Dynamics of Super-Enhancers Throughout Myogenesis

Basma Abdelkarim

A thesis submitted in partial fulfilment of the requirements for the degree of Masters in Cellular/Molecular Medicine with specialization in Bioinformatics

Department of Cellular and Molecular Medicine
Faculty of Medicine
University of Ottawa

Acknowledgements

I would like to take this opportunity to express my gratitude to my supervisor, Dr. Theodore J. Perkins for his assistance and valuable feedback throughout the course of my Masters, my fellow lab members and TAC members for their help and advice during my research. I would also like to thank my parents for their continuous support. Also, my best friend, Nada Radwan for always encouraging me to keep going and cheering me on when times got hard. Last, but certainly not least, I would like to thank my loving husband, Mohamed Kotb for always being there when I needed him and for his continuous help and support throughout this journey. And to whoever has taken the time and interest to read my thesis, I would like to express my sincere appreciation and gratitude.

Abstract

Recently, the term super-enhancer (SE) has been gaining more attention since its characterization in 2013 as a subset of enhancers that form large regulatory domains that regulate cell identity processes and coordinate cell development. Since then, SEs have been characterized in over 100 cell types, including diseased and tumor cells. In an attempt to standardize the method for identifying SEs, the ROSE algorithm was developed. This algorithm uses peak files for the transcription or epigenetic factor of interest and the sequence alignment of the ranking factor to identify SEs according to the signal density of the ranking factor. More recently, there has been interest in studying the dynamics of these SEs throughout development. In this study I introduce a novel algorithm, DYSE, that builds onto the functionality of the ROSE algorithm where it determines key changes in SEs as the cell transitions from state to state, using comparative analysis. Here I explain the features of this algorithm and I present my results from testing it using multiple transcription factors for a three-stage analysis of myogenesis. I characterized SEs that are lost, maintained or gained as the cell transitions from myoblast to early myotubes to late myotube stages. Through gene ontology (GO) enrichment analysis, I found that genes associated with SEs that are maintained between stages show more enrichment for myogenic processes than stage-specific ones.

Table of Contents

Acknowledgements II

Abstract III

List of Figures VI

List of Abbreviations VIII

CHAPTER 1 - Introduction 1

1.1 Transcriptional Regulation 1

1.2 Genomic sequence analysis 3

1.3 Skeletal muscle development 5

1.3.1 Myogenic regulatory factors (MRFs) 5

1.3.2 Other myogenic factors involved in myogenic development 9

1.4 Super-Enhancers overview 10

1.5 Super-Enhancers in differentiated cells 12

1.6 Super-Enhancers in cancers and developmental disorders 13

1.7 Rank Ordering of Super Enhancers (ROSE) algorithm 16

1.7.1 Overview 16

1.7.2 How it works 17

1.7.3 ROSE gene mapper 19

Research rationale 20

Hypothesis 20

Objectives 20

CHAPTER 2 - Materials and Methods 21

2.1 Dynamics of Super Enhancers (DYSE) algorithm 21

2.1.1 Overview 21

2.1.2 Using DYSE 25

2.1.2a Input files 25

2.1.2b Analysis type 26

2.1.3 How it works 27

2.1.3a Obtaining SE regions from ROSE 27

2.1.3b Determining changes in SEs between stages 28

2.1.3c Differential expression of SE associated genes between stages 28
2.1.3d Changes in SE activity between stages..... 29
2.2 Myogenic datasets..... 30
2.2 Preparing input data for DYSE 32

CHAPTER 3 – Results 35
3.1 Characterizing myogenic SEs 35
3.2 Myogenic SE associated genes..... 41
3.3 Changes in SE activity between stages..... 50

CHAPTER 4 – Discussion 53
Conclusion and Future Directions..... 57

Reference 60

Appendix 66
DYSE (Dynamics of Super-Enhancers) - Pseudo code ……. 66

List of Figures

Figure 1.1 **Illustration of a ChIP-seq experiment** 3

Figure 1. 2 **Myogenesis-associated factor** 6

Figure 1. 3 **ROSE Workflow**. 18

Figure 2.1 **DYSE Workflow** 22

Figure 2.2 **Flowchart of the DYSE algorithm** 25

Figure 2.3 **Illustration of determining SE activity changes** 30

Figure 3.1 **ROSE SE plots for all TFs** 36

Figure 3.2 **Comparing SEs between different stage** 40

Figure 3.3 **Gene ontology term enrichment** 49

Figure 3.4 **Changes in SE activity** 51

Figure 3.5 **Visualizing change in SE activity** 52

List of Tables

Table 1.1 **ROSE_main.py required parameters** 16

Table 1. 2 **ROSE_main.py optional parameters** 17

Table 2.1 **DYSE_main.py required parameters** 24

Table 2.2 **DYSE_main.py optional parameters** 24

Table 2.3 **Accession number list** 31

Table 2.4 **Growth and differentiation protocols** 33

Table 3. 1 **SE stats** 35

Table 3.2 **Single-enhancer SEs** 38

Table 3.3 **SE associated genes** 42

Table 3.4 **Differentially expressed genes** 50

Table 3.5 **Differentially expressed genes** 50

List of Abbreviations

Abbreviation	Phrase
bHLH	Basic helix-loop-helix
BRD4	Bromodomain-containing protein 4
ChIP-seq	Chromatin Immunoprecipitation followed by high throughput sequencing
DYSE	Dynamics of super-enhancers
ESC	Embryonic stem cell
GEO	Gene expression omnibus
GO	Gene ontology
GWAS	Genome wide association studies
H3K27ac	Histone H3 Lysine 27 acetylation
H3K27me3	Histone H3 Lysine 27 tri-methylation
H3K4me1	Histone H3 Lysine 4 mono-methylation
H3K4me3	Histone H3 Lysine 4 tri-methylation
HDAC	Histone deacetylase
HFSC	Hair follicle stem cell
IGV	Integrative genomic viewer
Med1	Mediator co-activator
Mef2D	Myocyte enhancer factor 2D
MIT	Massachusetts institute of technology
MRF	Myogenic regulatory factors
PcG	Polycomb
Pol II	RNA polymerase !!
ROSE	Rank ordering of super-enhancers
rpm	Read per million mapped reads
SE	Super-enhancer
SLE	Systemic lupus erythemtosus
SNP	Single nucleotide polymorphism

T-ALL	T-lineage acute lymphoblastic leukemia
TAC	Transit-amplifying cells
TE	Typical enhancers
TF	Transcription factor
Th	T-helper
TSS	Transcription start site
UCSC	University of California Santa Cruz

CHAPTER 1 - Introduction

1.1 Transcriptional Regulation

Different cell types in a eukaryotic organism originate from a single cell. They share the same genome yet acquire different structures and functions throughout the body. These variations between cell types arise in response to various factors including external signaling factors, mechanical forces, as well as differential gene expression. Differential gene expression is where each cell type expresses a different set of protein-coding genes (Shlyueva, et al., 2014). Different classes of transcriptional regulatory elements, such as enhancers, promoters, silencers, and insulators, work together to orchestrate the spatial and temporal patterns of gene expression to regulate cell function (Maston, et al., 2006).

The sequence directly neighboring the transcription start site (TSS) of a gene, known as the core promoter, is sufficient to independently assemble the RNA Polymerase II (Pol II) complex to initiate transcription. Gene expression is significantly weakened however, without the aid of other regulatory elements, such as enhancers.

Enhancers are distal regulatory sequences found in non-protein-coding regions of the DNA that are capable of upregulating gene expression. They could be upstream or downstream of the gene and they may also be located over a megabase (Mb) away from their target gene (Levine, 2010). The first enhancer discovered was a 72 bp segment of the SV40 viral DNA in HeLa cells. This sequence was found to upregulate both upstream and downstream reporter genes over one hundred-fold (Banerji, et al., 1981).

The interaction between distal enhancers and the promoter of the target gene is mediated by a subset of DNA-binding looping proteins (Shlyueva, et al., 2014). These looping proteins bring transcription factor (TF) bound enhancers into close proximity to the target gene promoter. They then enhance the formation

of the Pol II complex machinery to bring about transcription and upregulate the expression of the target gene. The chromatin structure, however, imposes significant obstacles on Pol II transcription.

In eukaryotic cells, several meters of DNA are packaged into compact units. With help from nucleosomes, they can fit in a few microns inside the nucleus of a cell. Nucleosomes are histone octamer complexes that are made up of two copies of each of histones H2A, H2B, H3 and H4. About 146 bp of DNA wrap around each nucleosome (Conaway, 2012). Across the DNA structure, arrays of nucleosomes create barriers that reduce the accessibility of the DNA for TF binding. Since enhancers need to be bound by their TFs in order to activate transcription, this compaction significantly hinders the process. Nucleosomes, however, are not static units, and can undergo dynamic changes. These changes are regulated by various proteins through mechanisms such as histone modifications and chromatin remodeling (Li, et al., 2007).

Histone modifications that were found to be correlated with regulatory elements mainly include methylation or acetylation of lysine (K) residues (Li, et al., 2007). Enhancers are normally devoid of nucleosomes to allow for TF binding. However, the nucleosomes in their vicinity are characterized by mono-methylation of Lysine 4 on histone 3 (H3K4me1). Promoters, on the other hand, are associated with di and tri-methylation of Lysine 4 on histone 3 (H3K4me2 and H3K4me3) (Heintzman, et al., 2007; Li, et al., 2007).

Another modification that has been found to play a key role in transcriptional regulation is Histone 3 Lysine 27 acetylation (H3K27ac). H3K27ac was found to differentiate between active and poised enhancers. Enhancers that are actively transcribing genes are surrounded by nucleosomes that typically have H3K4me1 and H3K27ac. Poised or temporarily inactive enhancers, are marked by H3K27me3 (Creyghton, et al., 2010; Rada-Iglesias, et al., 2011).

1.2 Genomic sequence analysis

Advanced high-throughput sequencing technologies have helped researchers understand the distribution of enhancers across the genome (Shlyueva, et al., 2014). This was facilitated by the ability to study the patterns of typical characteristics of enhancers, such as TF binding and histone modifications. Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) is a widely used method used to identify the genome-wide binding sites of single DNA-binding proteins or the locations of various histone modifications (Furey, 2012) (Figure 1.1).

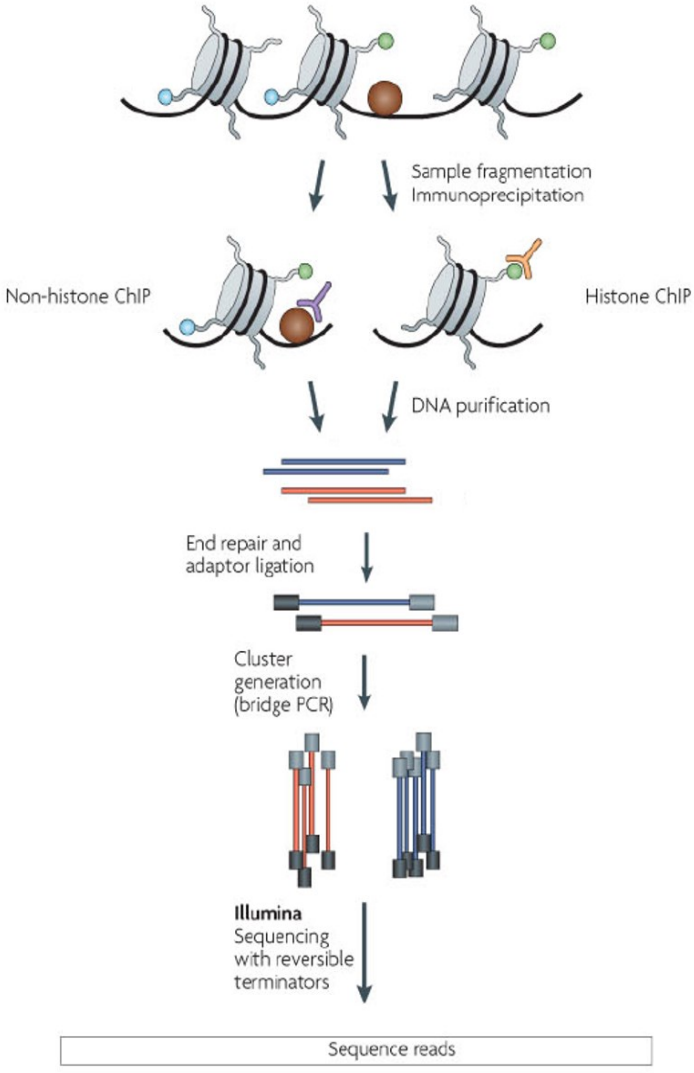


Figure 1.1 **Illustration of a ChIP-seq experiment** Chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing can be used to profile chromatin sites where DNA interacts with DNA-binding proteins and histone modifications. Adapted from *(Park, 2009)*.

Chromatin is first extracted from a cell culture to be used for ChIP-seq. As described earlier, chromatin consists of DNA wrapped around histone protein complexes. Transcription factor proteins and other enhancer related factors bind to open chromatin regions in the DNA that lack these histone proteins. One of the purposes of ChIP-seq is to find the genome-distribution of specific TFs and other enhancer related factors to help understand their role in regulating and maintaining various cell

processes and the genes that they regulate. It is also used to profile the distribution of various histone modifications across the genome. ChIP-seq is briefly described below as presented by Peter J. Park (Park, 2009; Figure 1.1). There are various technologies for high-throughput DNA sequencing. The following describes Illumina technology, which is currently the most widely used.

In the first step of the ChIP-seq process, formaldehyde is used to induce the proteins to bind DNA at their specific binding sites. This process preserves the bound proteins until they are ready for separation. However, this means that even proteins that are not targeted in the study would be bound as well. Once this step is complete the DNA is sheared into short fragments (~200-600 bp). The fragments bound by the TF of interest are then isolated using an antibody complementary to it.

The TF bound DNA fragments that are attached to the antibody are then extracted from the solution. The formaldehyde crosslinking is then reversed. After that the DNA fragments are separated from TFs and the TFs are washed away. From this process, millions of DNA fragments from many cells are produced.

These fragments are then used to create a sequencing library by adding sequencing adaptors to both ends of the DNA fragments (Ernst, et al., 2011). The fragments are then denatured, isolating the template strand preparing it for PCR bridge amplification. During PCR amplification, each DNA fragment is copied multiple times on a solid surface with covalently attached DNA linkers. The DNA linkers hybridize the adaptors attached to the DNA. This generates clusters of DNA fragments from each original fragment. Those clusters are then used for the sequencing step. The average sequencing depth for large genomes such as the human genome would be around 20 – 40 million mapped sequence reads (Furey, 2012).

The process used by Illumina's next generation sequencing (NGS) platforms, Sequencing by Synthesis, is as follows. High throughput sequencing by synthesis involves a step-by-step incorporation of reversibly fluorescent and terminated nucleotides. In each cycle, all four nucleotides are added to the sequencing chip at the same time and after nucleotide incorporation, the remaining nucleotides are washed away.

The fluorescent signals are read at each cluster and recorded. The fluorescent molecule and the terminator are then cleaved and washed away. This is repeated for multiple cycles until the whole DNA cluster is sequenced. In this process, tens of millions of DNA fragments are sequenced simultaneously, generating short sequence reads that can be used for further genomic analysis (Anon., 2019).

ChIP-seq can be done against TFs or histone marks. ChIP-seq has been widely used to study characteristics of regulatory regions in various cell types in order to understand gene expression patterns throughout cell development (Ernst, et al., 2011; Zhu, et al., 2013; Shlyueva, et al., 2014). Using ChIP-seq against histone modifications revealed that the patterns of promoter activity are more common across different cell types while the patterns of enhancer activity are more cell-type specific. This means that the genome wide distribution of the histone modification H3K4me3 is common across different cell types. Whereas, the distribution of H3K4me1 and H3K27ac are more commonly cell-type specific. For instance, the distribution pattern for muscle is distinct from red blood cell patterns.

1.3 Skeletal muscle development

1.3.1 Myogenic regulatory factors (MRFs)

In this thesis, I propose a new bioinformatic method for analyzing enhancers, and more specifically super enhancers. I apply the method I developed to a study of super enhancers in skeletal muscle development. Therefore, before going on to define super enhancers, we first review key facts about the transcriptional regulation underlying skeletal muscle development.

Skeletal muscle development, as for all other tissue types, involves dynamic changes in the chromatin state. As the cell transitions from one stage to another, the patterns of active enhancers change, causing different protein-coding genes to be expressed at each stage (Nord, et al., 2013).

Myogenic development involves a complex gene transcriptional network. This is still a focus of interest in many studies, and it serves as a model to study the mechanisms of molecular and cellular differentiation (Figure 1.2).

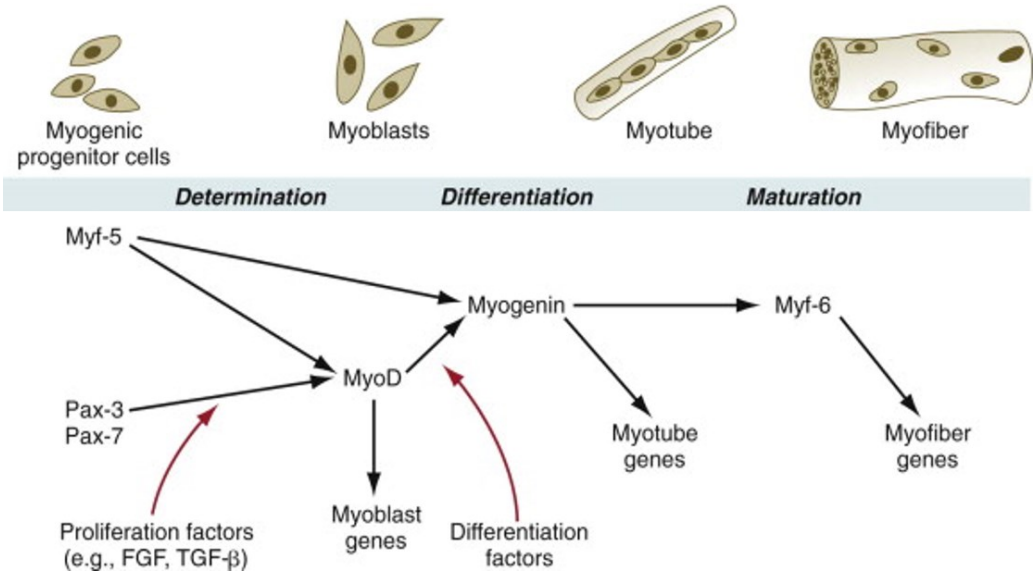


Figure 1. 2 Myogenesis-associated factors Different muscle-specific transcription factors are activated at different times throughout myogenic development to regulate proper development and function of myoblasts and myotubes. Adapted from *(Carlson, 2014)*

Skeletal muscles are striated muscles that are mainly used for voluntary skeletal movement. They are made up of segmented muscle fibers. These segments, also known as sarcomeres, are made up of actin and myosin filaments. Skeletal muscles are one of three major types of muscles, the others being cardiac and smooth muscles.

During skeletal muscle development and regeneration, progenitor cells, which are in this case satellite cells, undergo multiple changes, driven by activation of different sets of skeletal muscle-specific transcription factors. The myogenic regulatory factors (MRFs) are a family of basic helix-loop-helix (bHLH) muscle-specific proteins. MRFs include MyoD (Davis, et al., 1987), Myf5, Myogenin and Myogenic Regulatory Factor 4 (MRF4), also known as Myf6 (Braun, et al., 1989). These regulatory factors are capable

of transforming certain cell types such as embryonic fibroblasts into myogenic cells that can fuse into myotubes. Entry into the myogenic program is dependent on the expression of MyoD1 and Myogenin which confer skeletal muscle identity.

The paired-box homeodomain genes Pax3/7 and the MRF Myf5 control the entry of multipotent cells into the myogenic lineage (Hernández-Hernández, et al., 2017). Pax3 and Pax7 are paralogs with very similar genetic sequences. They share the same DNA binding sequence, yet, have different affinity for binding those regulatory regions. They have some overlapping but non-redundant functions in maintaining myogenic precursor cells (satellite cells) (Soleimani, et al., 2012).

Satellite cells lie under the basal lamina of muscle fibers and are maintained in a quiescent state (Relaix, 2006). Quiescent satellite cells can be activated to either proliferate and produce more satellite cells in response to muscle injury, or to exit the cell cycle and differentiate into myocytes, which then fuse into myotubes. These myotubes are then used to create more muscle fibers (Charge & Rudnicki, 2004). Both quiescent and actively proliferating satellite cells were found to highly express Pax7, but not MRFs. Pax3 is similarly co-expressed in satellite cells but only in a subset of muscles, such as the diaphragm, and expressed in low levels otherwise (Kassar-Duchossoy, et al., 2005).

Pax7 is required for satellite cell specification (Seale, et al., 2000). Further studies suggested that Pax7 also plays a key role in maintaining satellite cell survival and proliferation and inhibiting differentiation (Relaix, et al., 2006). Pax3, on the other hand, is involved in delamination and migration of embryonic myoblasts from the somite to limb bud during development. It is also capable of directly activating Myf5 and MyoD to initiate differentiation. Yet, it cannot compensate for loss of Pax7 (Soleimani, et al., 2012).

Pax3 and Pax7, which are upstream of MRFs in the myogenic transcriptional network, control the entry of satellite cells into the myogenic program by upregulating Myf5 and MyoD. These four factors together form a regulatory network that regulates satellite cell commitment to the myogenic lineage by the control

of various signaling pathways (Soleimani, 2012). With entry to myogenesis Pax3/Pax7 expression is down regulated and Myf5/MyoD expression is upregulated. These cells are then specified as myoblasts (Yokoyama & Asahara, 2011).

Myf5 and MyoD have somewhat redundant functions throughout myogenesis, and the inactivation of either one of them has no visible effect on skeletal muscles in the embryonic stage. The absence of Myf5 slows down myogenesis only up to the point at which MyoD is expressed (Braun, et al., 1992). On the other hand, cells lacking MyoD prolong the expression of Myf5 to make up for the lost MyoD (Rudnicki, et al., 1992). However, a Myf5/MyoD double knockout leads to the complete loss of myoblasts and skeletal muscle throughout the body (Rudnicki, et al., 1993). These two MRFs play a key role in activating downstream muscle-specific genes necessary for proper function and phenotype of muscle cells (Berkes & Tapscott, 2005).

More downstream in the transcriptional network is Myogenin. On the onset of differentiation, MyoD in proliferating myoblasts activates the expression of the Myogenin differentiation factor. Contrary to Myf5 and MyoD, Myogenin is essential for terminal differentiation of skeletal muscles. Studies have shown that even with the presence of proliferating myoblasts, mice lacking Myogenin have severely underdeveloped skeletal muscles (Berkes & Tapscott, 2005).

The last of the MRFs is expressed downstream of MyoD and Myf5 in the terminal differentiation process in later stages. It is involved in triggering muscle specific factors and inducing terminal differentiation of myoblasts marked for differentiations; those that are expressing MyoD and Myf5 (Bentzinger, et al., 2012).

1.3.2 Other myogenic factors involved in myogenic development

Other transcriptional regulators besides MRFs are indirectly involved in the myogenic transcriptional network to regulate development. Snai1/2 are highly expressed in myoblasts and bind the same DNA motifs as bHLH TFs, such as MyoD (Soleimani, et al., 2012). Their main role is to regulate the myogenic differentiation process.

If muscle fibers are made without enough myoblasts present, the muscle fibers produced would be poorly structured and would break apart. In order to properly regulate the development of muscle fibers, the differentiation process needs to be blocked when there is a shortage of myoblasts. This can be done by not giving MyoD access to the binding site in order to prevent it from expressing the genes needed for the differentiation of myoblasts. This allows for the expansion of the active myoblast pool prior to differentiation to allow for proper development of muscle fibers.

Snai1/2 prevents/delays differentiation by blocking MyoD from binding to its target sites. Snai1/2 first recruits histone deacetylases, HDAC1/2 and forms a repressive complex. HDAC1/2 remove the H3K27ac modification from nucleosomes surrounding the MyoD binding site, thereby making it inaccessible. The repressive complex specifically binds G/C rich enhancer box (E box) regions which are specifically associated with differentiation genes. Hence, delaying differentiation (Soleimani, et al., 2012).

MyoD activates Myogenin in later stages of myogenesis through two main mechanisms. It binds and induces chromatin remodeling at the Myogenin promoter to form the transcriptional complex and initiates transcription of Myogenin. In order for MyoD to stably bind the Myogenin promoter it requires an interaction with Pbx1, which is a TF containing a TALE (three-amino-acid loop extension) homeodomain (Berkes, et al., 2004). This interaction between Pbx1 and MyoD regulates the expression of multiple other MyoD-dependent genes. Hence, Pbx1 also plays an indirect, yet, important role in myogenic regulation (Fong, et al., 2015).

Along with Myogenin and other muscle-specific genes, MyoD induces the expression of Myocyte enhancer factor 2 (Mef2). Mef2 regulates the later stages in myogenesis by interacting with MRFs. On its own, it is incapable of determining the myogenic lineage, but it indirectly plays a key role in regulating the proper execution of the transcriptional network in control of myogenesis. Activation occurs when MyoD binds Mef2 forming an activation complex inducing transcriptional activation. In the case where this conformation does not occur, for instance, in the case where MyoD and Mef2 interaction is disrupted by the Notch signaling pathway, myogenesis is repressed (Kopantseva & Belyavsky, 2016).

Tead4 is another myogenic-related TF that acts in later stages to promote differentiation. It is involved in inducing cell cycle arrest by repressing the connective tissue growth factor (CTGF) (Benhaddou, et al., 2012). In the myogenic transcriptional network, Tead4 is downstream of MyoD and Myogenin. MyoD and Myogenin both directly bind the Tead4 promoter to initiate its expression during myotube differentiation (Joshi, et al., 2017).

1.4 Super-Enhancers overview

Similar to skeletal muscle cells, all other cell types have a specific transcriptional network that is carefully regulated to maintain proper structure and function. Understanding these transcriptional networks and identifying genes that play an important role in cell identity and development is a critical first step to understanding various developmental disorders.

Genome-wide profiling of putative enhancer regions has revealed that some enhancers have disproportionate binding of transcription factors compared to the rest of enhancers (Hnisz, et al., 2013; Whyte, et al., 2013; Lovén, et al., 2013). It has been proposed that clusters of enhancers enriched with Med1 co-activator form larger regulatory domains called super-enhancers (SEs) that regulate the

expression of cell type specific genes that are important for maintaining cell identity (Love´n, et al., 2013; Vahedi, et al., 2015).

SEs were first characterized in mouse embryonic stem cells (ESCs). ChIP-seq analysis of enhancer-associated Med1 binding showed exceptionally high levels of enrichment at large regulatory domains (Whyte, 2013). In ESCs, only a handful of master transcription factors regulate important gene expression programs to maintain a pluripotent state, some of which are Oct4, Sox2, Nanog, Klf4 and Essrb (Chen, et al., 2008).

When investigating genome-wide co-occupancy of OSN (Oct4, Sox2 and Nanog) in ESC, Whyte et al. identified a subsection of enhancers that had unusual enrichment of these factors (Whyte, et al., 2013). They also found that these regions are particularly enriched with TFs Klf4 and Essrb as well as Med1 co-activator, compared to other TFs. These were characterized as ESC SEs.

Out of 8,794 OSN bound enhancers, 231 were characterized as SEs. Further ChIP-seq analysis of ESC SEs showed that they contain factors that are typical of enhancers but on a much higher magnitude. These include H3K4me1, DNaseI hypersensitivity (Whyte, et al., 2013), H3K27ac (Creyghton, et al., 2010) and bromodomain-containing protein 4 (BRD4) (Niederriter, et al., 2015). The most striking difference between ESC typical and super enhancers observed was Klf4 and Essrb TF binding.

The genes associated with ESC SEs included the majority of the genes and TFs that have been linked to the control of ESC identity. These SE associated genes included the OSN master regulators. SE associated genes were more sensitive to perturbation in TF binding than other genes (Whyte, et al., 2013).

Further studies investigated different methods to identify SEs in a wider variety of differentiated cells (Niederriter, et al., 2015). In the original study where SEs were defined by Whyte and colleagues, Med1 coactivator ChIP-seq signal and distribution was the main factor used to distinguish between typical and super enhancers. In other studies, other enhancer associated factors including p300, BRD4, H3K4me1,

H3K27ac and DNase hypersensitivity (Heintzman, et al., 2007; Visel, et al., 2009) were tested for their ability to identify SEs. This is because master regulators that are likely to form SEs are not known for many cell types (Hnisz, et al., 2013).

Out of all marks, H3K27ac identified the majority of OSN-mediator SEs with the least extra sites. Henceforth, H3K27ac was used as a surrogate mark for identifying SEs. Using H3K27ac, Hnisz and colleagues created a catalog of SEs from over 80 human cell and tissue types. Complementary to previous studies, the majority of SEs identified were cell-type specific and they regulated genes that functioned predominantly in biological processes that largely define the identity of their respective tissue types (Hnisz, et al., 2013).

1.5 Super-Enhancers in differentiated cells

Whyte and colleagues went on to investigate the concept of SEs in differentiated cells. In murine progenitor B (pro-B) cells they defined 13,814 enhancers using ChIP-seq data for the master TF PU.1. From those, they identified 395 large regulatory domains sharing characteristics with SEs they characterized in ESCs. These regions also displayed high enrichment of Med1 coactivator. The constituent enhancers within SEs had a higher frequency of TF-binding motifs that correspond to TFs that are important for pro-B cells (Whyte, et al., 2013).

Whyte and colleagues also observed similar trends of clustered binding of lineage-specific master regulators in differentiated cells. In myotubes clusters of MyoD binding regions were observed, in T helper (Th) cells, clusters of T-bet binding sites were observed and in macrophages clusters of C/EBPa binding sites were observed. As seen in ESCs, master regulators in these differentiated cells accumulate at enhancers in close proximity forming large regulatory domains.

The SEs identified in these cell types were also found to be associated with genes that play a major role in the biology of these cells. In myotubes, MyoD, which is a master regulator, forms a circuit where it binds SEs to regulate the gene that codes for itself (Whyte, et al., 2013). In Th cells one of the SEs was associated with Tcf7 gene which encodes for a protein essential for the production of T cells in hematopoiesis (Tcf-1). In macrophages, the gene that encodes for the extracellular matrix glycoprotein Thbs-1 is regulated by a SE. The main observations from these analyses were that SE elements are majorly cell type-specific and they regulate lineage-specific genes that control cell identity and function.

In another study, the steps first defined to identify SEs were used to identify SEs in a multitude of differentiated cell types. Cell-type-specific enhancers were used to identify SEs in 86 human cell and tissue types including adipose tissue, heart, lung and multiple others. H3K27ac histone mark was then used, as previously described, as the ranking factor to distinguish between SEs and TEs. Next, SE associated genes were identified using gene ontology analysis. The gene ontology analysis showed that SE associated genes were mostly involved in cell-type-specific biological processes. Hence, SE associated genes were used to identify candidate master regulators for various tissue types. For example, TBX20, TBX, MEF2A, NKX2-5 and GATA4 were identified as candidate master regulators for heart tissue (Hnisz, et al., 2013).

This criterion proposed by Whyte et. al and Loven et. al to identify SEs has proved significant in the sense that it has been shown to identify majorly cell type-specific regulatory domains associated with key genes that regulate cell fate. The dbSuper online database has compiled over 82,000 SEs in 102 human and 25 mouse cell types, generated using different ranking factors (Khan & Zhang, 2016).

1.6 Super-Enhancers in cancers and developmental disorders

The strong correlation between enhancer activity and the proper execution of gene expression programs has led many researchers to investigate the correlation between mutations in non-coding regulatory

regions and the development of tumors and different developmental diseases (Visel, et al., 2009; Lee & Young, 2013).

A number of diseases and cancers are known to be a result of misregulation of gene expression networks. In many cases, this is caused by a variation in the regulatory elements associated with them (Lee & Young, 2013); that is because TFs rely heavily on the sequence of regulatory elements for binding. As previous analysis has shown, SEs contain many more master TF binding sites compared to TEs.

Studies have shown that in multiple cases mutations associated with developmental disorders appear in transcription factors and cofactors that accumulate at SE regions (Niederriter, et al., 2015). For example, mutations in chromatin remodeler CDH7 are affiliated with the development of CHARGE Syndrome. CHARGE Syndrome which is characterized by neuronal developmental defects, cardiac malformation, and vision and hearing loss (Vissers, 2004). Another example is Cohesin. Mutations in the cofactor Cohesin causes Cornelia de Lange Syndrome (CdLs). CdLS is a developmental disorder mainly characterized by developmental delays, structural birth malformations and upper limb malformations (Izumi, 2015). These cofactors along with others were found to be associated with SEs (Hnisz, et al., 2013).

Using genome-wide association studies (GWAS), analyses have indicated that more than half of the disease-associated single nucleotide polymorphisms (SNPs) that occur at non-coding regions fall in putative SE regions throughout the genome, implying an association between SEs and complex disorders and cancers. Moreover, SNPs that fall within SE domains tend to appear in disease-relevant cell types (Hnisz, et al., 2013). In other words, SNPs that are correlated to a certain condition like Alzheimer's are likely to be found at regulatory regions that control the expression of Brain-specific genes. Moreover, this would be observed in the tissue type that is mostly associated with the condition, which is in this case brain cells, as compared to other tissue types. As indicated by Hnisz et al. (Hnisz, et al., 2013), in Alzheimer's disease about 5 out of the 27 SNPs linked to Alzheimer's were found in SE regions in brain

cells. This trend was also reported in type 1 diabetes where 13 out of the 72 SEs linked to the disease were found in SEs of primary Th cells. In systemic lupus erythematosus (SLE), as many as 33% (22/67) of non-coding SNPs fall within SE domains in B cells (Hnisz, et al., 2013). This pattern has also been reported in various other diseases and syndromes (Achour, et al., 2015; Gelato, et al., 2018).

Moreover, studies also investigated how SEs are linked to the development of different cancers. They use SEs to identify possible or they use known oncogenes to identify SEs that drive them (Hnisz, et al., 2015; Niederriter, et al., 2015). Indeed, SEs were found to accumulate close to known oncogenes in various types of cancers (Hnisz, et al., 2013).

One interesting observation in cancer cells is that, compared to their healthy counterparts, they are either missing or have extra SE domains. Cancer cells tend to acquire SEs close to known oncogenes (Lovén, et al., 2013), likely due to genomic translocation or misregulated TF gene expression (Qian, et al., 2014; Niederriter, et al., 2015). Studies have shown that multiple cancer types are associated with disruption of SE regions including, but not limited to cancers in breast, colon (Hnisz, et al., 2015) and lung (Hnisz, et al., 2013) and various leukemias (Dawson, et al., 2014). This also applies to pediatric tumors like glioblastoma (Love´n, et al., 2013), neuroblastoma (Chipumuro, et al., 2014), T-lineage acute lymphoblastic leukemia (T-ALL) (Mansour, et al., 2014), etc.

The rising interest in using SEs to identify possible master regulators controlling cell identity and development, cofactors and TFs correlated to complex diseases, as well as oncogenes driving development of cancers increases the need for bioinformatics tools to compare SEs between different conditions and cell stages.

1.7 Rank Ordering of Super Enhancers (ROSE) algorithm

1.7.1 Overview

In an effort to automate and standardize the method used to identify SEs, Richard A. Young and his colleagues at the Massachusetts institute of technology (MIT) developed an algorithm widely known as ROSE (rank ordering of super enhancers) (Lovén, et al., 2013; Whyte, et al., 2013). This algorithm was developed using Python, a general-purpose programming language. It is run by calling the main Python script through command line. The command is used as follows:

```
python ROSE_main.py -g GENOME_BUILD -i INPUT_CONSTITUENT_GFF -r RANKING_BAM
          -o OUTPUT_DIRECTORY [optional: -s STITCHING_DISTANCE -t
                  TSS_EXCLUSION_ZONE_SIZE -c CONTROL_BAM]
```

The algorithm uses the same three-step method proposed by Whyte et. al to characterize SEs in, potentially, any given cell type (Lovén, et al., 2013; Whyte, et al., 2013). The description of the input needed for each parameter to run ROSE is in Tables 1.1 and 1.2.

Table 1.1 **ROSE_main.py required parameters**

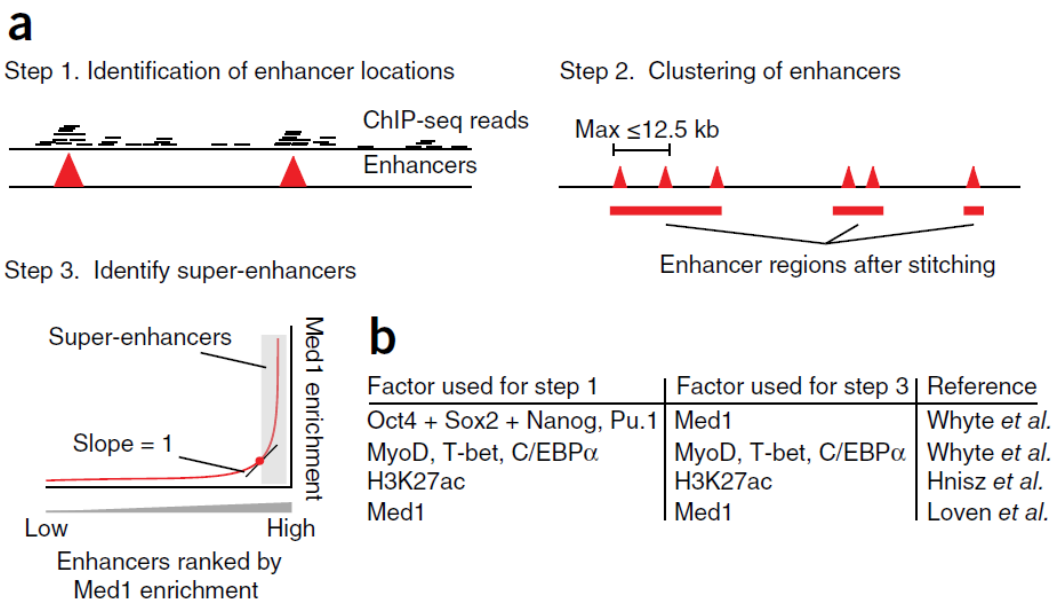
Parameter	Input
GENOME_BUILD	hg18, hg19, hg38, mm8, mm9 or mm10 referring to UCSC genome build (used for gene mapping). These RefSeq gene annotation files are included in the annotations folder in the ROSE algorithm repository
INPUT_CONSTITUENT_GFF	.gff format (more details in Chapter 2) of TF regions that are calculated to be enhancers using a peak calling algorithm (e.g. Model-based Analysis of CHIP-seq (MACS))
RANKING_BAM	ChIP-seq read alignments of the factor used for ranking enhancers by density
OUTPUT_DIRECTORY	Directory where the output folders are to be stored

Table 1. 2 **ROSE_main.py optional parameters**

Parameter	Input
STITCHING_DISTANCE	Maximum distance between two enhancer regions that are to be stitched (default is 12.5 kbp)
TSS_EXCLUSION_ZONE	Exclude regions within this distance up or downstream from a transcription start site (TSS) in order to account for promoter biases (default=0, recommended=2500)
CONTROL_BAM	.bam file to be used as a control and is subtracted from the density of the RANKING_BAM

1.7.2 How it works

The methodology used in the ROSE algorithm covers the characteristics that were attributed to super-enhancers such as clustering and histone marks. This is based on the idea that active enhancer regions in close proximity can collectively form a regulatory domain to control the expression of important cell identity and differentiation genes (Whyte, et al., 2013). In the first step, regions in the input constituent .gff file within the defined stitching distance are grouped together. These regions are more commonly known as stitched enhancers. In the case where the TSS exclusion value is not 0, the exclusion is done before the stitching step.



Factor used for step 1	Factor used for step 3	Reference
Oct4 + Sox2 + Nanog, Pu.1	Med1	Whyte *et al.*
MyoD, T-bet, C/EBPα	MyoD, T-bet, C/EBPα	Whyte *et al.*
H3K27ac	H3K27ac	Hnisz *et al.*
Med1	Med1	Loven *et al.*

Figure 1. 3 **ROSE Workflow** - Identifying SEs for a given TF: (a) Illustration of three step method used to characterize SEs. Step 1: enhancers are defined by peak calling ChIP-seq data for TF of interest. Step 2: stitching enhancers within 12.5kb. Step 3: Rank ordering of all enhancers based on Med1 signal and point of inflection of plot used to distinguish TEs and SEs. (b) Factor combinations used for steps 1-3 to identify SEs in different studies. Adapted from (Pott and Lieb, 2015).

Next, the collection of stitched and constituent enhancers is used to determine the signal of the ranking factor and the control factor, if provided by the user. SAMTools, a software that can generate alignments in a per-position format, is used for this step. It then identifies the ranking factor and control factor reads at each enhancer region.

ROSE then uses the reads to calculate the average density of the ranking factor by dividing the reads per million mapped reads (rpm) of the ranking factor by the length of each enhancer locus corresponding to the TF under study. Using the density of the reads rather the read counts as the factor determining the SE rank removes the bias of the SE size. If the read count was used instead, some SEs that are larger in size would rank higher by default when in fact there might be SEs consisting of single enhancers but having higher signal density and hence should rank higher in the SE list. In the case where the control factor is

used, the control factor signal density is subtracted from the ranking factor density and any negative value is converted to 0.

The algorithm then calls another script that is written in R, a programming language for statistical computing and graphical representations. This script takes in all the enhancers and ranks them according to the overall signal of the ranking factor (and control). Enhancers are then plotted on a scatter plot and the inflection point on the graph (the point where the slope is 1) is then used to distinguish typical and super-enhancers. Enhancer regions that have an overall ChIP-seq signal higher than the point of inflection are characterized as super-enhancers.

1.7.3 ROSE gene mapper

Another useful feature of ROSE is the gene mapper. It assigns super enhancers to genes. It uses the list of SE regions from the main ROSE algorithm and a RefSeq gene annotations file as input. The gene mapper first obtains the list of all genes from the annotation file and identifies three categories of SE-associated genes: overlapping genes, proximal genes within 50 Kb upstream or downstream from the TSS, and finally, distal genes within a 1Mb search window.

ROSE quickly became the most deployed algorithm to identify genome-wide SE regions. It has since been used in numerous studies to understand the function and involvement of SEs in various pathways associated with cancer and other diseases (Niederriter, et al., 2015). It has been shown to be useful in studying these relationships (Bhagwat, et al., 2016; Gelato, et al., 2018).

Research rationale

Super-enhancers have been widely studied since their characterization by Whyte et. al (2013). The studies ranged from identifying super-enhancers in various cell types including cancer cells and cells affected by other diseases (Niederriter, et al., 2015; Khan & Zhang, 2016) to comparative analysis of SEs between cells under different conditions or between healthy and diseased cells (Adam, et al., 2015; Pérez-Rico, et al., 2016). Even though ROSE plays a big role in identifying SEs in a given cell type, one would only be looking at SEs bound by a single TF. Moreover, it is only looking at a single cell condition or stage at a time.

Seeing how cell type specificity of SEs is emphasized, we believe that they are similarly stage-specific to an extent. Hence, rather than looking at SEs at a single stage, I believe that there is a lot of information to be gained by looking at SEs as a dynamic structure throughout cell development (Adam, et al., 2015).

Hypothesis

Using the ROSE algorithm over multiple stages and using a combination of TFs may identify novel SEs that are either lost, gained or maintained throughout cell development.

Objectives

Obj 1 - Writing an algorithm to computationally characterize the changes in super enhancers throughout development

Obj 2 - Characterizing myogenic super-enhancers and their dynamics by applying my algorithm to *Mus musculus* 3-stage myogenic data to identify skeletal muscle-associated SEs and their associated genes

Obj 3 - Analyzing changes in myogenic SE activation and repression throughout development as well as differential expression of SE associated genes between different stages

CHAPTER 2 - Materials and Methods

2.1 Dynamics of Super Enhancers (DYSE) algorithm

2.1.1 Overview

I developed a new SE comparative tool that builds on top of the ROSE algorithm. It determines changes in super-enhancer distributions between different cell stages (or conditions) and characterizes their dynamics as the cell transitions from one stage to another. The algorithm was built using the scripting language Python2.7.13 and utilizes the ChIP-seq analysis tools SAMTools and BEDTools. It uses ChIP-seq and RNA-seq data to identify changes in SE activity and their associated gene expression changes throughout cell development. Hence the name, DYSE (dynamics of super-enhancers).

The DYSE algorithm has two main parts. First, it identifies SEs using the ROSE algorithm for each cell condition or stage used and second, the downstream analysis that identifies overlaps of active SEs between the different stages (Figure 2.1). For the first step DYSE creates a sub-directory in the main output directory where it stores the raw output from ROSE for each TF. The SE bed files for all TFs associated with one state are then grouped together into one table and regions that overlap are then combined into one region while keeping track of the TFs that are bound, hence identifying SE regions that are co-bound by different factors. This is repeated for each state.

Next, to identify SEs that are specific to each stage I do the following: For each state, I compare the SE list for that state with the SEs from all other states. I group together all SEs from all states excluding that one and use bedtools to identify which regions do not overlap with any SE from the other cell states. Using this method, I determine which SEs are exclusively active in each state.

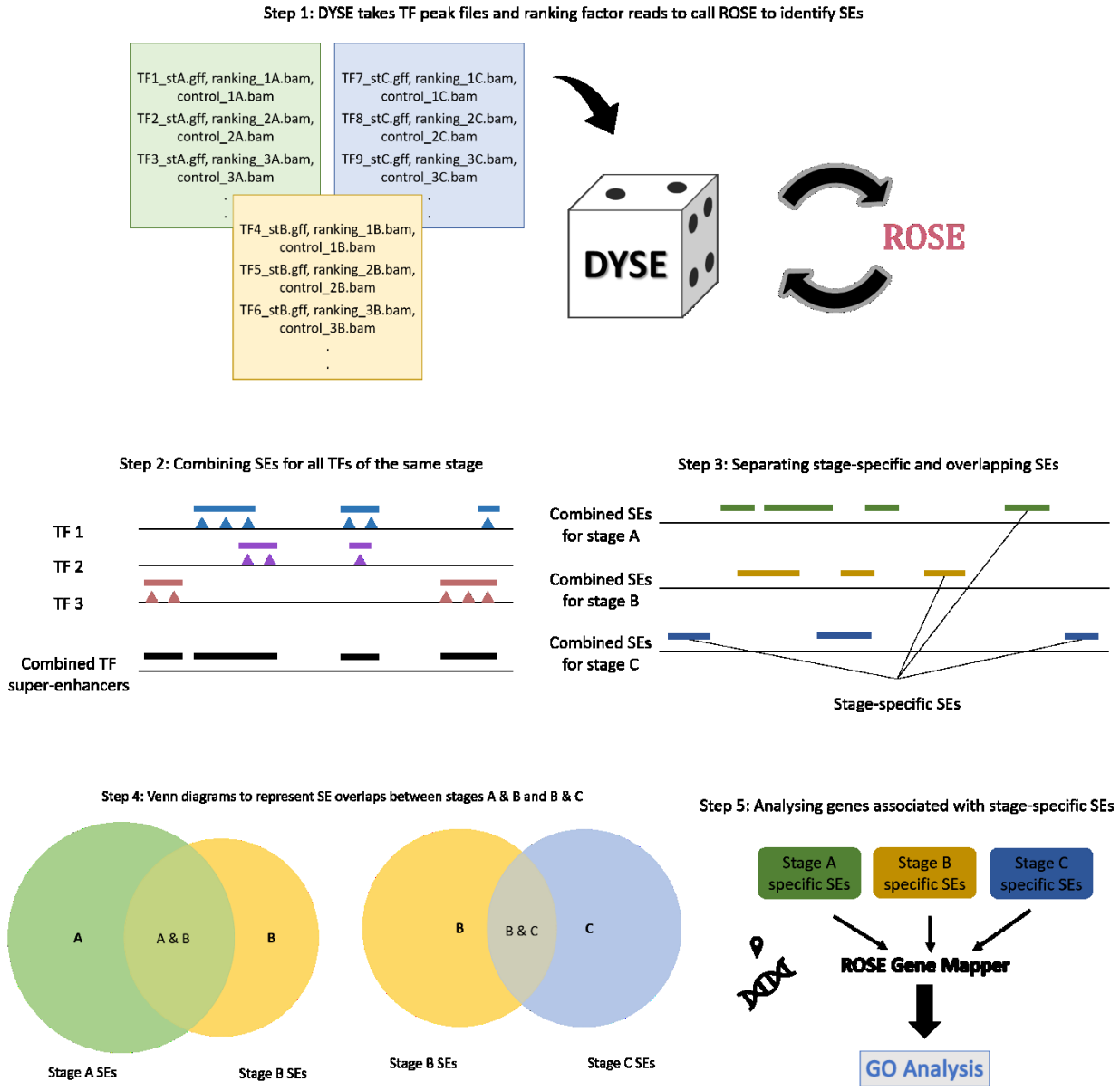


Figure 2.1 **DYSE Workflow** Step 1 – Identifying the SEs for each of the desired transcriptional factors by calling the ROSE algorithm (Pott and Lieb, 2015). Step 2 – For each stage, the SE regions for all transcription factors are grouped together to form one super-enhancer file for each stage. Stage 3 – SEs that appear in more than one stage are identified and separated from stage-specific ones. Stage 4 – The results for adjacent overlaps are visualized using proportional Venn diagrams. Step 5 – Identifying genes associated with SEs from each stage using the ROSE gene mapper script. Resulting genes can then be used for Gene ontology and differential gene expression analysis

Another aspect of my algorithm is to identify SEs that are potentially active in different cell states. There are two options that are provided by DYSE: to compare adjacent states only (according to the order of input files), or to compare all pairs of states. Stage-specific and multi-stage SEs are then sent to different output folders in the main output directory. The combined SEs for each stage are used to identify which SE regions overlap between different cell states, hence identifies SEs that are potentially similarly active during different stages or in different cell conditions. For that I use the bedtools intersect function using the option that shows the genomic regions in both cell states that are being compared, and also the option to see the extent of the overlap, as in, how many bps were involved in the overlap. Further details about the function and use of the algorithm is explained below.

I then use the ROSE gene-mapper script to identify which genes are associated with the multi-stage and stage-specific SEs. This gives us information about proximal, distal and overlapping genes for each of the SE regions used in the input. This information may be used to infer which genes may play a key role in regulation of cell identity and development. The algorithm I developed also provides a differential expression feature where this list of SE associated is compared with RNA-seq differentially expressed genes for that cell state to find out how the expression level of these genes changes/is different between different cell states.

The other feature that DYSE provides is determining changes in SE activity between different states. This function is based on the assumption that the user is using a histone mark or co-factor that is associated with repression of enhancer regions, like H3K27me3 for example. The idea behind this function is to compare the state of an SE from one cell state to another. In the step where we identify SEs we use a factor or histone mark that is associated with activating enhancers, such as H3K27ac. If, in the case of a multi-stage cell lineage, at a later state H3K27me3 is accumulated at that enhancer then that means that the enhancer gets repressed over time. The opposite case where an enhancer has H3K27me3 accumulated at an earlier state and then is enriched with H3K27ac at a later state, that means that it

becomes activated downstream in development. This method can also be used to identify how enhancer activity is different between healthy and diseased cells.

One of the main advantages of DYSE is that it provides two options: to call ROSE internally to determine SEs associated with each TF used or to skip that step, in case SEs have been previously identified (Figure 2.2). The input used for the algorithm depends on the option used to run the algorithm. These two cases are explained in more detail below. The main DYSE algorithm is called through command line using the command below. Tables 2.1 and 2.2 explain the use of each argument used in the algorithm:

```
python DYSE_main.py -i INPUT_FILES -o OUTPUT_FOLDER -t ANALYSIS_TYPE -g
      GENOME_BUILD [optional: -r ROSE_OUTPUT -s STITCH_DISTANCE]
```

Table 2.1 **DYSE_main.py required parameters**

Parameter	**Input**
INPUT_FILES	Input files to be run through ROSE or SE files to be used for analysis right away
OUTPUT_FOLDER	Directory where the output folders are to be stored.
ANALYSIS TYPE	1 – only comparing adjacent stages or cell conditions 2 – comparing all possible pairs of cell stages/conditions
GENOME_BUILD	hg18, hg19, hg38, mm8, mm9 or mm10 referring to UCSC genome annotation file (used for gene mapping). These RefSeq gene annotation files are included in the annotations folder in the ROSE algorithm repository

Table 2.2 **DYSE_main.py optional parameters**

Parameter	**Input**
STITCHING_DISTANCE	Maximum distance between two enhancer regions that are to be stitched (default is 12.5 kbp)
ROSE_OUTPUT	Name of sub-directory where ROSE output is to be stored. If not used, the input files are considered super-enhancer files and ROSE is not called

2.1.2 Using DYSE

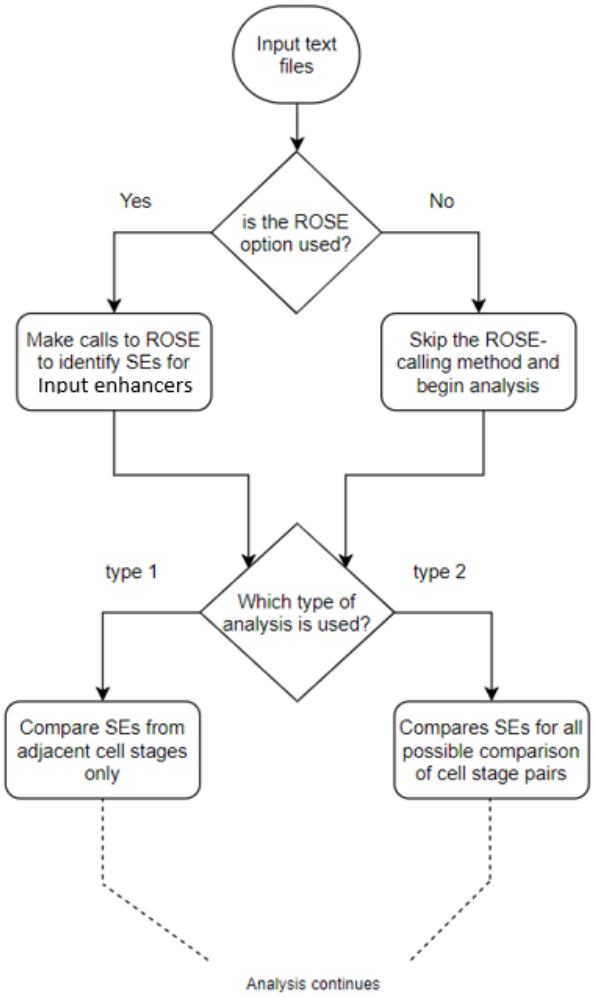


Figure 2.2 **Flowchart of the DYSE algorithm** DYSE allows the analysis to run on previously identified SEs. In that case, the ROSE-calling function is skipped, and the input data is considered to be SEs. DYSE also gives the flexibility of also choosing whether the SE activity comparisons will be between adjacent cell stages or between all different pairs of cell conditions or developmental stages

2.1.2a Input files

Case A: Using ROSE to identify SEs

Because DYSE handles multiple TF peak files, the input is taken in the form of text files listing the parameters that are to be used for the initial step of identifying SEs for each TF needed (Figure 2.1). Each input text file corresponds to a specific cell stage. In these files, each line consists of the path to the TF peak file in GFF format, the sequence alignment in BAM format for the ranking factor and the sequence alignment in BAM format for the control (optional). These files are used as input for ROSE to identify SEs as shown in Figure 2.1. An example of such files can be found in the git repository https://github.com/basmataher94/DYSE-Dynamics_of_Super_Enhancers

The way ROSE was designed, it only accepts GFF format for peak files. GFF consists of 9 columns, 6 of which are mandatory for the algorithm to work properly:

1. Chromosome number
2. Unique ID for each constituent enhancer region
3. N/A
4. Start of constituent
5. End of constituent
6. N/A
7. Strand (“+”/”-“/”.”)
8. N/A
9. Unique ID for each constituent enhancer region

Case B: Using previously identified SEs as input

In the case where SEs are pre-defined, BED files of SE peaks are used as input and the function calling ROSE is skipped. In this case the algorithm runs a lot faster as the SE identification step is the most time consuming.

2.1.2b Analysis type

In the case where more than 2 cell states/conditions are used for the analysis, DYSE gives the user two options:

1 - To do only adjacent comparisons of SEs
2 - To compare all different pairs of cell states.

As part of the analysis, the comparison between SEs of different states i.e. which ones overlap between stages and which ones are stage specific, is visualized in the form of Venn diagrams. Regardless of the number of cell states being studied, 2D Venn diagrams are generated only for adjacent cell state comparisons. DYSE does not allow for generating Venn diagrams comparing multiple states at a time because the complexity of the Venn diagrams increases with the number of datasets to compare.

2.1.3 How it works

2.1.3a Obtaining SE regions from ROSE

In the case where the ROSE option is used (case A), it is called internally from DYSE to identify SE regions for each TF individually. In the mapping step in ROSE, the sequence reads from the ranking factor and the control, if used, are mapped to all enhancer peaks (stitched and constituent ones). This identifies the density of the ranking factor and the control at each enhancer peak. If a control factor is used, the overall ranking factor density is calculated by subtracting the control factor signal density. In a use-case where DYSE calls ROSE for multiple TFs for multiple states, DYSE will take an excessively long time to run if it has to run ROSE for each TF sequentially.

During preliminary analysis where I manually analyzed myogenic data before developing DYSE, using ROSE to identify SEs using a single TF took about 20 minutes to compute on a Linux server using 16 CPUs with 16GB RAM each. By using ROSE to identify SEs for each of the TFs that I used for my analysis, it would take about 4 hours added to the time taken to perform the downstream analysis manually. This method would be very time consuming and redundant. To address this problem, I used parallel processing to run multiple ROSE calls simultaneously. This results in a significantly reduced run-time for DYSE. I use the Python package joblib which runs internal functions multiple times using different input values simultaneously. For each cell state, the ROSE calling function is called for all TFs in that cell state are run in parallel on separate threads simultaneously. Now, for 12 TFs covering 3 different myogenic stages, DYSE runs for about an hour and 45 minutes which is equivalent to about 8.75 minutes per TF including the downstream analysis.

GENOME_BUILD and STITCHING DISTANCE arguments are also used for running ROSE. The former one is the genome annotation that is used to create a loci list of all the known genes for that genome. This is used to determine TSS regions as they are excluded from the SE identification step. It is also used later for

gene mapping to identify genes associated with SEs based on proximity. The latter one is an optional argument that allows the user to define the stitching distance in case the default value is not used.

For each cell state, when SEs are identified for different TFs, the ones that are overlapping are combined into a single unit, thereby producing a single SE peak file per state in BED format. This determines which regulatory regions are co-bound by multiple TFs.

2.1.3b Determining changes in SEs between stages

Next, the SE peak files from each state are compared to identify which regions remain in an SE state as the cell develops from one stage to another. This step takes advantage of the "intersect" feature from BEDTools, which is another bioinformatics software used for ChIP-seq analysis. This feature identifies overlapping regions between multiple BED files and also shows the extent of the overlap between them. This allows us to see the similarity in SE distribution between stages.

To provide a visual representation of the results, 2D Venn diagrams are used. The Venn diagrams use the data from SE comparisons that are performed in the previous step to show the extent of SE overlaps between different stages. Since the Venn diagrams used are proportional, they show a perspective of the number of stage-specific SEs compared to multistage ones. These Venn diagrams are then stored in the form of PNG image files in another directory called "figures".

In the final step in the main DYSE algorithm, the ROSE gene mapper script is called for all stage-specific and multistage SE peak files. The gene mapper identifies SE-gene interactions using the proximity method. This is repeated for each output SE file from the analysis.

2.1.3c Differential expression of SE associated genes between stages

The DYSE algorithm provides a differential gene expression feature that uses the gene mapper output from the step explained above. Using RNA-seq data, it determines changes in the activity of SE associated

genes between different cell stages. The way this process works is by mapping cell stage-specific SE-associated genes to the RNA-seq data. This is to identify the genes that are differentially expressed between the different stages of the cell lineage. In this case, RNA-seq data comparing myoblasts, three-day differentiation and seven-day differentiation myotubes was used to compliment the three stages defined in this study. This feature can be run independently from the main DYSE program.

2.1.3d Changes in SE activity between stages

DYSE provides another feature that determines changes in SE activity between stages. When analyzing a multistage lineage, it can identify changes between each pair of adjacent stages. This means that it would identify which SEs were poised (temporarily inactive) at one stage and became activated in the following stage. Similarly, it also identifies the SEs that were active in one stage and became poised in the following stage.

For this feature SE peaks for each stage as well as H3K27me3 peaks are used as input. The method used to identify SEs that were repressed and activated between stages is by mapping SE peaks from one stage to H3K27me3 from the following stage and vice versa (Figure 2.3). The H3K27me3 histone mark is recommended to be used for this feature since it marks poised enhancers, which are temporarily inactive, as mentioned earlier. The BEDTools intersect function is used here again to search for the SE-associated genes in the list of differentially expressed genes.

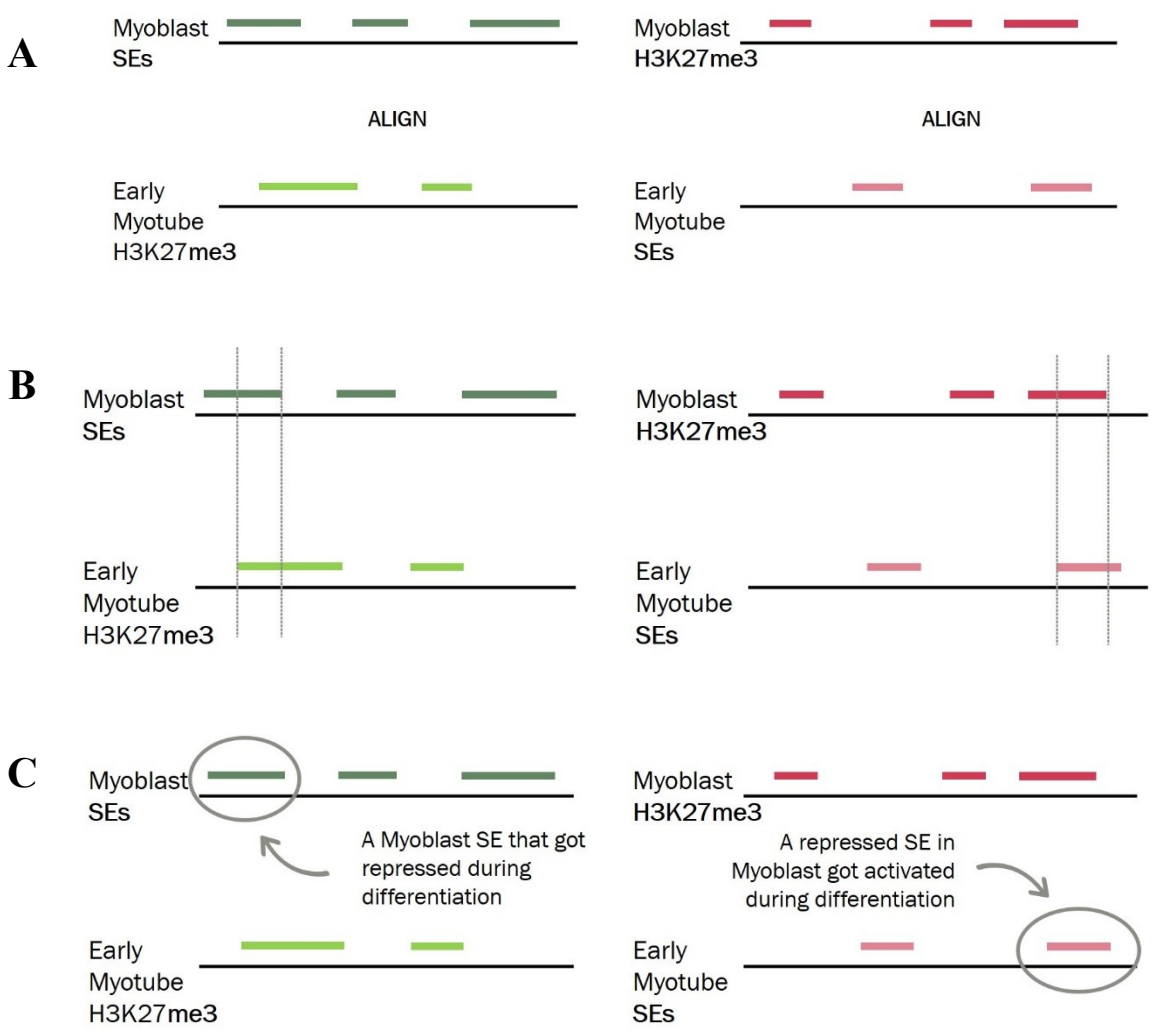


Figure 2.3 **Illustration of determining SE activity changes** This illustration shows the method used in DYSE to determine changes in SE activity between different cell states. A) The SE regions from stage A (Myoblasts) are aligned with H3K27me3 peaks from stage B (Early Myotubes) and SE regions from Myoblasts are aligned with H3K27me3 from Early Myotubes. B) The overlapping regions are identified. C) The overlapping regions represent SEs that were active and became poised (left) and those that were poised and became activated (right).

2.2 Myogenic datasets

To develop and test the DYSE algorithm, I used mouse (*Mus musculus*) myogenic data obtained from the Sequence Read Archive (SRA) on the National Center for Biotechnology Information (NCBI). I analyzed myogenic data for three stages: myoblasts, early myotubes and late myotubes.

Table 2.3 **Accession number list** This table shows the list of TFs, as well as the ranking factor and control used for myoblast, early and late myotube stages, and their accession numbers from the GEO database

Stage	Factors used	Tissue type	Accession number
Myoblast	Pax3	Primary myoblast	GSM615620
	Pax7	Primary myoblast	GSM615619
	MyoD	C2C12 myoblast	GSM1556148
	Pbx1	C2C12 myoblast	GSM2053055
	Snai1	Primary myoblast	GSM937909
	H3K27ac	C2C12 myoblast	GSM921130-31
	Input control	C2C12 myoblast	GSM721306
Early Myotube	MyoD	C2C12 48hr diff	GSM1556152
	Myogenin	C2C12 60hr diff	GSM1092114-16
	Pbx1	C2C12 3d diff	GSM2053056
	H3K27ac	C2C12 2d diff	GSM2464978
	Input control	C2C12 2d diff	GSM2971217
Late Myotube	Tead4	C2C12 6d diff	GSM2186203
	Myf6	Primary myotubes 5d diff	unpublished
	Mef2D	C2C12 5d diff	GSM1058956
	H3K27ac	C2C12 4d diff	GSM921132-33
	Input control	C2C12 5d diff	GSM2971219

The TFs we used for the analysis were of two types: known master regulators and factors that indirectly regulate cell development. ChIP-seq data for Pax3, Pax7, Snai1, Pbx1 and MyoD was used for myoblasts. Myogenin, MyoD and Pbx1 were used for early myotubes and Mef2D, Myf6 and Tead4 were used for late myotubes. H3K27ac was used as the ranking factor for the three stages. We used H3K27ac ChIP-seq data for myoblasts, 2d differentiated myotubes (early myotubes) and 5d differentiated myotubes (late

myotubes). All the datasets were downloaded from the GEO database as raw sequence reads in .sra format and pre-processed before SE analysis, as described in the next section. Table 2.3 shows the accession numbers for the datasets that were used in this study.

The datasets used for this study were chosen while taking into consideration the protocols used to obtain the data. For Pax3, Pax7, Snai1 and Myf6 DNA was extracted from C57BL/6 primary myoblasts cultured in Ham's complete medium. For the remaining TFs, DNA was extracted from C2C12 cells cultured in fetal calf/bovine serum with Dulbecco's modified Eagle's medium (DMEM; Gibco) supplemented with 1% penicillin/streptomycin. Myoblasts were extracted before reaching 100% confluence. To induce differentiation cells were cultured with DMEM and 2% horse serum (HS) and maintained till early or late myotube stage (Table 2.4).

2.2 Preparing input data for DYSE

I followed the ChIP-seq analysis pipeline described by Bardet (Bardet, et al., 2012). Raw ChIP-seq data for the different TFs were downloaded in the form of sequence reads from the sequence read archive on NCBI. Table 2.4 shows the list of studies from which the data was taken and the accession numbers linking to the ChIP-seq data files. I then used the fastq-dump feature from SRA-Toolkit software to convert the raw files into fastq format.

Table 2.4 **Growth and differentiation protocols** This table is an overview of the cell type used to extract DNA to be used for ChIP-seq of each factor and growth and differentiation protocols used for each dataset used in this study

Study	Factors and stage	Growth medium	Differentiation medium
(Soleimani, et al., 2012)	Pax3 Pax7	C57BL/6 primary myoblasts Collagen-coated plates Ham's complete medium	N/A
(Cui, et al., 2017)	MyoD myoblast And 48hr diff	C2C12myoblasts DMEM 1% pen/strep 10% FBS	DMEM 2% HS
(Dell'Orso, et al., 2016)	Pbx1 myoblast And 3d diff	C2C12 cells High-glucose DMEM 20% FBS	High-glucose DMEM 2% HS Insulin, Frasferrin, Selenium
(Joshi, et al., 2017)	Tead4 6d diff	C2C12 cells 20% FCS DMEM	DMEM 2% HS
(Soleimani, et al., 2012)	Snai1 myoblast	C57BL/6 primary myoblasts Collagen-coated plates Ham's complete medium	N/A
(Marinov, et al., 2014)	Myogenin 60hr diff	C2C12 cells DMEM	DMEM 2% Equine Serum
N/A	Myf6 5d diff	C57BL/6 primary myoblasts Collagen-coated plates Ham's complete medium	DMEM 5% HS
(Sebastian, et al., 2013)	Mef2D 5d diff	C2C12 myoblasts DMEM 10% FCS	DMEM
(Blum, et al., 2012)	H3K27ac myoblast and 4d diff	C2C12 myoblasts DMEM 10% FBS	DMEM 2% HS
(Peng, et al., 2017)	H3K27ac 2d diff	C2C12 myoblasts	DMEM 2% HS

(Asp, et al., 2011)	Input control mb	C2C12myoblasts DMEM 10% FBS	N/A
(Anan, et al., 2018)	Input control 2d and 5d diff	C2C12 cells DMEM 10% FBS	DMEM 2% HS 25mM glucose

Next, I mapped the sequence fragments to the mm10 reference genome using Bowtie2 and generated sequence alignment/map (SAM) files. Eight alignment threads were used for all sequence alignment processes. The sequence alignment files were then sorted and converted to the binary version (BAM file format) using SAMtools software with the default values. For the TF datasets, the MACS2 algorithm, which identifies read-enriched regions from ChIP-seq data (Feng, et al., 2012), was used with default parameters for peak calling.

The peak files generated by MACS2 were then converted to GFF file format. The main difference between these two formats is the order of the columns and the information provided. I used a text-editing script to simply re-order the columns to obtain the GFF format. This step is required as the ROSE algorithm depends on the GFF file format for input TF enhancer files. As peak files are not needed for the ranking factor (H3K27ac), based on the ROSE algorithm, I left them as binary alignment files in BAM format.

CHAPTER 3 – Results

3.1 Characterizing myogenic SEs

For the implementation of DYSE, I covered both options of using ROSE from within the algorithm to identify SEs or using previously identified SE bed files as the starting point. The way I did this was I did the run for DYSE using ROSE and then used the output SE files generated as the starting point for a second run for DYSE without ROSE. This generated the same results for the downstream analysis and from the second run, I noted that the analysis takes around four to five minutes to run separately. I also covered the option of comparing SEs between adjacent states and comparing all different cell states. For this option, the only difference would be that I would have proportional Venn diagrams to present the extent of SEs that are shared between adjacent stages only compared to all possible pairs.

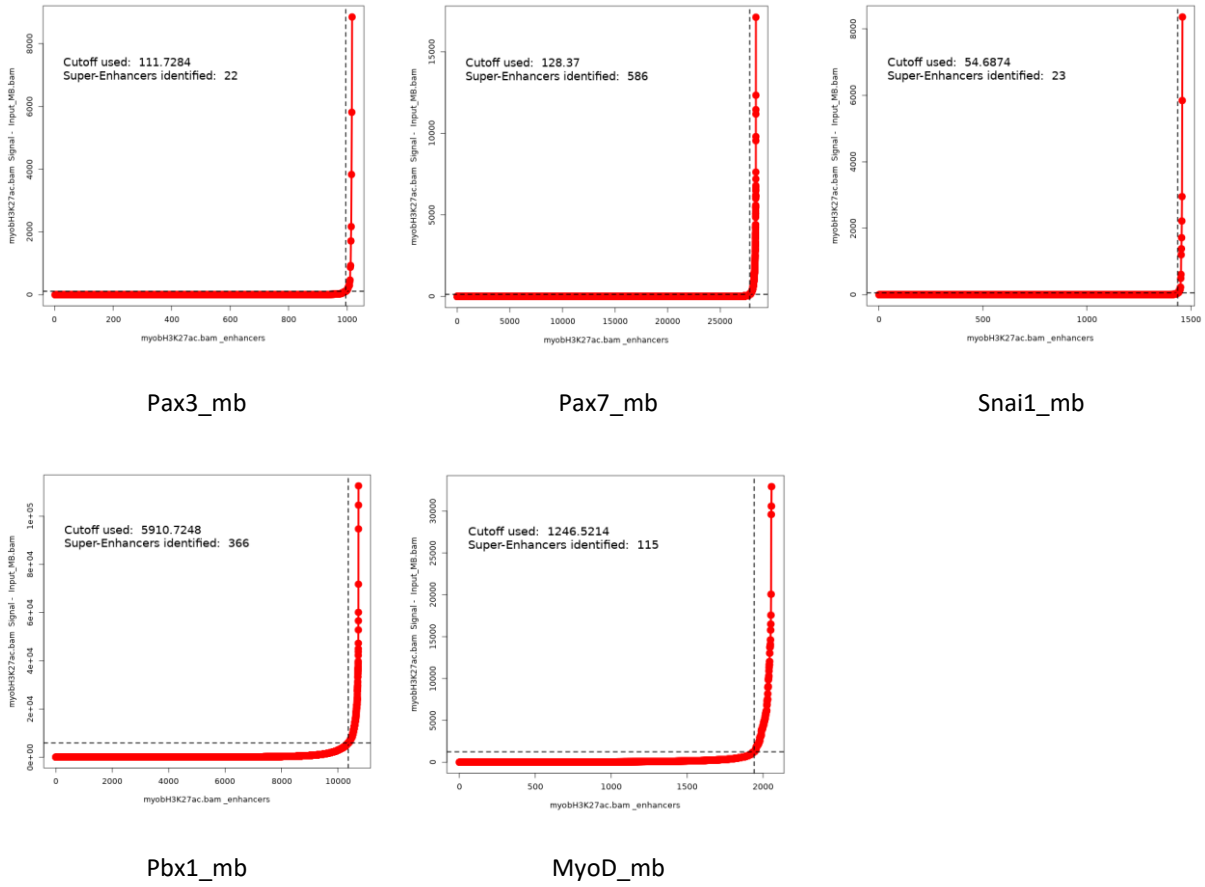
Table 3. 1 **SE stats** This table shows the number of putative enhancers corresponding to each TF used in this study and the total number of combined SEs for each stage.

Stage	TF	Num_peaks	Num_SEs	Total Enhancers	Grouped SEs	Stage-specific SEs
Myoblast	Pax3	1,360	19	42,545	973	623
	Pax7	38,699	535			
	Snai1	1,793	23			
	Pbx1	19,304	366			
	MyoD	2,679	115			
Early myotubes	Myogenin	42,512	1015	34,150	1,610	767
	MyoD	16,051	635			
	Pbx1	19,644	700			
Late myotubes	Myf6	13,634	896	15,401	1,264	595
	Mef2D	9,549	623			
	Tead4	364	6			

In the SE identification step in the DYSE algorithm, out of 42,545 enhancer peaks in myoblasts, 973 were characterized as SEs. In early myotubes, out of 34,150 enhancer peaks, 1,610 SEs were identified. In late myotubes, out of 15,401 enhancer peaks, 1,264 SEs were identified. The breakdown of these numbers among different TFs in each stage is presented in Table 3.1.

As will be described further along, some of these SEs were made up of multiple constituent enhancers bound by different TFs; meaning, they were co-occupied by multiple TFs. These regions, as mentioned earlier, were combined into a single unit. Thus, the sum of the SEs for all TFs in each stage is less than the total number of SEs. Figure 3.1 shows the SE plots generated by the ROSE algorithm from each TF used in the analysis.

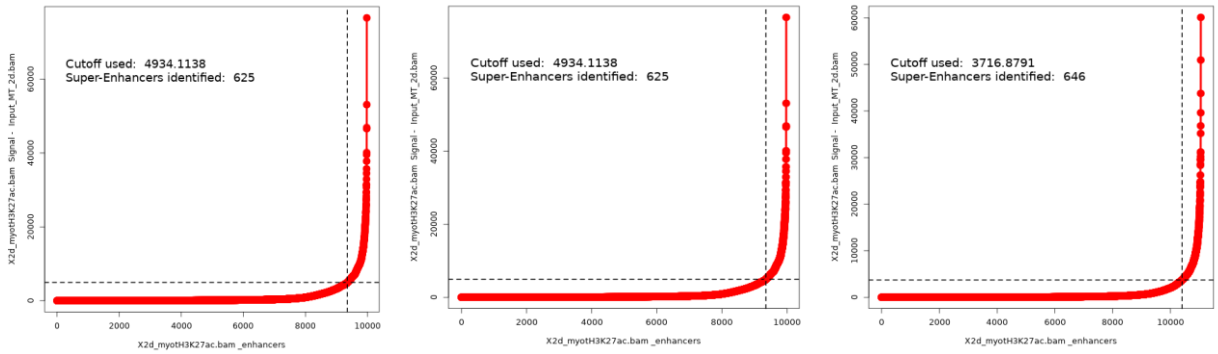
A



Pax3_mb Pax7_mb Snai1_mb

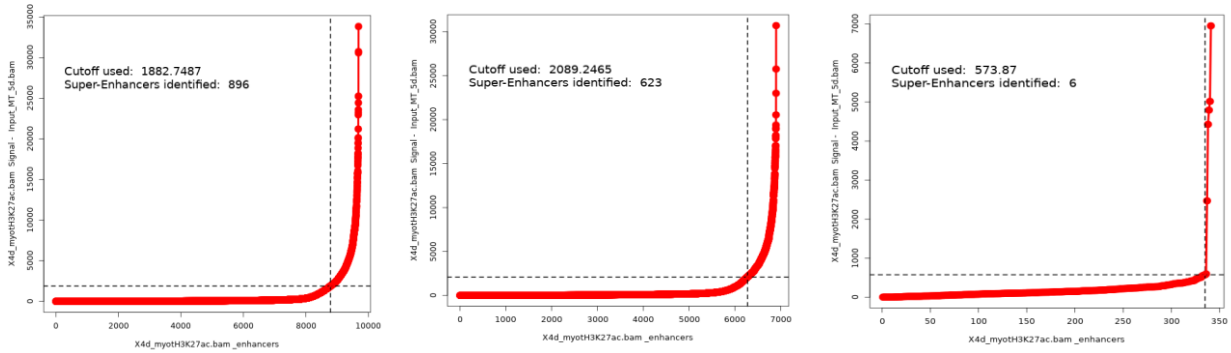
Pbx1_mb MyoD_mb

B



MyoD_2d Myog_60hr Pbx1_3d

C



Myf6_5d Mef2D_5d Tead4_6d

Figure 3.1 **ROSE SE plots for all TFs** In the third step in the SE identification method described by Whyte et al. (2013) all enhancers are ranked based on their H3K27ac signal density and plotted. The point of inflection is then calculated to distinguish between typical and super enhancers. Here I present the SE plots generated for each TF in (A) myoblasts, (B) early myotubes and (C) late myotubes

The output files from ROSE representing all SEs for each stage can be found in the git repository https://github.com/basmataher94/DYSE-Dynamics_of_Super_Enhancers . From the SE regions that were identified, some consisted of a single enhancer but had much higher H3K27ac signal density compared to

other TEs. In myoblasts, 73 SEs consisted of single enhancers. In early myotubes, 4 SEs were made up of a single enhancer and in late myotubes 1SE was made up of one enhancer (Table 3.2). However, the general trend observed in my results is that SEs are composed of multiple enhancers stitched together into one big domain. On the other hand, there were stitched enhancers that covered a large genomic area, yet, did not have an H3K27ac signal high enough to be categorized as an SE.

Table 3.2 **Single-enhancer SEs** This table shows the number of regions in each myogenic stage that are composed of a single enhancer region but have a strong enough H3K27ac signal to be considered SEs

Stage	**TF**	**Num of SE**
Myoblast	MyoD	2
	Pax7	50
	Pax3	10
	Snai1	11
Early myotube	Myog	1
	Pbx1	3
Late myotube	Mef2D	1

After SE regions were identified for all TFs in a given cell state using the ROSE algorithm, SEs that overlapped were combined into one regulatory region. This was done while keeping track of the TFs corresponding to the regions that overlapped. My results showed that in myoblasts, only 81 SEs were co-occupied, mostly by Pbx1 with MyoD (29 SEs) or with Pax7 (20 SEs). Furthermore, the majority of Pax3 SEs (16/22 SEs) overlapped with Pax7. In early myotube stage, almost a third of the identified SEs were bound by more than one TF. 60% of these were co-bound by MyoD and Myogenin. The rest were either co-bound by Pbx1 and MyoD or Pbx1 and Myogenin. In late myotubes, out of the 6 Tead4 SEs that were identified 2 co-occupied Myf6 SEs. Another 254 Myf6 SEs were co-bound by Mef2D.

In my analysis, I used ChIP-seq data for MyoD in both myoblast and early myotube stages. My results show that MyoD gives rise to five times more SEs in early myotubes compared to myoblasts. Similarly, Pbx1 gives rise to three times more SEs in early myotubes. On the other hand, Tead4 produced very few peaks in late myotubes an and almost no peaks in myoblasts. Moreover, Pax3 and Snai1 produced very few SEs in the myoblast stage (Table 3.1).

About half of the total grouped SEs identified at each stage were found to be stage-specific (Table 3.1, Figure 3.1). As for the remaining SEs, some of them are barely overlapping, while others almost fully overlap. In our algorithm, when we compare SEs from different stages, any overlap between the genomic region of an SE in stage A and an SE in stage B is considered an overlap regardless of the extent. And these regions are interpreted as SEs that are active in both stages. The most overlaps occur between early and late myotube stages, while the least overlaps occurred between myoblasts and late myotubes (Figure 3.2).

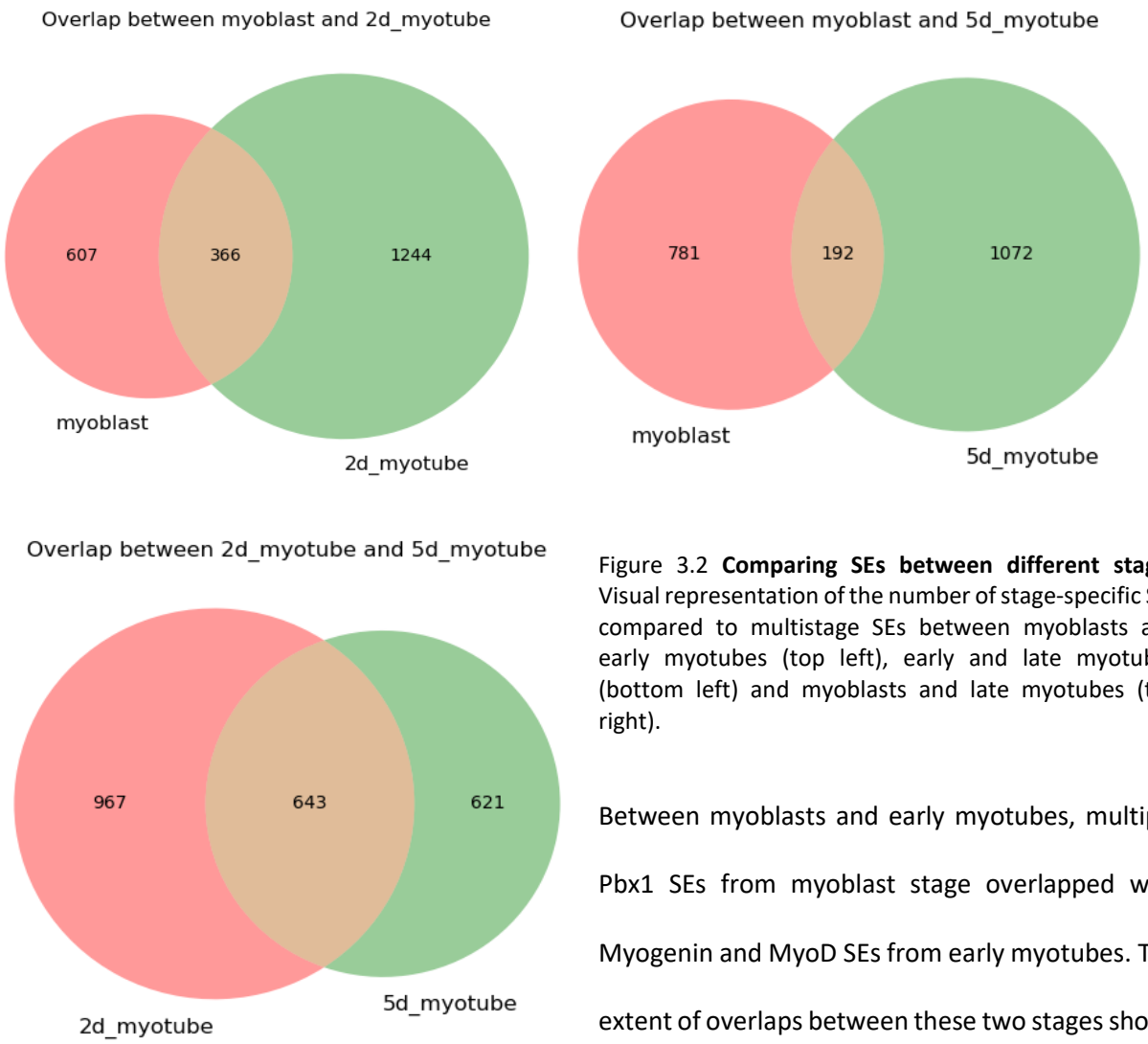


Figure 3.2 **Comparing SEs between different stages** Visual representation of the number of stage-specific SEs compared to multistage SEs between myoblasts and early myotubes (top left), early and late myotubes (bottom left) and myoblasts and late myotubes (top right).

Between myoblasts and early myotubes, multiple Pbx1 SEs from myoblast stage overlapped with Myogenin and MyoD SEs from early myotubes. The extent of overlaps between these two stages shows that the majority of the multi-stage SEs that are active in myoblast and early myotubes have a high extent of overlap, more than 95%. Between early and late myotubes, most of the shared SEs show a complete overlap where the late myotube SEs fall within early myotube SEs. Most of these overlaps involve Myogenin SEs from early myotubes. Unlike these two cases, myoblast and late myotube do not share many SEs and there was no visible trend in these overlaps.

3.2 Myogenic SE associated genes

In the following step DYSE uses all the SE peak files for both stage-specific and multi-stage SEs to identify genes that are associated with them. DYSE uses the gene mapper script from ROSE for this step, and the ROSE gene mapper determines whether or not a gene is associated with an SE based on proximity. DYSE identified 812, 1163 and 1111 genes that are linked to myoblast, early and late myotube stage-specific SEs, respectively.

Many of the genes that were associated with SEs that code for proteins that have been shown to be associated with myogenic developmental processes and pathways such as Ezh2, Neat1, Malat1, Snai1, Mir206, Myogenin and Myod1 (Hou, et al., 2012; Lu, et al., 2013). Surprisingly, however, major myogenic transcription factors were found to be associated with SEs that were not stage-specific. MyoD, Myogenin, Tead4 and Mef2D were found to be affiliated with SEs that are shared between early and late myotubes. Similarly, MyoD and Pbx1 were found among the genes affiliated with SEs shared between myoblasts and early myotubes. However, we did not find the Pax3 or Pax7 genes to be associated with any of the identified SEs across the three myogenic stages. We compared the stage-specific SE associated genes for early and late myotubes with the ones that were previously defined by Whyte and colleagues (Whyte, et al., 2013) and found that our results recognize about half of the previously identified genes (Table 3.3).

Table 3.3 SE associated genes Presented here are the previously identified myotube SE associated genes that were recognized in our analysis. (data obtained from dbSuper database *http://asntech.org/dbsuper/index.php*)

Chrom	Start	End	Gene	Early myotubes	Late myotubes
chr1	88330992	88337879	1700019O17Rik	✓	✓
chr1	91435101	91435589	Agap1	✓	
chr1	136091724	136093112	Mybph		✓
chr1	157077101	157081675	Stx6	✓	
chr1	172585526	172595867	Olfml2b	✓	
chr10	12540833	12543578	Utrn	✓	
chr10	25253621	25254155	Smlr1	✓	
chr10	45031396	45031863	Popdc3		✓
chr10	94012730	94031013	Tmcc3	✓	
chr11	53876192	53881532	Pdlim4		✓
chr11	58974937	58975364	Iba57		✓
chr11	66919161	66933530	Myh3		✓
chr11	88044607	88047774	Mrps23	✓	✓
chr11	88364523	88373381	Msi2		✓
chr11	102859929	102865040	Dcakd	✓	
chr11	116147026	116158026	Galr2		✓
chr11	119482621	119487108	Rptor	✓	
chr12	74342928	74343471	Trmt5	✓	
chr12	80145745	80162583	Plekhh1		✓
chr12	80612442	80613123	Rad51b	✓	
chr12	81224145	81225603	2310015A10Rik	✓	
chr12	85483217	85496481	Pnma1		✓
chr12	86568017	86568638	Eif2b2	✓	
chr12	114300276	114300903	Tex22	✓	
chr13	29829938	29838827	Cdkal1		✓

chr13	34202867	34204093	Tubb2b	✓	
chr13	37749877	37784339	Rreb1	✓	
chr13	46618074	46630793	Cap2		✓
chr13	49566135	49588504	Ecm2	✓	
chr13	52270976	52271494	Gadd45g	✓	
chr13	63209906	63214314	2010111I01Rik	✓	✓
chr13	98534106	98545417	Arhgef28	✓	
chr13	100029380	100032682	Zfp366	✓	
chr13	103677682	103690449	Cd180		✓
chr14	41769164	41776575	Tspan14	✓	
chr14	55616110	55624664	Myh7		✓
chr14	61747811	61748358	Sacs	✓	
chr15	7068879	7074377	Lifr	✓	
chr15	25691500	25704859	Fam134b	✓	
chr15	31137042	31137774	Dap	✓	
chr15	34336890	34342704	Rpl30	✓	
chr15	36802745	36823720	Ywhaz	✓	
chr15	38380691	38382979	Azin1		✓
chr15	41748393	41749002	Abra		✓
chr15	54960277	54965601	Deptor	✓	
chr15	55798276	55807810	Sntb1	✓	
chr15	63949071	63949832	Fam49b		✓
chr15	72949865	72962726	Chrac1		✓
chr15	76315656	76316013	Hsf1	✓	
chr15	76804578	76810538	1110038F14Rik		✓
chr15	76844203	76853374	Mb		✓
chr15	76986946	76994360	Rbfox2	✓	

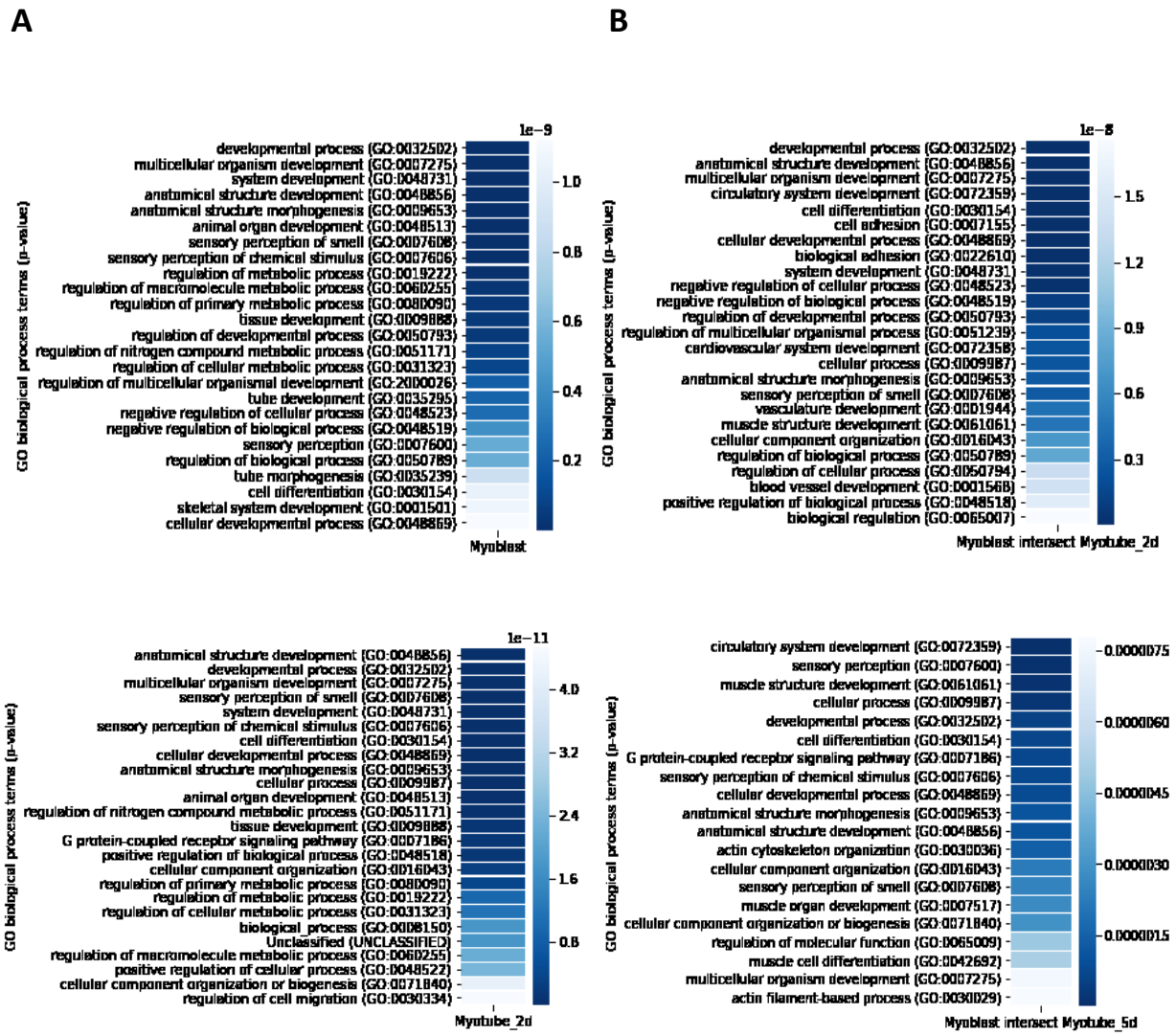
chr15	79234778	79249576	Tmem184b	✓	
chr15	81965772	81978537	Srebf2	✓	
chr15	83299746	83300385	Pacsin2	✓	
chr15	99182647	99183234	Fmnl3	✓	
chr15	99270768	99271924	Nckap5l	✓	
chr15	103000383	103003456	Smug1	✓	
chr15	103045629	103058332	Cbx5	✓	
chr16	4898828	4905886	Mgrn1		✓
chr16	24205604	24238685	Lpp	✓	✓
chr16	44238798	44239427	Gm608		✓
chr16	45336527	45337011	Cd200	✓	
chr16	84775095	84783893	Jam2		✓
chr17	16355489	16355982	Rgmb	✓	
chr17	25793086	25793484	Gng13	✓	
chr17	32052595	32053194	Sik1	✓	
chr17	32155000	32160396	Hsf2bp	✓	
chr17	44641879	44645509	Runx2	✓	
chr17	48569744	48577275	Apobec2		✓
chr17	64897842	64908896	Man2a1	✓	
chr17	68395629	68396019	Arhgap28	✓	✓
chr17	71388049	71388574	Myom1		✓
chr17	74663754	74670076	Memo1		✓
chr17	79754961	79764354	Cdc42ep3	✓	
chr17	84972672	84984813	Dync2li1	✓	✓
chr18	32538638	32551135	Bin1		✓
chr18	47314415	47330091	Commd10	✓	
chr18	67533767	67535129	Tubb6	✓	

chr18	77293359	77293939	Katnal2			✓
chr19	46260988	46261441	Gbf1	✓		
chr19	46592457	46593355	Trim8	✓		
chr19	47296888	47297499	Neurl1a	✓		
chr19	57047226	57059060	Afap1l2	✓		
chr2	6793366	6798573	Celf2	✓		
chr2	14058519	14058954	Tmem236	✓		
chr2	27326473	27327309	Brd3			✓
chr2	31912052	31920913	Fam78a			✓
chr2	32902511	32913479	Garnl3	✓		
chr2	49597473	49605922	Lypd6b	✓		
chr2	49961062	49961852	Lypd6	✓		
chr2	59523262	59523766	Tanc1	✓		
chr2	84474966	84486674	Ctnnd1	✓		
chr2	90875470	90875914	Rapsn	✓		
chr2	113961565	113969987	Aqr			✓
chr2	167142402	167142958	B4galt5			✓
chr2	173082951	173090699	Pmepa1			✓
chr2	180097287	180111708	Mir1a-1			✓
chr3	14787076	14788065	Car1	✓		
chr3	55054591	55077530	Dclk1			✓
chr3	88419788	88434464	Arhgef2	✓		
chr3	116631829	116632405	Palmd	✓		
chr3	121177601	121184847	Cnn3	✓		
chr3	121791170	121799627	Abca4	✓		
chr3	122116624	122118122	Bcar3	✓		
chr4	57614011	57624497	Palm2	✓		

chr4	63203102	63207827	Atp6v1g1	✓	
chr4	63708218	63720308	Tnc	✓	
chr4	95193273	95193666	Fggy	✓	
chr4	122903175	122905943	Heyl	✓	
chr4	129505401	129507350	Ptp4a2	✓	
chr4	132569234	132569775	Ahdc1	✓	
chr4	135720901	135734258	E2f2	✓	
chr4	151641691	151642242	Gpr153		✓
chr5	15472737	15478114	Cacna2d1	✓	
chr5	32249515	32253422	Fosl2	✓	
chr5	36715787	36734743	Sorcs2	✓	
chr5	64727997	64728401	Tbc1d1	✓	
chr5	125724271	125734679	Ncor2	✓	
chr5	134482580	134483038	Gatsl2		✓
chr5	136998708	137008373	Cux1		✓
chr5	137345694	137348829	Col26a1		✓
chr5	139170001	139170558	Fam20c	✓	✓
chr5	143736353	143736873	Fscn1	✓	
chr6	29385786	29389676	Flnc		✓
chr6	30570991	30578518	Cpa5	✓	
chr6	54319803	54320200	Prr15	✓	
chr6	67080571	67089117	Gadd45a		✓
chr6	72204026	72205184	Atoh8	✓	
chr6	90648777	90661936	Iqsec1	✓	
chr6	94587095	94610034	Lrig1		✓
chr7	51727407	51729053	Josd2		✓
chr7	74518180	74530835	Mef2a	✓	

chr7	119372763	119378341	Mical2		✓
chr7	149736644	149737393	Nctc1	✓	
chr7	149786291	149787122	H19	✓	
chr8	11382200	11387591	Col4a2	✓	
chr8	24249253	24249689	Mir486	✓	
chr8	55633629	55635917	Asb5	✓	
chr8	71656014	71659722	Lzts1		✓
chr8	73451917	73452368	Mir1969	✓	
chr8	81750984	81751500	1700011L22Rik	✓	
chr8	88024912	88029070	Gpt2		✓
chr8	125175309	125175878	Cbfa2t3		✓
chr8	129383532	129401142	Tomm20	✓	
chr9	14909520	14924332	Hephl1		✓
chr9	40266105	40277424	Gramd1b		✓
chr9	61281763	61287386	Tle3	✓	
chr9	62379528	62389526	Coro2b	✓	
chr9	66063206	66072379	Dapk2	✓	
chr9	67342049	67342369	Tln2	✓	
chr9	77485037	77485619	Klhl31		✓
chr9	94560055	94571076	Slc9a9	✓	
chr9	108175116	108182700	Dag1		✓
chrX	93315346	93327520	Msn	✓	✓

After I used DYSE to find SE associated genes, I used gene ontology (GO) enrichment analysis to identify the biological processes these genes show enrichment for. The GO analysis showed enrichment for myogenic development as well as other development and metabolic processes (Figure 3.3). The outcome showed that multi-stage SE-associated gene showed more enrichment for myogenic processes than for stage-specific ones.



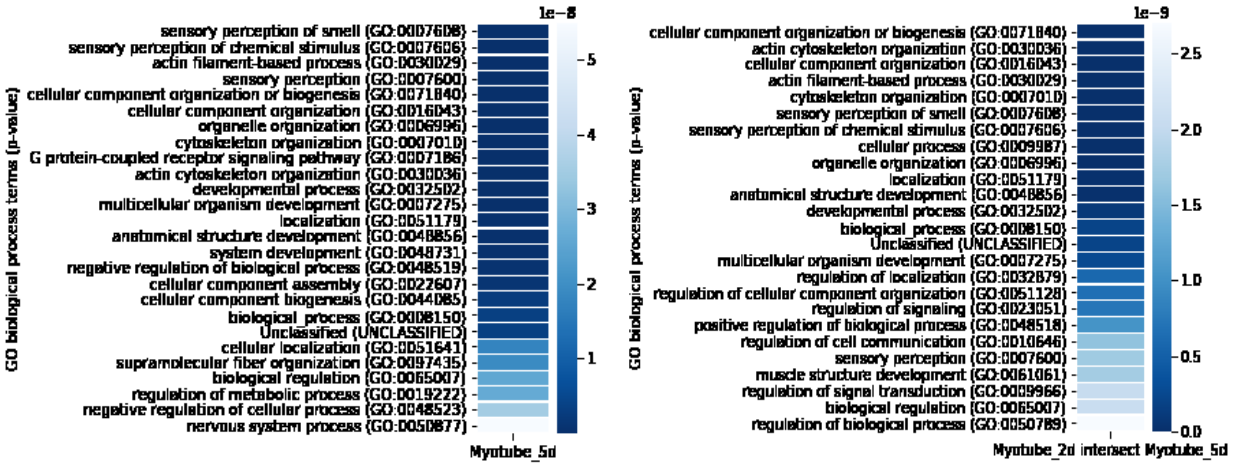


Figure 3.3 Gene ontology term enrichment The top 25 GO terms for biological processes that SE-associated genes show enrichment for in A) stage-specific SEs in myoblasts, early and late myotubes and B) multi-stage SEs active in myoblasts and early myotubes, early and late myotubes, and myoblasts and late myotubes. The GO terms are ranked by their calculated p-values using the online software GREAT (Genomic Regions Enrichment of Annotations Tool)

Next, I used the differential expression feature in DYSE to determine how the expression level of SE associated genes differs between stages. For this, I used published RNA-seq data for myogenic genes as input (Doynova, et al., 2017). The RNA-seq data used showed differential expression data between myoblasts, three-day differentiation myotubes and seven-day differentiation myotubes.

I applied this function to multi-stage SEs to see how their expression changes between the stages. My results showed that for SEs that are shared between myoblasts and early myotubes, out of the 574 associated genes which were defined using proximity, 98 are upregulated in myoblasts and 116 were upregulated in early myotubes. As for SEs shared between early and late myotubes, out of 1128 genes 189 were upregulated in early myotubes, while 65 were upregulated in late myotubes. For genes associated with SEs shared between myoblasts and late myotubes, out of 307 genes, 64 and 58 were upregulated in myoblasts and late myotubes, respectively (Table 3.4).

In an interest to find out whether stage-specific SE-associated genes were exclusively upregulated in their specified stage, I applied the same test for stage-specific SE associated genes. I found that for each stage

more than half of the differentially expressed genes were upregulated in their specified stage (Table 3.5). In myoblasts 226 of the differentially expressed genes were upregulated in myoblasts while 181 were upregulated in either early or late myotubes. For early myotubes, 251 genes were upregulated compared to 202 that were downregulated in early myotubes. For late myotubes, 58 genes were upregulated in late myotubes while 39 were downregulated.

Table 3.4 **Differentially expressed genes** This table shows the number of genes associated with multi-stage SEs that are differentially expressed between these stages

Stage A	**Stage B**	**Total num of SE genes**	**Upregulated in stage A**	**Upregulated in stage B**
Myoblast	Early myotubes	574	98	116
Early myotubes	Late myotubes	1128	189	65
Myoblasts	Late Myoblasts	307	64	58

Table 3.5 **Differentially expressed genes** This table shows the number of genes associated with stage-specific SEs that are differentially expressed between the three myogenic stages

Stage A	**Total num of SE genes**	**Upregulated**	**Downregulated**
Myoblast	812	226	181
Early myotubes	1163	251	202
Late Myotubes	1111	58	39

3.3 Changes in SE activity between stages

As explained in the previous subsection, the DYSE algorithm showed that some of the SEs were shared between stages and others are stage-specific. These SEs were analyzed further to determine changes in their activity as the cell transitions from one stage to another. The histone mark H3K27me3 was used in this step, as it is widely known to be affiliated with poised enhancers (enhancers that are in a temporarily inactive state). When there is more accumulation of this mark at enhancer regions it shows that it is in a

poised condition. In the context of this study, that would mean that either an enhancer is temporarily inactivated to be activated in a later stage. Or, it can also mean that an active enhancer was repressed.

For this feature, the TF or histone mark that is associated with inactive enhancers is used as input for each stage along with the list of SEs in BED file format for each stage being studied. The algorithm uses a "cross-alignment" method. In other words, as shown in Figure 2.3, DYSE identifies which H3K27me3 peaks overlap with SEs from the adjacent cell stages. My results showed that six early myotube SEs were poised due to H3K27me3 in the myoblast stage, while seven myoblast SEs were active and were repressed by H3K27me3. As for late myotubes, seven SEs were initially in a poised state in early myotubes and became activated. On the other hand, 3 early myotube SEs were repressed by H3K27me3 in late myotubes. This data is represented in more detail in Figure 3.4.

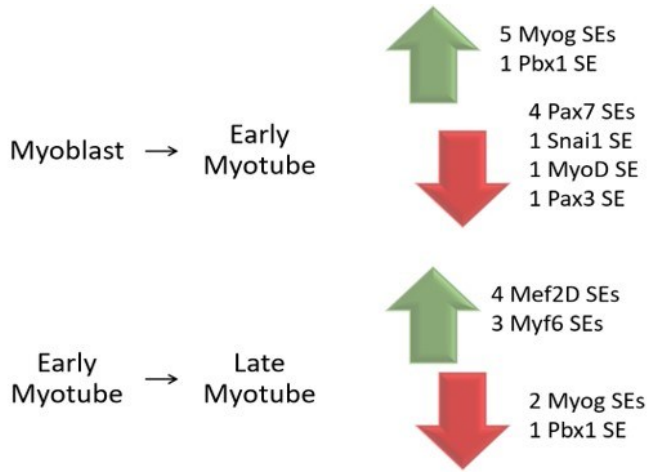


Figure 3.4 **Changes in SE activity** This illustration shows the number of SEs that were activated – had an H3K27me3 peak and in the former stage and in the same region gained a SE in the latter stage – (green) and repressed – had a SE in the former stage and then shows an H3K27me3 peak in the same region in the latter stage – (red) between myoblasts and early myotubes (top) and early and late myotubes (bottom)

I used a software called Integrative Genome Viewer (IGV) to represent the change in H3K27ac peaks at different SEs to show examples of changes in SE activity across different stages (Figure 3.5).

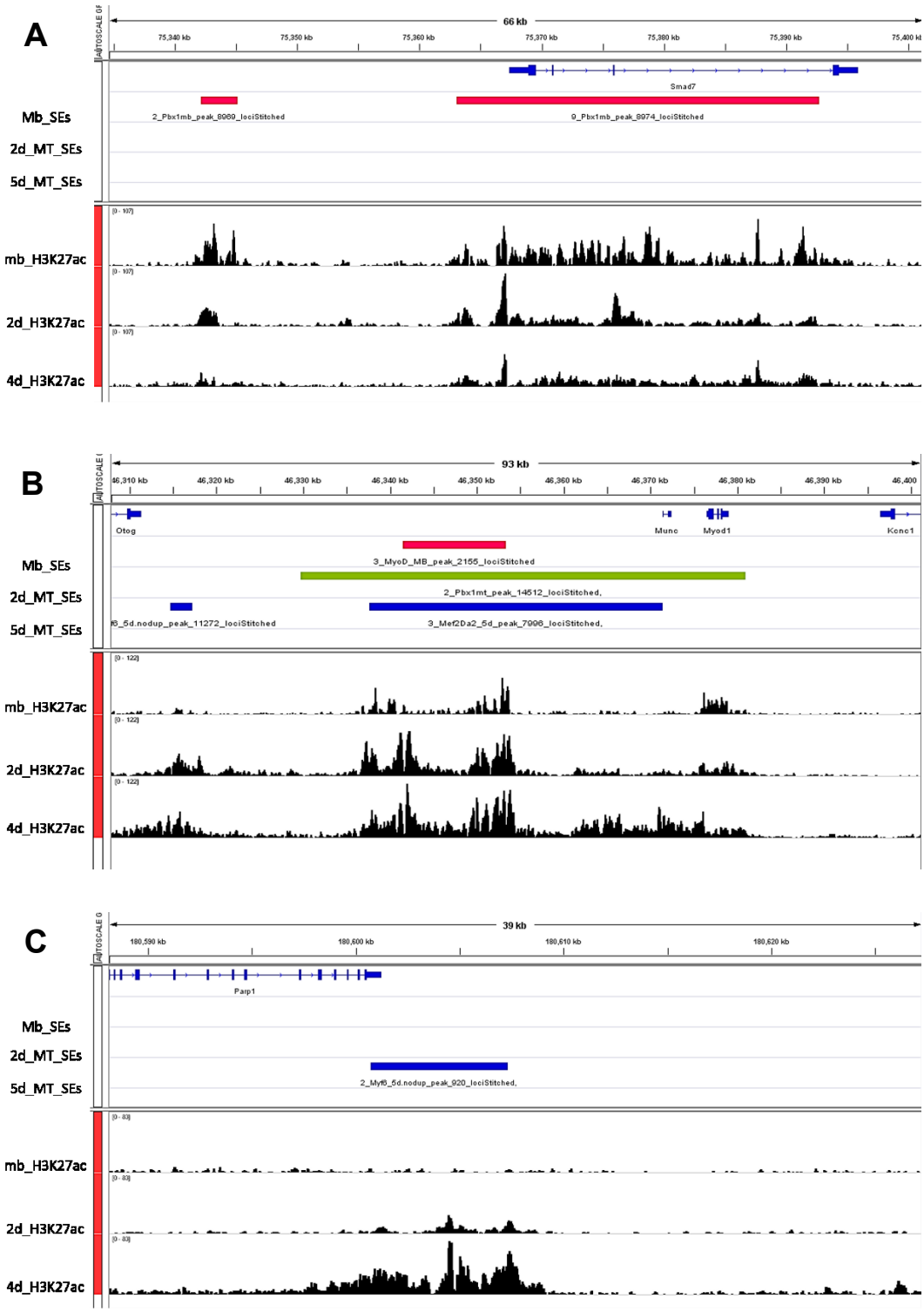


Figure 3. 5 **Visualizing change in SE activity** IGV (integrative genome viewer) representation of H3K27ac enrichment. The SE tracks that were generated show changes in SE activity between myogenic stages. In some cases they are activated as the cell develops and in other cases they get deactivated. Here we show an example of a A) myoblast SE, B) multistage SE and C) late myotube SE

CHAPTER 4 – Discussion

Multiple studies focused on understanding genome wide regulation networks in hope of understanding the role of various genes in cellular processes and pathways and how their misregulation is correlated to different diseases (Lee & Young, 2013; Izumi, 2016). A few years ago, Whyte and colleges characterized a subset of regulatory elements as super enhancers (SEs) (Whyte, et al., 2013). These SEs are, in a sense, large scale enhancers. They are large regulatory domains, usually consisting of multiple enhancers, with disproportionate binding of TFs and high H3K27ac density (Hnisz, et al., 2013; Love´n, et al., 2013; Whyte, et al., 2013). Another characteristic of SEs is that they mostly regulate genes that play a major role in defining cell identity and development (Lovén, et al., 2013).

The interest in understanding SEs and how their associated genes are involved in developmental and cell-type specific pathways has risen over the past few years, since SEs were first characterized. The previously published ROSE (Rank Ordering of Super Enhancers) algorithm (Lovén, et al., 2013; Whyte, et al., 2013) presents a useful tool that identifies SEs in a given cell type using a specified ranking factor.

Majority of SE studies focus on analyzing SE-associated genes and mutations in distinct cells in a single time phase, and many times they use single TFs to identify SEs. However, recently, multiple studies have shown interest in identifying SEs in multiple cell stages and different cell states (Adam, et al., 2015; Peng, et al., 2017). Some are focused on comparing SEs and their associated genes between healthy and diseased cells and others focus more on changes in SEs over multiple cell stages (Adam, et al., 2015; Lopes-Novo, et al., 2018). This thesis introduces DYSE, the first computational tool specifically designed for the comparison of SEs in multiple cell types or conditions.

I tested my DYSE algorithm using muscle development as my model and investigating three stages of myogenesis: myoblasts, early myotubes and late myotubes. Seeing that my results complimented current findings, I then automated my method into a Python algorithm that builds onto ROSE to automate analysis of SE dynamics in various cell lineages.

The advancement that DYSE delivers is that it takes SE analysis to another level and uses ROSE on a wider scale to investigate the remodeling of SEs throughout development. It does so by running the analysis using multiple TFs over multiple stages. Also, DYSE is a generic algorithm, which means that it can potentially run on any cell lineage and it allows analysis using multiple TFs for multiple stages.

The main focus of this study was to look at SEs as a dynamically changing system. Understanding how SE activity changes from state to state throughout development would shed light on how SEs are involved in maintaining cell identity pathways and how they control the triggering of pluripotent cells to differentiate. I believe that the dynamic changes SEs undergo as the cell transitions from stage to stage carries a lot of information about the cellular processes and genes involved in these changes.

There have been various thoughts about the functionality and definition of SEs since it was first introduced in 2013 (Hamdan & Johnsen, 2017). It hasn't yet been determined whether or not SEs are functionally separate entities from TEs (Pott & Lieb, 2015; Moorthy, et al., 2017). However, studies have proven that, using the current method (Figure 4.1) presented in ROSE to define SEs (Hnisz, et al., 2013; Lovén, et al., 2013; Whyte, et al., 2013) identifies a sub-category of enhancers that is likely to play a significant role in determining cell fate and regulating key gene expression programs.

The patterns observed in TF co-binding complemented the myogenic transcriptional regulatory network described earlier in the introduction. Pax3 and Pax7 TFs co-occupied some of the identified SEs in the myoblast stage. Pbx1 which is a TF known to indirectly influence myogenic regulation was found to co-

occupy multiple SEs that were bound by the MRFs MyoD or Myogenin in the myotube stage. To add to that, multiple SEs in early myotube stage were co-bound by MyoD and Myogenin.

The results presented reflect known interactions between myogenic TFs. As mentioned earlier, as myoblasts prepare for differentiation, Pbx1 interacts with MyoD to ensure that it stably binds to Myogenin to induce differentiation. As for the genes that are regulated by SEs, our results showed that Myogenin, MyoD, Tead4 and Mef2D were regulated by SEs that are active in both early and late myogenic stages. On the other hand, MyoD and Pbx1 were regulated by SEs that were active in both myoblasts and early myotubes. Contrarily, Pax3, Pax7 and Snai1 which are major players in the regulation of myoblasts were not found to be associated to any SEs.

My results reveal that SEs do potentially regulate cell identity genes in various stages of development and complement know patterns in myogenic TF activity. Compared to the results from Whyte et. al's study very few myotube SE regions overlapped. However, about half of the myotube SE-associated genes that were previously identified (Whyte, et al., 2013; Khan & Zhang, 2016) were recognized in our analysis as SE-associated genes in either early or late myotube stages (Table 3.3).

The reason for that could be due to the differences in the method used to execute the analysis (Whyte, et al., 2013). In our case, we used MyoD as well as other myogenic regulatory factors to identify enhancer regions and then used H3K27ac as a ranking factor to determine the cutoff that differentiates between TEs and SEs. On the other hand, Whyte and colleagues used MyoD as both the enhancer-determining factor and the ranking factor used to differentiate between TEs and SEs. Hence DYSE is likely to have identified a wider range of SE associated genes that may have not been recognized in the previously published data. The fact that many of SE associated gene that were previously identified were also identified in this study implies the importance of these genes during myogenic development and cell differentiation.

To add to that, GO (gene ontology) analysis showed that the SE-associated genes that were identified in this study were correlated to myogenic processes and cell components. They showed enrichment for GO terms associated with development and skeletal muscles (Figure 3.3).

For differential expression analysis, we ran the analysis on multi-stage SE-associated genes. The results showed that a subsection of the genes that are associated with shared SEs are differentially expressed. That may reflect that either the remaining SE associated genes are equally active in both stages or that the SEs may target different genes in different stages altogether.

On the other hand, for stage-specific SE associated gene the results were different from what I expected. Given that these genes are associated with SEs that are specific to one stage, I expected that the genes corresponding to each stage would be upregulated in that stage compared to the other stages. However, that was not the case that was observed. The results for this section showed that the differentially expressed genes were almost evenly distributed with respect to whether they are up or down regulated. In myoblasts for example, around half of the differentially expressed genes were downregulated in myotubes and the other half was upregulated in myotubes.

The exact cause behind these results is not clear. However, possible causes could be that the SE-gene association may not have been very accurate given that proximity was used to associate SEs to their target genes. Moreover, a recent study has shown that the activity of SEs in different genes does not necessarily mean that they will act on the same targets. On the contrary, they may act upon different distant promoters that are in close physical proximity (Novo, et al., 2018).

Other more advanced methods like chromatin conformation capture (3C) may be used as an alternative to proximity to locate interactions between SEs and their target genes. In 3C, identification of potential interactions between DNA segments is dependent on physical proximity of these regions in the 3D space (Furey, 2012). An adaptation of this method (Hi-C) was then developed where 3C is combined with ligation

purification and followed by high throughput sequencing to enable unbiased genome-wide identification of 3D chromatin interactions (Lieberman-Aiden, et al., 2009; Soon, et al., 2013).

Briefly, Hi-C involves cross-linking DNA in close physical proximity to fix it then fragmenting the DNA using endonucleases. The edges of the cross-linked DNA fragments are then ligated. The ligation sites are marked with biotin which is used to pull down these fragments that are then taken for sequencing. A recent study has investigated the accuracy of this method in identifying SE-promoter interactions to find genes that are regulated by SEs (Lopes-Novo, et al., 2018).

They characterized about 200 new SE-promoter interactions while taking out about 42 others that were identified using linear proximity. What was concluded from that study is that it is a major possibility for an SE to skip the nearest promoter and to interact with multiple others. In addition to that, they concluded that SEs may be active in different cell stages but interacting with a different set of promoters and thereby regulating a different set of genes.

An important deduction to be noted was that the clustering of enhancers is not necessary nor sufficient to identify a SE. As mentioned previously, we had identified regions which consisted of a singular enhancer with an unexpectedly high H3K27ac signal density, closer in characteristics to SE regions. On the other hand, some stitched enhancer covered larger areas, yet, had a weaker H3K27ac signal density and therefore were not considered SEs.

Conclusion and Future Directions

Since DYSE was developed as a generic algorithm, it could potentially run the analysis on any cell lineage. Moreover, it can be used to compare SEs between different cell stages of the same lineage or to between healthy and diseased cells. However, it would be restricted to datasets that ROSE can run as it relies on the ROSE algorithm for the initial SE identification step. The pros of using DYSE is that it runs the analysis

on multiple TFs for multiple cell stages in one run, hence, hastening the preliminary analysis of SE dynamics between different cell stages. Furthermore, as mentioned earlier, it is designed to run ROSE-calls in parallel for TFs within the same cell stage. This algorithm was developed to simplify the preliminary analysis of identifying changes in SE activity as well as changes in SE-associated genes throughout development.

Nonetheless, it is not free of limitations. Analyzing multiple TF ChIP-seq data adds the risk of skewing of the results due to datasets generated using different ChIP-seq sequencing depths. In addition, as it relies of the ROSE gene mapper, currently, SE-associated genes can only be identified using linear proximity. Another limitation or restriction to DYSE is that it assumes the order of stages is as input by the user. In other words, if the user does not order the input files, the wrong order will not be recognized by the algorithm. As for the feature that determines changes in SE activity between stages, it is based on the assumption that the TF of histone mark the user uses is associated with repressing enhancers, and also it currently only allows for the use of one factor.

As this is the first iteration of the DYSE algorithm, there remains space for improvement of the algorithm to improve its performance and accuracy in identifying cell development related changes in SE activity and changes in SE-promoter interactions. One such improvement would be to integrate Hi-c analysis in the step identifying SE-associated genes. In addition, adding the capability of using a combination of different cofactors and histone marks associated with suppression of enhancers to the feature that compares SE activity between different stages will produce more reliable conclusions as to how SEs act differently in different stages. Another possible field of improvement is to add to the intelligence of the algorithm where it would by itself recognize the correct order of the data based on the name to perform the correct analysis. However, that is currently beyond the scope of this project.

For more accurate results in identifying SEs it is recommended to use ChIP-seq data generated in the same study where the same machinery and sequencing depth was used. Moreover, more reliable results would be produced if the control datasets were also generated using the same protocol and conditions for the other ChIP-seq data.

In the field of SEs, the DYSE algorithm presents a new, easy-to-use computational tool that can be used to study the dynamics of SEs throughout cell development. This study presents a method for automating SE analysis to aid scientists in studying and further analyzing genes and biological processes that are associated with SEs. In my preliminary analysis, when I used the ROSE algorithm to identify SEs using a single TF it took around 20 minutes using 16 CPUs each with 16GB of RAM. Using DYSE I was able to parallelize the step for identifying SEs such that it took around an hour and 45 minutes to identify SEs for 12 TFs covering three myogenic stages. Moreover, running DYSE with the option to not run the ROSE algorithm and using previously identified SEs took between four to five minutes to complete the downstream analysis.

Even though I highlight the role of SEs throughout cell development in our study, I do not disregard the importance of enhancers that fall outside the category of SEs in regulating key cellular processes and pathways. I rather believe that SE comparative analysis helps in narrowing down a subset of genes that is likely to reveal important information about various cell processes and pathways involved in maintaining cell identity and regulating development.

Reference

Achour, M. et al., 2015. Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of Huntington's disease mice. *Human Molecular Genetics,* 3, Volume 24, pp. 3481-3496.

Adam, R. C. et al., 2015. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature,* Volume 521, pp. 366-370.

Anan, K. et al., 2018. LSD1 mediates metabolic reprogramming by glucocorticoids during myogenic differentiation. *Nucleic Acid Research,* 3.

Anon., 2019. *An introduction to Next-Generation Sequencing Technology.* s.l.:s.n.

Asp, P. et al., 2011. Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *PNAS,* Volume 108.

Banerji, J., Rusconi, S. & Schaffner, W., 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell,* 27(2), pp. 299-308.

Bardet, A. F., He, Q., Zeitlinger, J. & Stark, A., 2012. A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols,* , 7(1), pp. 45-61.

Benhaddou, A. et al., 2012. Transcription factor TEAD4 regulates expression of Myogenin and the unfolded protein response genes during C2C12 cell differentiation. *Cell Death & Differentiation,* 19(2), pp. 220-231.

Bentzinger, C. F., Wang, Y. X. & Rudnicki, M. A., 2012. Building Muscle: Molecular Regulation of Myogenesis. *Cold Spring Harbor Perspectives in Biology,* Volume 4.

Berkes, A. A. & Tapscott, S. J., 2005. MyoD and the transcriptional control of myogenesis. *Elsevier,* 16(4-5), pp. 585-595.

Berkes, C. A. et al., 2004. Pbx Marks Genes for Activation by MyoD Indicating a Role for a Homeodomain Protein in Establishing Myogenic Potential. *Molecular Cell,* 14(4), pp. 465-477.

Bhagwat, A. S. et al., 2016. BET Bromodomain Inhibition Releases the Mediator Complex from Select cis-Regulatory Elements. *Cell Reports,* Volume 15, pp. 519-530.

Blum, R. et al., 2012. Genome-wide identification of enhancers in skeletal muscle the role of MyoD1. *Genes & Development,* Volume 26, pp. 2763-2779.

Braun, B., Rudnicki, M. A., Arnold, H.-H. & Jaenisch, R., 1992. Targeted inactivation of the muscle regulatory gene Myf-5 results in abnormal rib development and perinatal death. *Cell,* 71(3), pp. 369-382.

Braun, T. et al., 1989. Differential expression of myogenic determination genes in muscle cells: possible autoactivation by the Myf gene products. *EMBO,* 8(12), pp. 3617-3625.

Buckingham, M. & Relaix, F., 2015. PAX3 and PAX7 as upstream regulators of myogenesis. *Seminars in Cell and Developmental Biology,* 9, Volume 44, pp. 115-125.

Calhabeu, F. et al., 2012. Alveolar rhabdomyosarcoma-associated proteins PAX3/FOXO1A and PAX7/FOXO1A suppress the transcriptional activity of MyoD-target genes in muscle stem cells. *Oncogene,* 6, Volume 32, pp. 651-662.

Carlson, B. M., 2014. Chapter 9 - Integumentary, Skeletal, and Muscular Systems. In: Human Embryology and Developmental Biology (Fifth Edition). s.l.:Elsevier, pp. 156-192.

Charge, S. B. P. & Rudnicki, M. A., 2004. Cellular and Molecular Regulation of Muscle Regeneration. *Physiological Reviews,* Volume 84, pp. 209-238.

Chen, X. et al., 2008. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell,* 6, Volume 133, pp. 1106-1117.

Chipumuro, E. et al., 2014. CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell,* 11, Volume 159, pp. 1126-1139.

Conaway, J. W., 2012. Introduction to Theme "Chromatin, Epigenetics, and Transcription". *Annual Review of Biochemistry,* 81(1), pp. 61-64.

Creyghton, M. P. et al., 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS,* Volume 107, pp. 21931–-21936.

Cui, H. et al., 2017. Muscle-relevant genes marked by stable H3K4me2/3 profiles and enriched MyoD binding during myogenic differentiation. *Plos One,* 6.Volume 12.

Davis, R. L., Weintraub, H. & Lassar, A. B., 1987. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell,* 51(6), pp. 987-1000.

Dawson, M. A. et al., 2014. Recurrent mutations, including NPM1c, activate a BRD4-dependent core transcriptional program in acute myeloid leukemia. *Leukemia,* 12, Volume 28, pp. 311-320.

Dell'Orso, S. et al., 2016. The Histone Variant MacroH2A1.2 Is Necessary for the Activation of Muscle Enhancers and Recruitment of the Transcription Factor Pbx1. *Cell Reports,* 2, Volume 14, pp. 1156-1168.

Dixon, J. R. et al., 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature,* Volume 518.

Doynova, M. D. et al., 2017. Linkages between changes in the 3D organization of the genome and transcription during myotube differentiation in vitro. *Skeletal Muscle*, 7(5).

Ernst, J. et al., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *NAture,* Volume 473, pp. 43-49.

Feng, J. et al., 2012. Identifying ChIP-seq enrichment using MACS. *Nature Protocols,* 9, Volume 7, pp. 1728-1740.

Fong, A. P. et al., 2015. Conversion of MyoD to a Neurogenic Factor - Binding Site Specificity Determines Lineage. *Cell Reports,* 10(12), pp. 1937-1946.

Furey, T. S., 2012. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics,* Volume 13, pp. 840-852.

Gelato, K. A. et al., 2018. Super-enhancers define a proliferative PGC-1α-expressing melanoma subgroup sensitive to BET inhibition. *Oncogene,* Volume 37, pp. 512-521.

Gryder, B. E. et al., 2017. PAX3–FOXO1 Establishes Myogenic Super Enhancers and Confers BET Bromodomain Vulnerability. *Cancer Dicovery,* 4, Volume 7, pp. 884-899.

Hamdan, F. H. & Johnsen, S. A., 2017. Super enhancers – new analyses and perspectives on the low hanging fruit. *Transcription,* 11.

Heintzman, N. D. et al., 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature,* 5, Volume 459, pp. 108-112.

Heintzman, N. D. et al., 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics,* Volume 39.

Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K., 2015. The selection and function of cell type-specific enhancers. *NAture Reviews. Molecular Cell Biology,* 3, Volume 16, pp. 144-154.

Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K., 2015. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology,* Volume 16, pp. 144-154.

Hernández-Hernández, J. M., García-González, E. G., Brun, C. E. & Rudnicki, M. A., 2017. The myogenic regulatory factors, determinants of muscle development, cell identity and regeneration. *Seminars in Cell and Developmental Biology,* 11, Volume 72, pp. 10-18.

Hnisz, D. et al., 2013. Super-Enhancers in the Control of Cell Identity and Disease. *Cell,* Volume 155, pp. 934-947.

Hnisz, D. et al., 2015. Convergence of Developmental and Oncogenic Signalling Pathways at Transcriptionl Super-Enhancers. *Molecular Cell,* Volume 58, pp. 362-379.

Hou, X. et al., 2012. Discovery of MicroRNAs Associated with Myogenesis by Deep Sequencing of Serial Developmental Skeletal Muscles in Pigs. PLoS ONE, 7(12).

Izumi, K., 2016. Disorders of Transcriptional Regulation: An Emerging Category of Multiple Malformation Syndromes. *Molecular Syndromology,* Volume 7, pp. 262-273.

Je Yeong, K., Sumin, O. & Kyung Hyun, Y., 2017. Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. *Molecules and cells,* Volume 40, pp. 169-177.

Joshi, S. et al., 2017. TEAD transcription factors are required for normal primary myoblast differentiation in vitro and muscle regeneration in vivo. *Plos Genetics,* 2.Volume 13.

Kassar-Duchossoy, L. et al., 2005. Pax3/Pax7 mark a novel population of primitive myogenic cells during development. *Genes and Development,* Volume 19, pp. 1426-1431.

Khan, A. & Zhang, X., 2016. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Research,* Volume 44, pp. 164-171.

Kopantseva, E. E. & Belyavsky, A. V., 2016. Key Regulators of Skeletal Myogenesis. *Molecular Biology,* 5, Volume 50, pp. 169-192.

Lee, T. I. & Young, R. A., 2013. Transcriptional Regulation and Its Misregulation in Disease. *Cell,* Volume 152.

Levine, M., 2010. Transcriptional Enhancers in Animal Development and Evolution. *Current Biology,* 9, Volume 20, pp. R754--R763.

Li, B., Carey, M. & Workman, J. L., 2007. The Role of Chromatin during Transcription. *Cell,* 128(4), pp. 707-719.

Lieberman-Aiden, E. et al., 2009. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science,* 326(5950), pp. 289-293.

Lopes-Novo, C. et al., 2018. Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. *Cell Reports,* 22(10), pp. 2615-2627.

Love´n, J. et al., 2013. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell,* Volume 153, pp. 320-334.

Lovén, J. et al., 2013. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell,* 153(2), pp. 320-334.

Lu, L. et al., 2013. Genome-wide survey by ChIP-seq reveals YY1 regulation of lincRNAs in skeletal myogenesis. The EMB O Journal, 32(19), pp. 2575-2588.

Mansour, M. R. et al., 2014. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science,* 12, Volume 346, pp. 1373-1377.

Marinov, G. K., Kundaje, A., Park, P. J. & Wold, B. J., 2014. Large-scale quality analysis of published ChIP-seq data. *G3: Genes, Genomes, Genetics,* 2, Volume 4, pp. 209-223.

Maston, G. A., Evans, S. K. & Green, M. R., 2006. Transcriptional regulatory elements in the Human Genome. *Annual Review of Genomics and Human Genetics,* 7(1).

Moorthy, S. D. et al., 2017. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Research,* Volume 27, pp. 246-258.

Niederriter, A. R., Varshney, A. & Donna M. Martin, S. C. J. P., 2015. Super Enhancers in Cancers, Complex Disease, and Developmental Disorders. *Genes,* Volume 6, pp. 1183-1200.

Nord, A. S. et al., 2013. Rapid and Pervasive Changes in Genome-Wide Enhancer Usage During Mammalian Development. *Cell,* Volume 155, pp. 1521-1531.

Novo, C. L. et al., 2018. Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. *Cell Reports,* , 22(10), pp. 2615-2627.

Park, P. J., 2009. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics,* Volume 10, pp. 669-680.

Peng, X. L. et al., 2017. MyoD- and FoxO3-mediated hotspot interaction orchestrates super-enhancer activity during myogenic differentiation. *Nucleic Acid Research,* 6, Volume 45, pp. 8785-8805.

Pérez-Rico, Y. A. et al., 2016. Comparative analyses of super-enhancers reveal conserved elements in vertebrate genomes. *Genome Research,* Volume 27, pp. 259-268.

Pott, S. & Lieb, J. D., 2015. What are super-enhancers?. *Nature Genetics,* Volume 47, pp. 8-12.

Qian, J. et al., 2014. B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. *Cell,* 12, Volume 159, pp. 1524-1537.

Rada-Iglesias, A. et al., 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature,* 470(7333), pp. 279-283.

Reinert, K., Langmead, B., DavidWeese & Evers, D. J., 2015. Alignment of Next-Generation Sequencing Reads. *The Annual Reviewa of Genomics and Human Genetics,* 5, Volume 16, pp. 133-151.

Relaix, F., 2006. Skeletal muscle progenitor cells: from embryo to adult. *Cellular and Molecular Life Sciences,* 63(11), pp. 1221-1225.

Relaix, F. et al., 2006. Pax3 and Pax7 have distinct and overlapping functions in adult muscle progenitor cells. *The Journal of Cell Biology,* 172(1).

Rudnicki, M. A., Braun, T., Hinuma, S. & Jaenisch, R., 1992. Inactivation of MyoD in mice leads to up-regulation of the myogenic HLH gene Myf-5 and results in apparently normal muscle development. *Cell,* 71(3), pp. 383-390.

Rudnicki, M. A. et al., 1993. MyoD or Myf-5 is required for the formation of skeletal muscle. *Cell,* 75(7), pp. 1351-1359.

Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J., 2012. The long-range interaction landscape of gene promoters.. *Nature,* Volume 48.

Seale, P. et al., 2000. Pax7 Is Required for the Specification of Myogenic Satellite Cells. *Cell,* 102(6), pp. 777-786.

Sebastian, S. et al., 2013. Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation. *Genes and development,* 6, Volume 27, pp. 1247-1259.

Shlyueva, D., Stampfel, G. & Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics,* 15(4), pp. 272-286.

Shlyueva, D., Stampfel, G. & Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics,* Volume 15.

Soleimani, V. D. et al., 2012. Transcriptional Dominance of Pax7 in Adult Myogenesis is Due to High-Affinity Recognition of Homeodomain Motifs. *Developmental Cell,* 6, Volume 22, pp. 1208-1220.

Soleimani, V. D. et al., 2012. Snail regulates MyoD binding-site occupancy to direct enhancer switching and differentiation-specific transcription in myogenesis. *Molecular Cell,* 8, Volume 47, pp. 457-468.

Soon, W. W., Hariharan, M. & Snyde, M. P., 2013. High-throughput sequencing for biology and medicine. *Molecular Systems Biology,* Volume 9, p. 640.

Vahedi, G. et al., 2015. Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature,* 4, Volume 520, pp. 558-562.

Visel, A. et al., 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature,* 2, Volume 457, p. 8.

Whyte, W. et al., 2013. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell,* Volume 153, pp. 307-319.

Yokoyama, S. & Asahara, H., 2011. The myogenic transcriptional network. *Cellular and Molecular Life Sciences,* 68(11), pp. 1843-1849.

Zhu, J. et al., 2013. Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. *Cell,* Volume 152, pp. 642-654.

Appendix

DYSE (Dynamics of Super-Enhancers) – Pseudo code

```
for each stage:
      if useROSE:
            parallel (call ROSE for each TF)
            group all SEs
       else:
            input is SEs → group all SEs
       add all SEs to BED file
for each BED file:
      combine SEs from all other BED files
      sort combined SEs
      identify SEs specific to that stage
if type=2:
      compare all possible pairs of BED files
else:
      compare pairs of BED files in order
generate proportional Venn diagrams
call ROSE_gene_mapper for stage specific and shared SEs
```