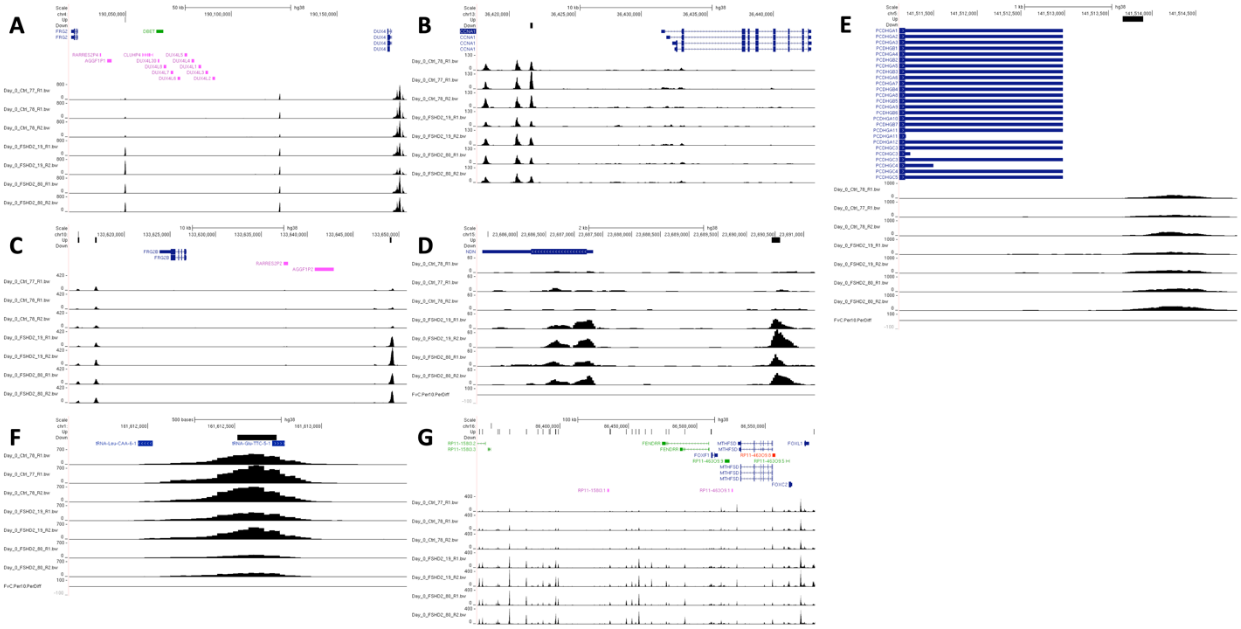


Figure S4.3: Chromatin accessibility at SMCDH1 regulated regions (A) UCSC genome browser shot from Figure 2B with all 7 samples. (B) UCSC genome browser shot from Figure 2C with all 7 samples. (C) UCSC genome browser shot of chromatin accessibility in the D4Z4 region of chromosome 10. Regions with significant accessibility changes are marked with a line in the rows labelled “Up” for higher accessibility in FSHD2 or “Down” for less accessibility in FSHD2. (D) UCSC genome browser shot of chromatin accessibility in the SNRPN gene cluster zoomed in to the DAR near the *NDN* gene. (E) UCSC genome browser shot of chromatin accessibility in the PCDH gene cluster on chromosome 5 zoomed in to the DAR which follows the 3’ exon of the PCDHgamma genes. (F) UCSC genome browser shot of chromatin accessibility in the tRNA gene cluster zoomed in to the DAR. (G) UCSC genome browser shot of the region on chromosome 16 with higher accessibility in FSHD2 than control.

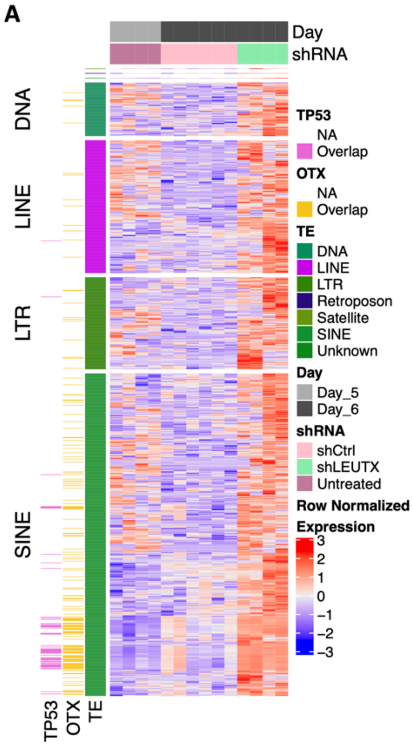


Figure S4.4: TEs with higher expression upon *LEUTX* depletion (A) Heatmap of row normalized expression of TE loci with higher expression following *LEUTX* depletion. The class of TE is shown in respective colors on the left of the plot. Loci enriched for an OTX motif (OTX1 or OTX2) are marked in gold. Loci enriched for the TP53 motif are marked in magenta.

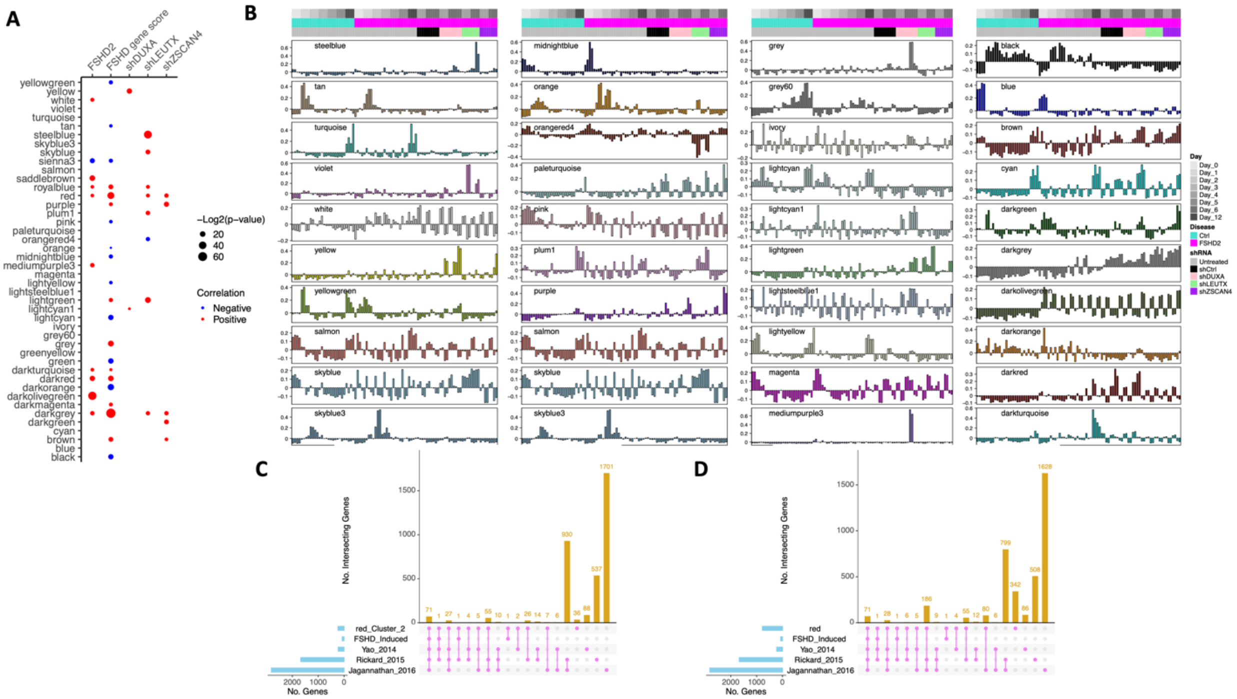


Figure S4.5: Data from all modules identified by WGCNA (A) Correlation of all gene modules with traits (p-value <0.01). Modules positively correlated with the given trait are marked in red, and negatively marked in blue. (B) Eigen values of all the gene module for individual samples. Y-axis scale is unitless. (C + D) Overlap of genes from (C) the red module subcluster 2 from Figure 5D or (D) the entire red module with genes upregulated in *in vitro* differentiation of FSHD myoblasts from Chapter 3 or from [65] (labelled Yao 2014), DUX4 expressing myocytes identified by a reporter (labelled Rickard 2015) [11], and overexpression of DUX4 (labelled Jagannathan 2016) [35].

4.7 Tables

Table 4.1: shRNA sequences used for TF depletion

shRNA	**Sequence**	**ID**
shDUXA #2	5'-CTAGATTACTTCTCCAGAGAA-3'	TRCN0000017664
shDUXA #3	5'- TCGAAGAGCTAGGCACGGATT-3'	TRCN0000017666
shLEUTX	5'-CCTGGAATCTCTGATGCAAAT-3'	TRCN0000336862
shZSCAN4	5'-CCCAAGATACTTCCTTAGAAA-3'	TRCN0000016848
shRNA Control		SHC002, Sigma-Aldrich

Table S4.1: GO terms for individual batches of *DUXA* depletion

	GO BP Term	Overlap	Adjusted.P.value	Top 5 Genes
Batch 2	positive regulation of cytokine production (GO:0001819)	34/220	5.83E-10	CEBPB, FLT4, GATA3, NOD2, PTGS2
	regulation of type I interferon production (GO:0032479)	21/85	9.89E-10	ZBP1, UBA7, DDX58. TREX1. NLRC5
	positive regulation of intracellular signal transduction (GO:1902533)	51/479	4.42E-09	CSF3, HIP1, FLT4, IFIT5, SECTM
Batch 1	None	None	None	None
Batch 3	None	None	None	None

Table S4.2: P-values for enrichment of TEs with lower expression following *DUXA*, *LEUTX* or *ZSCAN4* depletion at day 6 of differentiation Only loci with p-value <0.01 for any comparison are shown.

	shDUXA_Day6	shLEUTX_Day6	shZSCAN4_Day6
(GAATG)n	1.00E+00	7.41E-02	1.59E-03
AluSq	4.65E-01	3.33E-01	8.57E-03
ERVL-int	4.19E-03	1.00E+00	1.00E+00
HAL1b	4.88E-01	4.94E-01	2.31E-03
HERVH-int	1.04E-05	5.14E-02	2.79E-02
HERVIP10F-int	6.63E-03	1.00E+00	1.00E+00
L1MA9	8.68E-01	4.67E-03	4.99E-01
LTR1C1	1.00E+00	1.58E-03	5.12E-02
LTR3A	5.35E-03	1.00E+00	1.00E+00
LTR51	1.00E+00	2.03E-05	1.00E+00
LTR85a	6.63E-03	1.00E+00	1.00E+00
MamRep564	6.63E-03	1.20E-01	9.97E-02
MER112	9.13E-03	2.01E-01	4.95E-01
MER117	1.00E+00	3.81E-03	9.63E-02
MER5C1	4.19E-03	1.00E+00	1.00E+00
MIRb	3.28E-02	7.81E-02	7.05E-03
MIRc	8.73E-01	1.66E-03	1.86E-01
MLT1A0	6.90E-05	9.56E-01	9.22E-01
MLT1D	6.00E-07	6.88E-01	5.54E-01
MLT1E3	1.82E-05	1.00E+00	1.00E+00
MLT1F2	9.47E-03	7.30E-01	1.00E+00
MLT1L	1.00E-02	3.79E-02	7.80E-01
MLT1M	6.23E-03	1.00E+00	1.00E+00
MLT2A1	2.94E-07	3.33E-01	1.00E+00
MLT2A2	1.30E-02	3.97E-06	1.74E-01
SAR	1.00E+00	1.00E+00	1.09E-04
THE1C	9.19E-08	8.62E-01	4.79E-01
THE1C-int	1.19E-05	1.00E+00	1.00E+00
THE1D	2.49E-21	5.09E-01	8.93E-01
THE1D-int	2.38E-07	4.81E-01	1.00E+00
Tigger8	5.35E-03	1.00E+00	1.00E+00

Table S4.3: Motifs enriched in promoters of genes with higher expression upon *LEUTX* depletion at day 6 of differentiation

rank	motif_DB	motif_ID	motif_alt_ID	adj_p-value
1	JASPAR2020_HumanOnly	MA1579.1	ZBTB26	4.58E-43
2	JASPAR2020_HumanOnly	MA0106.1	TP53	1.99E-40
3	JASPAR2020_HumanOnly	MA0024.2	E2F1	1.56E-34
4	JASPAR2020_HumanOnly	MA0470.1	E2F4	1.36E-30
5	JASPAR2020_HumanOnly	MA0471.1	E2F6	3.48E-29
6	JASPAR2020_HumanOnly	MA0719.1	RHOXF1	1.74E-27
7	JASPAR2020_HumanOnly	MA0471.2	E2F6	4.14E-24
8	JASPAR2020_HumanOnly	MA1547.1	PITX2	9.24E-23
9	JASPAR2020_HumanOnly	MA0682.2	PITX1	1.31E-22
10	JASPAR2020_HumanOnly	MA1553.1	RARG(var.3)	2.53E-20
11	JASPAR2020_HumanOnly	MA1552.1	RARB(var.3)	2.97E-20
12	JASPAR2020_HumanOnly	MA1581.1	ZBTB6	9.95E-18
13	JASPAR2020_HumanOnly	MA0014.3	PAX5	3.07E-17
14	JASPAR2020_HumanOnly	MA1555.1	RXRB(var.2)	3.96E-17
15	JASPAR2020_HumanOnly	MA0712.2	OTX2	1.94E-15
16	JASPAR2020_HumanOnly	MA0711.1	OTX1	2.31E-14
17	JASPAR2020_HumanOnly	MA1122.1	TFDP1	2.65E-14
18	JASPAR2020_HumanOnly	MA0469.1	E2F3	3.02E-13
19	JASPAR2020_HumanOnly	MA0478.1	FOSL2	5.58E-13
20	JASPAR2020_HumanOnly	MA0516.1	SP2	1.28E-09
21	JASPAR2020_HumanOnly	MA0714.1	PITX3	4.74E-09
22	JASPAR2020_HumanOnly	MA0099.1	JUN::FOS	6.62E-09
23	JASPAR2020_HumanOnly	MA0712.1	OTX2	7.80E-09
24	JASPAR2020_HumanOnly	MA0112.2	ESR1	2.82E-07
25	JASPAR2020_HumanOnly	MA0891.1	GSC2	3.26E-07
26	JASPAR2020_HumanOnly	MA1142.1	FOSL1::JUND	4.79E-07
27	JASPAR2020_HumanOnly	MA0648.1	GSC	2.86E-06
28	JASPAR2020_HumanOnly	MA0528.2	ZNF263	5.03E-06
29	JASPAR2020_HumanOnly	MA0154.1	EBF1	2.14E-05
30	JASPAR2020_HumanOnly	MA1137.1	FOSL1::JUNB	2.83E-05
31	JASPAR2020_HumanOnly	MA0637.1	CENPB	5.73E-04
32	JASPAR2020_HumanOnly	MA1513.1	KLF15	6.71E-04
33	JASPAR2020_HumanOnly	MA1650.1	ZBTB14	1.84E-03
34	JASPAR2020_HumanOnly	MA1132.1	JUN::JUNB	2.33E-03
35	JASPAR2020_HumanOnly	MA0258.1	ESR2	3.15E-03

36	JASPAR2020_HumanOnly	MA0098.1	ETS1	7.29E-03
37	JASPAR2020_HumanOnly	MA1516.1	KLF3	8.55E-03

Table S4.4: P-values for enrichment of TEs with differential accessibility in FSHD2 myoblasts compared to control Only TEs types with p-value <0.01 are shown.

	HigherInControl	HigherInFSHD2
Charlie3	1.00E+00	6.74E-03
ERVL-MaLR	1.29E-06	2.09E-01
Gypsy	1.17E-04	9.44E-01
L1	4.52E-01	5.66E-08
L1MA8	2.60E-01	3.58E-03
L1MB4	6.89E-01	6.91E-03
L1MB7	9.49E-01	3.78E-03
L1MC1	9.07E-01	7.80E-04
L1ME3A	9.80E-01	2.18E-04
L1PA8	5.64E-03	1.00E+00
LTR14B	1.06E-03	1.00E+00
LTR18C	1.00E+00	6.54E-03
LTR1D	1.00E+00	3.35E-03
LTR26	3.39E-04	5.59E-01
LTR81B	7.30E-03	1.00E+00
MER11D	2.99E-03	1.00E+00
MER1B	6.57E-01	4.73E-03
MER58A	9.55E-01	3.57E-06
MER66-int	6.04E-04	1.00E+00
MER81	2.94E-03	6.06E-01
MER95	1.00E+00	3.35E-03
TcMar-Tigger	1.00E+00	5.62E-14
telo	1.00E+00	6.74E-03
THE1B	1.28E-04	8.08E-01
Tigger2b_Pri	2.06E-03	1.00E+00
Tigger3	7.60E-01	3.47E-03
Tigger3b	9.99E-01	7.06E-26
Tigger3c	1.00E+00	1.02E-07

Table S4.5: Genes from the red module which are clustered together on the chromosome Genes in the same region have the same cluster number. Genes which are in the FSHD-induced gene list from Chapter 3 are labelled “FIgene”.

chr	start	cluster	GeneName	GeneFull	FIgene
chr1	9922113	1	LZIC	ENSG00000162441.11_LZIC	
chr1	9997206	1	RBP7	ENSG00000162444.11_RBP7	
chr1	12774841	2	PRAMEF12	ENSG00000116726.4_PRAMEF12	FIgene
chr1	12791397	2	PRAMEF1	ENSG00000116721.9_PRAMEF1	FIgene
chr1	12824605	2	PRAMEF11	ENSG00000239810.3_PRAMEF11	FIgene
chr1	12847408	2	HNRNPCL1	ENSG00000179172.9_HNRNPCL1	
chr1	12857086	2	PRAMEF2	ENSG00000120952.4_PRAMEF2	FIgene
chr1	12879224	2	PRAMEF4	ENSG00000243073.3_PRAMEF4	FIgene
chr1	12892896	2	PRAMEF10	ENSG00000187545.5_PRAMEF10	FIgene
chr1	12916610	2	PRAMEF7	ENSG00000204510.5_PRAMEF7	FIgene
chr1	12938472	2	PRAMEF6	ENSG00000232423.6_PRAMEF6	FIgene
chr1	13049476	3	PRAMEF27	ENSG00000274764.5_PRAMEF27	FIgene
chr1	13060869	3	HNRNPCL3	ENSG00000277058.2_HNRNPCL3	FIgene
chr1	13068677	3	PRAMEF25	ENSG00000229571.7_PRAMEF25	FIgene
chr1	13115496	3	HNRNPCL2	ENSG00000275774.2_HNRNPCL2	FIgene
chr1	13148905	3	PRAMEF26	ENSG00000280267.4_PRAMEF26	FIgene
chr1	13164586	3	HNRNPCL4	ENSG00000179412.10_HNRNPCL4	
chr1	13172455	3	PRAMEF9	ENSG00000204505.4_PRAMEF9	FIgene
chr1	13196330	3	PRAMEF13	ENSG00000279169.2_PRAMEF13	FIgene
chr1	13222705	3	PRAMEF18	ENSG00000279804.2_PRAMEF18	FIgene
chr1	13254212	3	PRAMEF5	ENSG00000270601.4_PRAMEF5	FIgene
chr1	13281035	3	PRAMEF8	ENSG00000182330.10_PRAMEF8	FIgene
chr1	13303539	3	RP11-219C24.6	ENSG00000237700.2_RP11-219C24.6	FIgene
chr1	13315581	3	PRAMEF15	ENSG00000204501.7_PRAMEF15	FIgene
chr1	13342034	3	PRAMEF14	ENSG00000204481.7_PRAMEF14	FIgene
chr1	13369067	3	PRAMEF19	ENSG00000204480.7_PRAMEF19	FIgene
chr1	13389632	3	PRAMEF17	ENSG00000204479.4_PRAMEF17	FIgene
chr1	13410450	3	PRAMEF20	ENSG00000204478.9_PRAMEF20	FIgene
chr1	44118850	6	KLF17	ENSG00000171872.4_KLF17	FIgene
chr1	44137821	6	KLF18	ENSG00000283039.1_KLF18	FIgene
chr11	4233288	26	RP11-437G21.4	ENSG00000284018.1_RP11-437G21.4	FIgene
chr11	4242056	26	RP11-437G21.2	ENSG00000284306.1_RP11-437G21.2	FIgene
chr11	4287499	26	RP11-437G21.3	ENSG00000283873.1_RP11-437G21.3	FIgene
chr11	4329865	26	RP11-437G21.5	ENSG00000284546.1_RP11-437G21.5	

chr11	4338660	26	RP11-437G21.1	ENSG00000284438.1_RP11-437G21.1	
chr11	89797655	34	TRIM49	ENSG00000168930.13_TRIM49	FIgene
chr11	89869282	34	TRIM64B	ENSG00000189253.7_TRIM64B	FIgene
chr11	89911111	34	TRIM49D1	ENSG00000223417.8_TRIM49D1	FIgene
chr11	89924064	34	TRIM49D2	ENSG00000233802.8_TRIM49D2	FIgene
chr11	90031106	35	TRIM49L2	ENSG00000204449.3_TRIM49L2	FIgene
chr11	90085950	35	UBTFL1	ENSG00000255009.4_UBTFL1	FIgene
chr11	95025258	36	KDM4E	ENSG00000235268.2_KDM4E	FIgene
chr11	95049422	36	RP11-735A19.2	ENSG00000255855.2_RP11-735A19.2	FIgene
chr11	125893485	39	PUS3	ENSG00000110060.8_PUS3	
chr11	125903348	39	DDX25	ENSG00000109832.13_DDX25	
chr16	75660227	67	RP11-490B18.9	ENSG00000284484.1_RP11-490B18.9	FIgene
chr16	75693929	67	DUXB	ENSG00000282757.3_DUXB	FIgene
chr16	75714224	67	CPHXL	ENSG00000283755.1_CPHXL	
chr18	31942575	74	RP11-326K13.4	ENSG00000263823.1_RP11-326K13.4	
chr18	32018372	74	RNF125	ENSG00000101695.8_RNF125	
chr18	32018829	74	RP11-53I6.2	ENSG00000263917.1_RP11-53I6.2	
chr19	7018969	76	CTB-25J19.1	ENSG00000196589.6_CTB-25J19.1	FIgene
chr19	7030578	76	MBD3L5	ENSG00000237247.6_MBD3L5	FIgene
chr19	7049321	76	MBD3L2	ENSG00000230522.5_MBD3L2	FIgene
chr19	7056209	76	MBD3L3	ENSG00000182315.9_MBD3L3	FIgene
chr19	39731250	80	CLC	ENSG00000105205.6_CLC	
chr19	39776595	80	LEUTX	ENSG00000213921.7_LEUTX	FIgene
chr19	55759014	83	RFPL4A	ENSG00000223638.3_RFPL4A	FIgene
chr19	55769141	83	RFPL4AL1	ENSG00000229292.1_RFPL4AL1	FIgene
chr19	56189570	84	ZSCAN5B	ENSG00000197213.9_ZSCAN5B	
chr19	56202301	84	ZSCAN5C	ENSG00000204532.6_ZSCAN5C	
chr19	57119138	85	USP29	ENSG00000131864.10_USP29	
chr19	57134096	85	ZIM3	ENSG00000141946.1_ZIM3	FIgene
chr19	57154021	85	DUXA	ENSG00000258873.2_DUXA	FIgene
chr2	18555545	88	NT5C1B-RDH14	ENSG00000250741.6_NT5C1B-RDH14	
chr2	18562872	88	NT5C1B	ENSG00000185013.16_NT5C1B	
chr2	232378534	100	ALPP	ENSG00000163283.6_ALPP	
chr2	232406843	100	ALPPL2	ENSG00000163286.8_ALPPL2	FIgene
chr5	17604177	129	RP11-432M8.22	ENSG00000283740.1_RP11-432M8.22	FIgene
chr5	17610496	129	RP11-432M8.9	ENSG00000249156.2_RP11-432M8.9	FIgene
chr5	17632088	129	RP11-432M8.12	ENSG00000283776.1_RP11-432M8.12	
chr5	17634460	129	RP11-432M8.13	ENSG00000250782.1_RP11-432M8.13	

chr5	17654870	129	RP11-432M8.17	ENSG00000269466.3_RP11-432M8.17	FIgene
chr6	73209746	142	RP11-257K9.8	ENSG00000243501.5_RP11-257K9.8	FIgene
chr6	73223544	142	KHDC1L	ENSG00000256980.4_KHDC1L	FIgene
chr8	7355517	163	ZNF705G	ENSG00000215372.6_ZNF705G	
chr8	7428888	163	DEFB103B	ENSG00000177243.3_DEFB103B	

4.8 References

1. Snider L, Geng LN, Lemmers RJLF, Kyba M, Ware CB, Nelson AM, et al. Facioscapulohumeral Dystrophy: Incomplete Suppression of a Retrotransposed Gene. Pearson CE, editor. PLoS Genet. 2010 Oct 28;6(10):e1001181.
2. Vanderplanck C, Ansseau E, Charron S, Stricwant N, Tassin A, Laoudj-Chenivesse D, et al. The FSHD Atrophic Myotube Phenotype Is Caused by DUX4 Expression. Chadwick BP, editor. PLoS One. 2011 Oct 28;6(10):e26820.
3. Statland JM, Tawil R. Facio-scapulo-humeral muscular dystrophy. Indian J Pediatr. 2008;34(5):186–8.
4. Himeda CL, Jones TI, Jones PL. Facioscapulohumeral muscular dystrophy as a model for epigenetic regulation and disease. Antioxid Redox Signal. 2015 Jun 1;22(16):1463–82.
5. Wang C-Y, Brand H, Shaw ND, Talkowski ME, Lee JT. Role of the Chromosome Architectural Factor SMCHD1 in X-Chromosome Inactivation, Gene Regulation, and Disease in Humans. Genetics. 2019 Oct 1;213(2):685–703.
6. Lemmers RJLF, Tawil R, Petek LM, Balog J, Block GJ, Santen GWE, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat Genet. 2012 Dec;44(12):1370–4.
7. Mason AG, Slieker RC, Balog J, Lemmers RJLF, Wong CJ, Yao Z, et al. SMCHD1 regulates a limited set of gene clusters on autosomal chromosomes. Skelet Muscle. 2017 Jun 6;7(1).
8. Tassin A, Laoudj-Chenivesse D, Vanderplanck C, Barro M, Charron S, Ansseau E, et al. DUX4 expression in FSHD muscle cells: How could such a rare protein cause a myopathy? J Cell Mol Med. 2013 Jan;17(1):76–89.
9. Williams K, Jiang S, Kong X, Zeng W, Nguyen NV, Ma X, et al. Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. PLoS Genet. 2020;16(5):1–26.
10. van den Heuvel A, Mahfouz A, Kloet SL, Balog J, van Engelen BGM, Tawil R, et al. Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. Hum Mol Genet. 2018 Nov 16;
11. Rickard AM, Petek LM, Miller DG. Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. Hum Mol Genet. 2015 Jun 5;24(20):5901–14.
12. Resnick R, Wong C-J, Hamm DC, Bennett SR, Skene PJ, Hake SB, et al. DUX4-Induced Histone Variants H3.X and H3.Y Mark DUX4 Target Genes for Expression. Cell Rep. 2019 Nov 12;29(7):1812-1820.e5.
13. Hendrickson PG, Doráis JA, Grow EJ, Whiddon JL, Lim JW, Wike CL, et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat Genet. 2017 Jun 1;49(6):925–34.
14. Töhönen V, Katayama S, Vesterlund L, Jouhilahti E-M, Sheikhi M, Madissoon E, et al. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun. 2015 Nov 11;6(1):8207.
15. Srinivasan R, Nady N, Arora N, Hsieh LJ, Swigut T, Narlikar GJ, et al. Zscan4 binds nucleosomal microsatellite DNA and protects mouse two-cell embryos from DNA damage. Sci Adv. 2020;6(12).
16. Zalzman M, Falco G, Sharova L V., Nishiyama A, Thomas M, Lee SL, et al. Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature. 2010;464(7290):858–63.

17. Hirata T, Amano T, Nakatake Y, Amano M, Piao Y, Hoang HG, et al. Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells. Sci Rep. 2012;2:1–11.
18. Jouhilahti E-M, Madissoon E, Vesterlund L, Töhönen V, Krjutškov K, Plaza Reyes A, et al. The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation . Development. 2016 Aug 31;143(19):3459–69.
19. Madissoon E, Jouhilahti EM, Vesterlund L, Töhönen V, Krjutškov K, Petropoulous S, et al. Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci Rep. 2016 Jul 14;6.
20. Chau J, Kong X, Viet Nguyen N, Williams K, Ball M, Tawil R, et al. Relationship of DUX4 and target gene expression in FSHD myocytes. Hum Mutat. 2021;
21. Young JM, Whiddon JL, Yao Z, Kasinathan B, Snider L, Geng LN, et al. DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. PLoS Genet. 2013 Nov;9(11).
22. Akiyama T, Xin L, Oda M, Sharov AA, Amano M, Piao Y, et al. Transient bursts of Zscan4 expression are accompanied by the rapid derepression of heterochromatin in mouse embryonic stem cells. DNA Res. 2015;22(5):307–18.
23. Zeng W, De Greef JC, Chen YY, Chien R, Kong X, Gregson HC, et al. Specific loss of histone H3 lysine 9 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). PLoS Genet. 2009 Jul;5(7).
24. Zeng W, Chen YY, Newkirk DA, Wu B, Balog J, Kong X, et al. Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs. Hum Mutat. 2014;35(8):998–1010.
25. Grote P, Herrmann BG. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. 2013;(October):1579–85.
26. Haynes P, Bomsztyk K, Miller DG. Sporadic DUX4 expression in FSHD myocytes is associated with incomplete repression by the PRC2 complex and gain of H3K9 acetylation on the contracted D4Z4 allele. Epigenetics and Chromatin. 2018 Aug 20;11(1).
27. Buckingham M, Rigby PWJ. Gene Regulatory Networks and Transcriptional Mechanisms that Control Myogenesis. Vol. 28, Developmental Cell. 2014. p. 225–38.
28. Mariot V, Roche S, Hourdé C, Portilho D, Sacconi S, Puppo F, et al. Correlation between low FAT1 expression and early affected muscle in facioscapulohumeral muscular dystrophy. Ann Neurol. 2015;78(3):387–400.
29. Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L, et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nat Commun. 2019 Dec 21;10(1):364.
30. Zhang W, Chen F, Chen R, Xie D, Yang J, Zhao X, et al. Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. Nucleic Acids Res. 2019;47(16):8485–501.
31. Geng LN, Yao Z, Snider L, Fong AP, Cech JN, Young JM, et al. DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy. Dev Cell. 2012 Jan 17;22(1):38–51.
32. Whiddon JL, Langford AT, Wong CJ, Zhong JW, Tapscott SJ. Conservation and innovation in the DUX4-family gene network. Nat Genet. 2017 Jun 1;49(6):935–40.
33. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network

analysis. BMC Bioinformatics. 2008;9.

34. Campbell AE, Belleville AE, Resnick R, Shadle SC, Tapscott SJ. Facioscapulohumeral dystrophy: activating an early embryonic transcriptional program in human skeletal muscle. Hum Mol Genet. 2018 Aug 1;27(R2):R153–62.
35. Jagannathan S, Shadle SC, Resnick R, Snider L, Tawil RN, van der Maarel SM, et al. Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. Hum Mol Genet. 2016 Aug 17;ddw271.
36. Liu H, Zhang Z, Han Y, Fan A, Liu H, Zhang X, et al. The FENDRR/FOXC2 Axis Contributes to Multidrug Resistance in Gastric Cancer and Correlates With Poor Prognosis. Front Oncol. 2021;11(March):1–14.
37. Gozo MC, Aspuria PJ, Cheon DJ, Walts AE, Berel D, Miura N, et al. Foxc2 induces Wnt4 and Bmp4 expression during muscle regeneration and osteogenesis. Cell Death Differ. 2013;20(8):1031–42.
38. Balog J, Thijssen PE, Shadle S, Straasheijm KR, van der Vliet PJ, Krom YD, et al. Increased DUX4 expression during muscle differentiation correlates with decreased SMCHD1 protein levels at D4Z4. Epigenetics. 2015 Dec 2;10(12):1133–42.
39. Singh B, Coffey RJ. From wavy hair to naked proteins: The role of transforming growth factor alpha in health and disease. Bone. 2008;23(1):1–7.
40. Zhang X, Liu J, Liang X, Chen J, Hong J, Li L, et al. History and progression of fat cadherins in health and disease. Onco Targets Ther. 2016;9:7337–43.
41. Tanoue T, Takeichi M. New insights into fat cadherins. J Cell Sci. 2005;118(11):2347–53.
42. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171–81.
43. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.
44. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Dec 4;12(1):323.
45. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.
46. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016 Sep 15;32(18):2847–9.
47. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
48. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013;9(8):1–10.
49. McLeay RC, Bailey TL. Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. BMC Bioinformatics. 2010;11.
50. Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48(D1):D87–92.
51. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. 2009;37(SUPPL. 2):202–8.
52. Active Motif I. ATAC-Seq Kit Manual Catalog No. 53150 (version B5). Vol. 53150. 2020.
53. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.

2012;9(4):357–9.
54. Broad Institute. Picard [Internet]. Available from: http://broadinstitute.github.io/picard/
55. Berg DE, Schmandt MA, Lowe JB. Specificity of transposon Tn5 insertion. Genetics. 1983;105(4):813–28.
56. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44(W1):W160–5.
57. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12(6):996–1006.
58. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28;38(4):576–89.
59. Allison K. homer-idr: Second pass updated.
60. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019;9(1):1–5.
61. Jurka J. Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet. 2000;16(9):418–20.
62. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma. 2020;2(3):1–10.
63. Langfelder P, Horvath S. Tutorial for the WGCNA package for R: I. Network analysis of liver expression data in female mice. 2014;1–5.
64. Jawaid W. enrichR: Provides an R Interface to “Enrichr.” R package version 2.1. 2019.
65. Yao Z, Snider L, Balog J, Lemmers RJLF, Van Der Maarel SM, Tawil R, et al. DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. Hum Mol Genet. 2014 Oct 15;23(20):5342–52.

CHAPTER 5

Summary and Conclusions

Chapter 5

Summary and Conclusions

5.1 Introduction

In the previous chapters I have explored the molecular basis for susceptibility and progression of FSHD but have only provided a partial picture. I have surveyed four muscle groups, but we still do not fully understand the mechanisms by which expression of TFs used during muscle development is sustained in adult tissue or why. Further understanding the molecular basis can guide our search for treatments and biomarkers. I have examined the roles of a handful of DUX4 target gene TFs in FSHD gene dysregulation, but how this gene dysregulation functionally contributes to pathogenesis is unclear, making identification of therapeutic targets difficult as well as limiting the usefulness of disease models. However, current high resolution sequencing technologies could be used in FSHD and other systems to better understand the molecular mechanisms underlying disease progression.

5.2 Muscle group specific gene regulatory networks

Location and muscle group specific GRNs are best understood in embryogenesis of model systems but less studied in developed tissue including muscle. I have provided evidence of sustained location specific DNA modifications and gene expression of developmentally associated TFs in myoblasts from adult human tissue, and others have provided similar evidence in model organisms. Aside from DNA methylation, the maintained expression of these TFs may be in part regulated by extracellular signaling as transplantation studies in mice have found that transplanted satellite cells express genes specific to their new location [1]. *In vitro* experiments with satellite cells could test the role of potential signals by measuring associated expression changes. A comprehensive profiling survey of gene expression, chromatin accessibility and histone marks would inform the epigenetic differences of muscle groups throughout the body and help us to understand location specific postnatal gene regulation. Importantly, this work could reveal key differences that contribute to susceptibility to myopathies.

The sustained expression of developmental TFs in adult tissue suggests a potential functional role. Expression of HOX genes in other adult organs can function in stem cells regulating lineage and differentiation and have been proposed to function similarly in skeletal muscle [2–5]. Satellite cells from different muscles vary in their capacity to renew and differentiate, which may contribute to a less severe phenotype in myopathies [6]. Whether developmental TFs contribute to susceptibility remains unknown, but their role could be probed through *in vitro* perturbation experiments followed by functional readouts, such as doubling time, or by sequencing assays, such as ChIP-seq and RNA-seq. Indeed, determining whether their contribution is at the level of regeneration or at the intersection with disease specific gene networks would reveal the basis for susceptibility to disease and therefore inform therapy strategies.

5.3 Therapies for FSHD

With the advent of gene therapies, many groups are targeting *DUX4* expression as potential treatments for FSHD [7–10]. However, the functional effects of limiting *DUX4* expression in FSHD-affected muscle have not been well characterized. DUX4 initiates gene dysregulation, but DUX4 expression is sparse and burst-like [11,12]. Additionally, the gene dysregulation does not appear to be reliant on continued DUX4 expression but rather is self-sustained in which case blocking the expression or action of DUX4 would not stop pathogenesis [13–15]. Suppressing DUX4 may limit progression into new muscles, but DUX4 associated gene expression has been found in even mildly affected FSHD muscle [16]. If seemingly unaffected muscles have already initiated the pathogenic program with expression of DUX4, then therapies limiting its expression may not be effective. Rather, attention may need to be turned to factors downstream of DUX4 that are responsible for dystrophy.

5.4 Role of downstream TFs in FSHD gene dysregulation

The expression of DUX4 initiates zygotic genome activation (ZGA) but its expression wanes quickly similar to its expression in FSHD [11,17]. The role of the transcriptional regulators activated by DUX4 in embryogenesis is only partially understood. The role of TFs such as LEUTX during embryogenesis are of particular interest as it is primate specific [18,19]. The role of these TFs in FSHD is even less well understood. We and others have provided evidence for these transcriptional regulators in sustaining the DUX4 activated gene dysregulation [13,15]. Our evidence included depletion assays in FSHD myoblasts, but assays such as ChIP-seq and further validation studies with overexpression in muscle and other cell types could help us understand the magnitude of these regulators on the DUX4 gene network and other genes. ChIP-seq studies for some of these TFs are difficult due to sequence homology so may require tagging for efficient IP [20]. Since these target genes are only activated in a subset of cells, single cell or single nucleus assays such as scATAC-seq or CUT&TAG may be necessary to gain the resolution for expression of these TFs in a native context [21].

5.5 Potential insights from *in vivo* studies

To that effect, the *in vitro* cultures that we have used here have provided incredible insights into the initial dysregulation following DUX4 expression and candidates for mechanisms that may cause dystrophy *in vivo*. However, much remains to be understood regarding the mechanisms by which DUX4 activated gene dysregulation lead to the hallmarks of FSHD. Many mechanisms have been proposed (for a review see [22]) but most of these findings rely on *in vitro* systems or mouse models. As DUX4 and some of its target genes are primate-specific, mouse models of FSHD are of questionable translatability [18–20]. However, *in vitro* culture does not recapitulate a fully differentiated myofiber or the complex microenvironment with other cell types and extracellular signals. Muscle is an endocrine tissue, secreting signals important in normal function, and as discussed, location-specific extracellular signals appear to influence TF expression [23]. The role, if any, of extracellular signals in FSHD is not understood, but could benefit from *in vitro* proteomic and/or metabolic assays. In terms of non-muscle cells, immune cells have been implicated in FSHD but under-explored [16,19]. Biopsy samples from muscles with early signs of disease and low levels of DUX4 target genes have potential immune infiltrates, and immune cells might activate DUX4 and its target gene expression [16,24]. Whether immune cells contribute to pathology remains unclear, but could be understood with single cell RNA-seq.

5.6 High resolution transcriptomics use in FSHD

High resolution transcriptomics from muscle biopsies could answer many open questions in the field. Single cell or single nucleus RNA-seq can be used to identify heterogeneous muscle and non-muscle cell types, which could identify the type of immune cells present in diseased tissue and their contributions to DUX4 related gene expression. Mononuclear muscle cells identified from scRNA-seq or snRNA-seq can be identified in the stage of myogenesis to ascertain the extent of repair and regeneration occurring in dystrophic tissue [25,26]. Importantly, we can use snRNA-seq to correlate nuclei that have activated the DUX4 gene program with other specific profiles. For example, correlating pathogenic nuclei with *MYH* expression can identify the source myofiber type, and correlating with markers of myofiber subpopulations can reveal the potential interactions between pathogenic nuclei and non-muscle cell types such as neurons and tendons. Spatial transcriptomics could give incredible insights by correlating pathology with expression thereby revealing the path from DUX4 initiation to the dystrophic phenotype.

5.7 Application of techniques to other systems and future assays

The assays I have applied and outlined here for use in FSHD can be expanded to further applications in skeletal muscle and other tissues. For example, single-nucleus RNA-seq can be used to assess NMJ nuclei in SMA following muscle denervation. I have discussed the use of genomic and transcriptomic assays, but for diseases such as FSHD which occur in rare populations of nuclei these assays are best performed at the single nucleus level. Assays, such as scNMT-seq, enable profiling of multiple features from the same cell but have not yet been applied to multinucleated skeletal muscle [27]. These assays have an important advantage for rare nuclei populations and would be exciting to use in the context of skeletal muscle. Ultimately the most informative assays for muscle will be those that can retain spatial information not just for relation to pathology but for relating nuclei to the cell of origin. Currently, these combinatorial assays have no way of preserving the spatial context. The application of high-resolution, spatially-resolved genomics in the study of myopathies and rare disease will surely answer many outstanding questions.

5.8 References

1. Evano B, Gill D, Hernando-Herraez I, Comai G, Stubbs TM, Commere PH, et al. Transcriptome and epigenome diversity and plasticity of muscle stem cells following transplantation. PLoS Genet. 2020 Oct 30;16(10):e1009022.
2. Song JY, Pineault KM, Dones JM, Raines RT, Wellik DM. Hox genes maintain critical roles in the adult skeleton. Proc Natl Acad Sci U S A. 2020;117(13):7296–304.
3. Houghton L, Rosenthal N. Regulation of a muscle-specific transgene by persistent expression of Hox genes in postnatal murine limb muscle. Dev Dyn. 1999;216(4–5):385–97.
4. Morgan R. Hox genes: a continuation of embryonic patterning? Trends Genet. 2006;22(2):67–9.
5. Hutlet B, Theys N, Coste C, Ahn MT, Doshishti-Agolli K, Lizen B, et al. Systematic expression analysis of Hox genes at adulthood reveals novel patterns in the central nervous system. Brain Struct Funct. 2016;221(3):1223–43.
6. Stuelsatz P, Shearer A, Li Y, Muir LA, Ieronimakis N, Shen QW, et al. Extraocular muscle satellite cells are high performance myo-engines retaining efficient regenerative capacity in dystrophin deficiency. Dev Biol. 2015;397(1):31–44.
7. Lim KRQ, Maruyama R, Echigoya Y, Nguyen Q, Zhang A, Khawaja H, et al. Inhibition of DUX4 expression with antisense LNA gapmers as a therapy for facioscapulohumeral muscular dystrophy. Proc Natl Acad Sci U S A. 2020;117(28):16509–16515.
8. Himeda CL, Jones TI, Jones PL. Targeted epigenetic repression by CRISPR/dSaCas9 suppresses pathogenic DUX4-fl expression in FSHD. Mol Ther - Methods Clin Dev. 2021;20(March):298–311.
9. Le Gall L, Sidlauskaite E, Mariot V, Dumonceaux J. Therapeutic Strategies Targeting DUX4 in FSHD. J Clin Med. 2020;9(9):2886.
10. Hamel J, Tawil R. Facioscapulohumeral Muscular Dystrophy: Update on Pathogenesis and Future Treatments. Neurotherapeutics. 2018 Oct 25;15(4):863–71.
11. Rickard AM, Petek LM, Miller DG. Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. Hum Mol Genet. 2015 Jun 5;24(20):5901–14.
12. Tassin A, Laoudj-Chenivesse D, Vanderplanck C, Barro M, Charron S, Ansseau E, et al. DUX4 expression in FSHD muscle cells: How could such a rare protein cause a myopathy? J Cell Mol Med. 2013 Jan;17(1):76–89.
13. Williams K, Jiang S, Kong X, Zeng W, Nguyen NV, Ma X, et al. Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. PLoS Genet. 2020;16(5):1–26.
14. Chau J, Kong X, Viet Nguyen N, Williams K, Ball M, Tawil R, et al. Relationship of DUX4 and target gene expression in FSHD myocytes. Hum Mutat. 2021;
15. Resnick R, Wong C-J, Hamm DC, Bennett SR, Skene PJ, Hake SB, et al. DUX4-Induced Histone Variants H3.X and H3.Y Mark DUX4 Target Genes for Expression. Cell Rep. 2019 Nov 12;29(7):1812-1820.e5.
16. Wong CJ, Wang LH, Friedman SD, Shaw D, Campbell AE, Budech CB, et al. Longitudinal measures of RNA expression and disease activity in FSHD muscle biopsies. Hum Mol Genet. 2020;29(6):1030–44.
17. Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L, et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nat Commun. 2019 Dec 21;10(1):364.

18. Katayama S, Ranga V, Jouhilahti E-M, Airenne TT, Johnson MS, Mukherjee K, et al. Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. Sci Rep. 2018 Dec 27;8(1):17421.
19. Yao Z, Snider L, Balog J, Lemmers RJLF, Van Der Maarel SM, Tawil R, et al. DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. Hum Mol Genet. 2014 Oct 15;23(20):5342–52.
20. Leidenroth A, Hewitt JE. A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol Biol. 2010 Nov 26;10(1):364.
21. Ludwig CH, Bintu L. Mapping chromatin modifications at the single cell level. Development. 2019;146(12):dev170217.
22. Campbell AE, Belleville AE, Resnick R, Shadle SC, Tapscott SJ. Facioscapulohumeral dystrophy: activating an early embryonic transcriptional program in human skeletal muscle. Hum Mol Genet. 2018 Aug 1;27(R2):R153–62.
23. Lightfoot AP, Cooper RG. The role of myokines in muscle health and disease. Curr Opin Rheumatol. 2016;28(6):661–6.
24. Banerji CRS, Panamarova M, Zammit PS. DUX4 expressing immortalized FSHD lymphoblastoid cells express genes elevated in FSHD muscle biopsies, correlating with the early stages of inflammation. Hum Mol Genet. 2020;29(14):2285–99.
25. Barruet E, Garcia SM, Striedinger K, Wu J, Lee S, Byrnes L, et al. Functionally heterogeneous human satellite cells identified by single cell RNA sequencing. Elife. 2020 Apr 1;9.
26. De Micheli AJ, Laurilliard EJ, Heinke CL, Ravichandran H, Fraczek P, Soueid-Baumgarten S, et al. Single-Cell Analysis of the Muscle Stem Cell Hierarchy Identifies Heterotypic Communication Signals Involved in Skeletal Muscle Regeneration. Cell Rep. 2020;30(10):3583-3595.e5.
27. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. Nat Commun. 2018 Dec 1;9(1).