©Copyright 2026

Andrew Mullen

Single Cell Methods to Learn Transcription Factor Interactions and Necessary Noncoding DNA During Zebrafish Somitogenesis

Andrew C. Mullen

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Bill Noble, Chair

Cole Trapnell

Sheng Wang

Robert Cornell

Program Authorized to Offer Degree:

Computer Science and Engineering

University of Washington

Abstract

Single Cell Methods to Learn Transcription Factor Interactions and Necessary Noncoding DNA During Zebrafish Somitogenesis

Andrew C. Mullen

Chair of the Supervisory Committee:
Professor Bill Noble
Department of Genome Sciences & Department of Computer Science and Engineering

The precise control of gene expression during embryonic development is orchestrated by complex networks of transcription factors (TFs) and their interactions with noncoding regulatory DNA elements. However, our understanding of how these elements function *in vivo* and how TFs cooperate to drive cell fate decisions remains incomplete and limited in scope. In this thesis, I will present a method sciPlex ATAC-seq, a multiplexed low cost method for single-cell assay for transposable accessible chromatin. Pairing this method with sciPlex RNA-seq, a similar method for transcriptomic measurement, allowed me to build an integrated single-cell atlas of zebrafish embryogenesis and organogenesis from unpaired data to map cis-regulatory elements during zebrafish development. This approach enables single-cell resolution mapping of accessible regulatory DNA in hundreds to thousands of individually indexed embryos, allowing the identification of enhancers and TF binding motifs across hundreds of cell types. I further integrate these data with deep learning models to infer TF-TF cooperativity and test mechanistic predictions using F0 CRISPR injected zebrafish embryos. Following up on novel interactions from our single-cell perturbation experiments, I developed and applied methods to identify TF perturbation responsive noncoding DNA elements and tested their sufficiency and necessity during zebrafish somitogenesis. This framework offers a low-cost and scalable approach to

decode the regulatory logic of development and provides a blueprint for functionally annotating the noncoding genomes of multicellular models.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

ACKNOWLEDGMENTS

This dissertation marks a major milestone in my growth as a molecular biologist. I am deeply grateful to the many people who supported me throughout my PhD journey.

First and foremost, I am grateful to Cole Trapnell and David Kimelman, who mentored me in complementary ways and took a chance on me when I had little laboratory experience. They met my early, sometimes costly mistakes with patience and continued to give me opportunities to learn and improve. I especially admire Cole's ability to read into the tea leaves of emerging technologies and develop rigorous statistical tools that democratize genomics for the broader scientific community. David taught me the practical craft of molecular biology; more importantly, he gave me the confidence to pursue my own research interests and the skills to build the tools I need. Importantly, he modeled the long term discipline and personal sacrifice required to sustain progress in experimental biology.

I am deeply indebted to the University of Washington MD–PhD Program for supporting a major shift in my training from systems neuroscience to molecular biology. That flexibility fundamentally changed my trajectory as a physician-scientist, and I hope to demonstrate over my career that their trust in me was well placed.

I am grateful to my family for valuing education and giving me the curiosity and perspective to find beauty in the world around us. I am also thankful to the members of the Asimov Collective for giving me an outlet to sustain creativity alongside academic training. Finally, I owe my deepest thanks to my wife, Kelly, for her steady encouragement and understanding, especially when experiments or deployed code demanded evenings and weekends. Our daily morning and evening walks were a grounding constant through the volatility of experimental work.

Chapter 1 : Introduction

My thesis work is motivated by two beliefs: (i) functional annotation of the coding and noncoding genome is essential for improving clinical outcomes when using genomic data; (ii) decoding *cis*-regulatory DNA activity is key to identifying disease mechanisms and ultimately therapeutic targets.

Although the cost of sequencing continues to fall, the clinical utility of genomic data remains limited by our ability to interpret it, particularly in the vast noncoding regions that represent over 90 percent of the human genome. The major barrier is no longer reading genomes but understanding the mechanisms of gene regulation and how they can go awry. Overcoming this barrier requires advances in both molecular and computational methods capable of measuring, quantifying, and ultimately functionally annotating genomes in a scalable manner.

Most human diseases arise from disruptions in gene regulation, yet the vast majority (90%) of disease variants map to noncoding regions of the genome (*1*). Unlike coding mutations, which often affect protein structure in predictable ways, noncoding variants influence gene expression across multiple regulatory layers: altering 3D genome enhancer-promoter contacts, chromatin state, transcription factor (TF) binding, and RNA processing and stability. Of this vast range, we focus on TF binding at *cis*-regulatory DNA where variation can be interpreted in terms of TF recognition and binding to noncoding DNA. The sequence space of noncoding DNA is deeply combinatorial. Even a single 200bp accessible patch of DNA, the size of a small ATAC-seq or ChIP-seq peak, has a sequence space (4^200), far larger than the number of atoms in the universe (10^72). TFs further increase this complexity by binding DNA as multiunit complexes with other TFs, sometimes within or between peaks, further expanding the combinatorial complexity of TF-DNA regulatory interactions (*2*). Despite decades of effort, we

lack a predictive quantitative framework that explains how sequence features of noncoding DNA and TF combinations encode transcriptional outputs in different cell types *in vivo*. This gap in understanding greatly limits our ability to interpret genetic variation and to design therapies that act through regulatory pathways.

1.1 Zebrafish as a Model Organism in Genomics Technology Development for the Study of Noncoding DNA

Embryonic development in laboratory model organisms provides a powerful system for studying gene regulation and noncoding DNA. Genetically identical embryos are not only anatomically consistent, but produce common cell fate commitments, gene expression trajectories, and accessible noncoding DNA regulatory programs, enabling comparisons across time and perturbations. Because development unfolds through well-defined stages of tissue development, regulatory events can be mapped at high resolution and interpreted within cell fate commitment. In contrast to human clinical samples, where sampling is limited and confounded by environmental exposures, embryonic development provides an experimentally tractable, temporally resolved, and biologically coherent model for cell fate decisions and gene-regulation.

Zebrafish development is an excellent model to study TF interactions and cis-regulatory elements (CREs) because TFs are required for zebrafish embryonic differentiation and when knocked out, entire lineages of cells fail to form (*3*–*5*). Zebrafish are a well-established model of vertebrate development. In addition to making nearly all of the same organs and tissues as humans, our aquatic friends share genetic and epigenetic properties like DNA cytosine methylation, similar histone modifying proteins, and have genome sizes similar in scale to humans (*6*). Importantly, they are experimentally tractable for rapid genome manipulation, and large colonies can be easily maintained. Lastly, their “goldilocks” cell-count number during

embryogenesis enables well-controlled and well-powered multiplexed experiments with multiple genetic or chemical perturbations in one experiment (*7*). For these reasons, zebrafish embryogenesis is an ideal model system for developing and validating high-throughput functional genomics methods and computational tools.

1.2 High-throughput multiplexed methods for single-cell RNA-sequencing and single-cell ATAC-sequencing

The last decade has been a technological renaissance for molecular biology measurement, especially in single-cells. Measurements are cheaper, easier, higher throughput, and more multiplexed than ever before (*8–12*). Molecular biology journals are constantly publishing new assays, technologies, and the latest protocols. Open science is flourishing and biologists and the medical field are benefiting. Additionally, commercial kits are available for purchase, and library preparation techniques have been democratized. Scientists no longer need to be experts in molecular biology and microfluidics to perform single-cell sequencing experiments.

The expansion of genomics and single-cell methods reflects two intertwined stories of technology. One story is molecular, namely increasingly sophisticated barcoding schemes that allow thousands to millions of reactions to be combined into a single library preparation. The other is engineering, in the form of sequencing platforms that have leveraged advances in semiconductor manufacturing and photolithography for computer processors to make DNA sequencing flow cells exponentially denser, cheaper, faster, and larger in scale (*13*, *14*). As of 2026, in 24-48 hours the NovaSeqX can generate 20 billion paired end reads for a ~2x100bp read length. While silicon manufacturing continues to improve, so will next generation sequencing technologies. One can eagerly imagine what the next decade will bring for magnifying the scale and lowering the cost of sequencing technologies. This exciting future does

not even include mentioning the numerous innovative technologies that give higher quality reads with nanoballs, lower reagent costs with wafers, or emerging long read and ion based sequencing strategies. The future of biological DNA library preparation and sequencing is bright.

In one part of this thesis, I describe my efforts to improve a component of single-cell library preparation methods, alongside broader work on single-cell data representation. These advances were built on decades of work in molecular biology and have been accelerated by the parallel development of high-throughput sequencing chemistry and instruments in industry. At a high level, DNA library preparation transforms a biological property of interest, RNA transcripts, chromatin state, lineage, or perturbation identity, into a DNA sequence that can be amplified and sequenced once pooled (*8*, *11*). These transformations are compatible with adding barcodes to track unique molecules and assign sequencing reads back to their single cell of origin.

Single-cell sequencing has become a popular method of biologic analysis because it resolves organs and tissues into their individual cellular components, revealing cell types, states, and lineage dynamics that were previously obscured by bulk measurements. These higher resolution representations enable mechanistic insights into biology, especially for heterogeneous tissue or rare cell types. Wherever diverse biological processes are happening simultaneously, like during embryonic development, single-cell sequencing is a valuable tool. There are many protocols and techniques for single-cell library preparation; however, in general they fall into two categories, droplet-based and plate-based methods. In droplet-based methods, dissociated cells are physically isolated into oil droplets using microfluidics and mixed with solutions that contain enzymes and uniquely barcoded oligonucleotide beads. Enzymatic reactions are performed to encode the biologic property of interest into DNA as well as attach barcodes. After

the required enzymatic reactions are complete, the droplets are lysed and the user is left with a pooled tube of barcoded DNA. In plate-based methods, specifically single-cell combinatorial indexing (sci-), nuclei are distributed as pools across multi-well plates with enzymes and in addition to molecular transformations of DNA, they are iteratively labeled with barcodes through rounds of splitting and pooling, generating high combinatorial diversity such that a likelihood of collision, i.e., two cells sharing the same barcodes, is statistically low (*15*, *16*). Both techniques have unique advantages and drawbacks, but notably, plate-based sci- methods scale combinatorially while droplet-based techniques are limited to linear scaling. On a per cell measured basis, sci- based library preparations are low cost and require minimal specialized equipment (*11*).

Single-cell RNA-seq (scRNA-seq) measures RNA transcriptomes from individual cells, enabling identification of cell types, quantification of transcriptional abundances under perturbation, inferring developmental trajectories, and constructing gene-regulatory networks (*17*). Sample pooling techniques that 'hash' nuclei enable multiplexing (-Plex) specimens or whole embryos, which makes large experiments more tractable and reduces variation due to batch and sequencing depth (*18*, *19*). The Trapnell lab recently used sci-Plex RNA-seq to profile over one million cells during zebrafish organogenesis at whole embryo scale (*7*). This generated a wild-type reference atlas of several hundred individually indexed zebrafish embryos. The Trapnell lab has also produced sciPlex RNA-seq of chemically and genetically perturbed F0 knockout zebrafish embryos (*7*, *20*). The high replicate count in sciPlex RNA-seq zebrafish embryogenesis experiments enables robust statistics in order to detect lower magnitude effects of perturbations and construct developmental lineage trees (*21*).

scATAC-seq (Assay for Transposase-Accessible Chromatin) allows the user to measure individual cells' chromatin accessibility, enabling the identification of noncoding regulatory elements. The assay uses a hyperactive Tn5 transposase to insert sequencing adapters into open regions of the genome (*22*). These adapters can then be barcoded, PCR amplified, and sequenced (*8*, *23*). This enables discovery of cell type specific *cis*-regulatory elements (CREs) like enhancers, silencers, and insulators, building promoter-enhancer contact maps, as well as inferring DNA motif gene regulatory syntax (compositions and orientations of motifs in CREs that drive behavior). Recently, a multiplexing technique, sci-Plex ATAC-seq, for ATAC-seq was developed in the Trapnell lab (*18*). It was applied to cell culture, but had not been adapted to function for individually indexing zebrafish embryos, like sciPlex RNA-seq had been.

Single cell multiomic assays, like 10X Multiome, sci-CAR (*24*), or SHARE-seq (*25*), can measure multiple genomic layers of a single-cell at once. By splitting the libraries during preparation, one can capture both scRNA and scATAC-seq because cell ID barcodes remain the same and the two measurements can be computationally linked post hoc. One of the greatest challenges with these multiomic assays is finding optimized molecular biology conditions to satisfy multiple classes of molecules at the same time, for instance mRNA transcripts, genomic DNA, and histone bound DNA in a cell prefer different buffers, temperatures, and handling. Specifically in scATAC-seq and scRNA-seq coassays, the nuclei require lysis for tagmentation, elevated temperatures, chelation of ions required for protective proteins, and mechanical handling before reverse transcription in a droplet or plate. These many steps degrade RNA quality and increase the degree of background in the measurements when compared to a single assay alone. There is no such thing as a free lunch, especially in the case of single-cell assays. (*26–29*).

1.3 Computational Methods for Interpretation and Analysis of scATAC-seq Data

1.3.1 Methods of Linking Unpaired scRNA-seq and scATAC-seq Data

Attaching cell-type labels to scATAC-seq is required to effectively interpret the data. However, attaching cell-type labels to scATAC-seq is a challenging task because there are not well-curated catalogs of cell-type specific *cis*-regulatory elements, like there are marker genes for analyzing scRNA-seq data. For this reason, linking single-cell clusters in scRNA-seq and scATAC-seq datasets is a very common and important early task in scATAC-seq analysis.

Optimal strategies for linking unpaired scRNA-seq and scATAC-seq is still an active area of research in computational biology. There are several common and state-of-the-art methods for linking unpaired scRNA-seq and scATAC-seq data. Most techniques rely on an intermediate step of estimating gene transcription through methods called gene-activity scores. These metrics aggregate promoter, gene body, and proximal accessible peaks into a single number per cell to estimate gene transcription in that cell. These methods are poor predictors of transcription because while the absence of promoter accessibility correlates with an absence of transcription, the presence of promoter accessibility does not positively correlate with transcription (*30*). Here, I review a short list of methods for scRNA/scATAC-seq integration, focusing on approaches that i) are widely cited and used, ii) support cell type level mapping, and iii) are computable at atlas scale.

- NNLS - In Domcke et al.(*23*), for each cell, they first create gene-activity scores for each gene from scATAC-seq data. They then filter to several thousand highly variable genes. They do a similar filtering on an unpaired, cell type annotated scRNA-seq dataset. For the scRNA-seq data, they compute non-negative least squares (NNLS) weighted gene loadings in order to map cells to cluster labels. They then apply those calculated gene

loadings to the gene-activity scores from the scATAC-seq data and compute k-nearest neighbors to assign cell types to scATAC-seq data.

- CCA- SeuratV3(*31*) developed a method for unpaired scRNA-seq and scATAC-seq integration that uses canonical correlation analysis (CCA), a statistical method that finds linear combinations of dataset features such that the resulting combinations are maximally correlated. After computing gene-activity scores for scATAC-seq, the highly variable gene-activity scores undergo CCA with highly variable scRNA-seq genes. This maximally correlated component set then provides a latent space for mutual nearest neighbor assignments. For very large datasets, anything over 500k single cells, this final step of mutual nearest neighbors is not computationally tractable without specialized computing hardware.
- GLUE - Cao et. al., developed graph-linked unified embedding, GLUE, a neural network architecture designed to make a unified common embedding space between unpaired scATAC-seq and scRNA-seq datasets (*32*). There are three sub neural networks in GLUE: i) a data variational autoencoder, which functions as a single-cell encoder, ii) a graph autoencoder, which functions as a peak-gene feature encoder, and iii) a discriminator for adversarial training. The data autoencoder takes as input the Latent Semantic Index (LSI) and PCA matrix values from scATAC-seq and scRNA-seq respectively. Instead of relying on conventional gene-activity scores to turn peak level counts to RNA-seq approximations, GLUE defines a graph, linking nearby peaks to genes. This graph serves as the input to the graph autoencoder. The feature encoder latent space acts a prior in the cell decoder. Lastly, the discriminator's goal is to align the scATAC-seq and scRNA-seq cell embeddings through adversarial training.

1.3.2 Methods of Sequence to Function Modeling for Noncoding DNA

A fundamental goal of the study of gene-regulation is to learn the regulatory "grammar" of how noncoding CREs encode rules via DNA to turn genes on and off in the correct temporal, spatial, and signaling environment. Sequence to function modeling is a broad class of computational tools designed to decode that grammar from weakly annotated noncoding DNA sequences. This allows predictions to be made about noncoding variants as well as predict how accessible DNA sequences may predict biologic properties of cells. There are many methods, recently reviewed (*33*), but I highlight three widely cited, field defining examples that mark key conceptual jumps in noncoding DNA regulatory sequence modeling using deep neural networks:

- DeepSea - DeepSea is a convolutional neural network designed as a multi-assay chromatin feature predictor trained on ENCODE consortium data such as TF binding, histone mark, and DNase hypersensitivity data. The model was primarily focused on noncoding variant effect prediction (*34*).
- Basset/scBasset - Basset is also a convolutional neural network; however, it is designed to predict DNA hypersensitivity sites across different cell types, rather than different measured chromatin properties. The accessible regions of chromatin vary by cell type, so by modeling the large distribution of cell-states, Basset can learn specific as well as more general properties of DNA sequence grammar. An updated version of Basset, scBasset, used scATAC-seq data instead of bulk DNase hypersensitivity tracks. In addition to capturing the variability of cell states that can exist in scATAC-seq, the model produces a latent space, informed by DNA sequence, for single-cell representation, imputation, and inference (*35*, *36*).

- Enformer - Enformer is a transformer neural network, the same neural network unit underlying large language models, designed for multi-assay feature prediction. Notably, Enformer incorporates a larger receptive field than DeepSea and Basset, thereby allowing distal noncoding variants to influence model predictions (*37*).

The sequence-to-function model research landscape is rapidly expanding, with many methods that differ in architecture, training data, and specific inference tasks. However, they all share a common goal, modeling noncoding DNA to predict biologic properties of interest (TF binding, histone state, or transcription). These models are most valuable when treated as hypothesis generators for experimental follow-up. By testing predictions in the laboratory, we can identify which model and data assumptions generalize and which do not. Through iterations of modeling and biologic validation, we can learn the sequence features that matter *in vivo* and steadily improve our ability to quantitatively link noncoding DNA sequence to gene regulation.

1.4 Reverse Genetics For the Study of Noncoding DNA

Mechanistic developmental genetics has historically concentrated on coding regions of the genome, in part because early genetic screens preferentially surfaced penetrant morphological phenotypes by disrupting essential and pleiotropic genes. By contrast, genome-wide association studies (GWAS) implicated noncoding DNA variation as the dominant contributor to disease risk, with subtle and context-specific regulatory changes below the level of sensitivity of early genetic screens. This mismatch between tools for probing genetic mechanisms and the biology underlying common human disease has slowed efforts to translate genetic association into mechanism and ultimately treatment. The first FDA-approved gene therapy for sickle cell disease is particularly notable because it works by editing a noncoding

erythroid enhancer of BCL11A in an intron, to induce reactivation of fetal hemoglobin, illustrating how regulatory sequence can be the most clinically actionable lever.

Experimental progress in developmental genetics is largely built on two complementary strategies: forward genetics and reverse genetics. In forward genetic screens, induced random variation is linked to phenotype, and the causal locus is identified through positional cloning. Historically, forward genetics became the bedrock of the field because it was lower cost and produced dramatic phenotypes seen by eye. However, forward genetic screens are structurally biased toward discoveries where a small number of nucleotide changes produces large effects. These conditions are much more easily satisfied by coding loss-of-function mutations.

These features make forward genetics poorly matched for studying the noncoding regulatory genome. Many noncoding elements act in a cell type and stage-specific manner. Enhancer logic can be redundant or buffered such that single nucleotide mutations have small or negligible effects. As a result, the space of regulatory variants that have functional consequences is systematically underexplored by forward screening strategies.

CRISPR has been a panacea for transforming what is feasible in reverse genetics, targeted perturbation followed by phenotyping. It has made reverse genetics scalable not only to all genes but also for noncoding regulatory elements, enabling experimental access to noncoding DNA changes with direct clinical relevance.

Because of the previously mentioned challenges of forward genetics for detecting functional noncoding DNA, it has been rare to find noncoding DNA elements that induce transcriptional or morphological changes in vertebrate development. The most famous example of noncoding DNA affecting vertebrate morphology is the zone of polarizing activity regulatory sequence, also known as the ZRS. It is a ~1kb enhancer that is nearly 1Mb away from the gene it

regulates in an intron of a distal gene. It was identified over a decade of investigation by connecting clinical genetic samples from patients with polydactyly, extra fingers or toes, to a region of DNA quite distal to the gene it regulates, sonic hedgehog, *SHH*. Deeper investigation over the ensuing three decades has shown that swapping the native enhancer in a mouse to that of a snake was able to produce a "serpentized" mouse, where limb buds failed to differentiate *in vivo*. The serpentized ZRS is a powerful example of how evolution can exert its effects through the genome. On a few other occasions, a similar phenomenon has been observed where disruptions to a noncoding element is able to cause a cell type loss or change in morphology of an *in vivo* organism.

In addition to tools for perturbation, higher resolution phenotyping tools like improved cellular and molecular phenotyping through single-cell sequencing assays as well as improvements in whole embryo live-imaging and cell-lineage tracking enables the identification of more subtle changes that are not embryonic lethal but potentially clinically actionable.

Nearly four decades ago, it was shown that expressing *MYOD1* in fibroblast cell lines converted them to myoblasts *in vitro*. That finding has changed our collective understanding of how a single protein can act as a master regulator to determine cell fate. Since that finding, scientists have been searching to identify other transcription factors that define cell fate during embryonic development. The deep scientific literature surrounding myogenesis makes it an attractive target to validate methods for TF-TF interactions. From a clinical perspective, sarcopenia, an age associated loss of muscle mass, substantially reduces quality of life and imposes a growing economic burden as the US population ages, yet there are few effective pharmacological interventions. Understanding the basic biology of myogenesis and muscle stem-cell population renewal is critical to developing therapeutics.

Developmentally upstream of myogenesis is somitogenesis and anterior-posterior (AP) patterning. The rhythmic gene regulatory network that operates in the presomitic mesoderm (PSM) of vertebrates to create the sequential body segments along the A-P axis has long fascinated developmental biologists. Over the last several decades, many gene components have been identified and how their protein-protein interactions occur has been studied. What remains unclear is the role of noncoding DNA in mediating and organizing these protein-protein interactions. Understanding the stem-cells that produce the entire somitic lineage may hold the keys to understanding cell fate renewal and fate commitment in myogenesis.

1.5 About this Dissertation

In this dissertation, I present techniques for generating and analyzing scATAC-seq during zebrafish embryogenesis. Along the way I optimized molecular methods, extended analysis frameworks, and tried to uncover novel *cis*-regualtory elements that drive zebrafish somitogenesis. In Chapter 2, I present sciPlex-ATAC-seq to generate sciATAC-seq libraries of hundreds of thousands of cells from individually indexed zebrafish embryos. I use state-of-the-art computational tools to integrate these scATAC-seq data with scRNA-seq to assign cell type and identify cell type specific *cis*-regulatory elements. In Chapter 3, I extend the neural network architecture scBasset to predict cell type specific transcription factor protein-protein interactions and validate the predictions in CRISPR injected F0 zebrafish embryos. In Chapter 4, I perform mechanistic followup on the transcriptomics and *cis*-regulatory dynamics of the phenotypes identified from the sequence to function neural network validation experiments. In the final chapter, I discuss how these methods can be used to build cell type specific reporters for any cell type and how, when paired with CRISPR perturbation methods,

these techniques represent a step towards building a suite of technologies to functionally annotate the noncoding genome at scale.

Chapter 2: Building an Integrated scATAC and scRNA-seq Atlas of Zebrafish Embryogenesis

2.1 INTRODUCTION

To detect noncoding DNA accessibility and transcriptomic changes during embryonic development and during perturbation, we adapted sci-Plex ATAC-seq and sciPlex RNA-seq, a workflow for multiplexing hundreds to thousands of samples during scATAC-seq and scRNA-seq to capture single-nucleus accessibility profiles and transcriptomes from whole organisms respectively. We optimized whole-embryo dissociations followed by oligonucleotide hashing to label each nucleus with an embryo specific barcode, finding that we can unambiguously recover the embryo of origin for around 75% of cells of scATAC-seq and 70% of cells of scRNA-seq.

Existing single-cell atlases of zebrafish development have primarily focused on either scRNA-seq (Farrell et al., 2018; Wagner et al., 2018; Saunders et al., 2023) or scATAC-seq (McGarvey et. a., 2022, Sun et al., 2024), but not both in a unified framework. As a result, there is no direct alignment between gene expression and chromatin accessibility at the single-cell level, hindering our ability to fully resolve the regulatory logic of cell fate decisions and make predictions about transcription factor biology. A recent multiomic atlas (Liu et al., 2024) captured ~40,000 cells from gastrulation through early somitogenesis; however, due to the linear scaling associated with droplet based methods, the data is limited in the number of cells it profiles. While these datasets resolved diverse cellular states identified during zebrafish embryogenesis, we produced orders of magnitude more cells, across a wider range of stages including later into organogenesis while individually indexing embryos.

Individually indexing embryos allows us to identify heterogeneity that exists in the transcriptomic and cis-regulatory landscape of cellular differentiation and perturbation. Additionally, previous studies primarily focused on early gastrulation and early somitogenesis, as well as being better powered to identify cis-regulatory elements and transcripts that change during genetic perturbation. Because of the combinatorial scaling of our methods, we believed that we could fill the gap in the literature linking chromatin accessibility and transcriptomic data later into zebrafish organogenesis allowing us to make predictions about transcription factor interactions at later stages and in more diverse cell-types.

2.2 RESULTS

2.2.1. Hashing accurately recovers replicate identity in pooled sciATAC-seq single-cell libraries

sciPlex-ATACseq is designed to multiplex many samples together in preparation of sci-ATAC-seq. We first demonstrated its feasibility to individually index zebrafish embryos in a pilot mixed species “fishbowl” experiment, mixing human HEK293 cells and individually indexed zebrafish embryos. Tm optimized hash oligonucleotide DNA sequences with locked nucleic acid, conformationally locked ribose nucleotides, capture sequences allow accurate species identification and allow indexing of wells or embryos with sciATAC-seq readout at single-cell resolution (Fig. 2.1). Unlike in sciPlex RNA-seq, the library preparation of sciPlex ATAC-seq requires hash molecules to form duplexes and stay intact throughout the sci- plate based single-cell library preparation (Fig. 2.1A). In order to reduce barcode swapping, the hash molecules were optimized by exchanging the existing polyA tail for a melting temperature (Tm) optimized sequence with locked nucleic acids (LNA), raising the Tm from 48°C to 86°C. This reduced the likelihood of barcode swapping between cells enabling correct assignment conditions to single-cells. Previous work on sciPlex ATAC-seq had a correct assignment rate of

~10%, these optimizations yielded a hash assignment rate of ~75%, and a nearly 90% species assignment post QC. (Fig. 1C,E). In this proof of concept experiment, we were able to recover 7000 cells (Fig. 2.1D) and a median unique accessible chromatin fragment count of 20120 per cell and a median hash UMI of 3023 per cell (Fig. 2.1B).

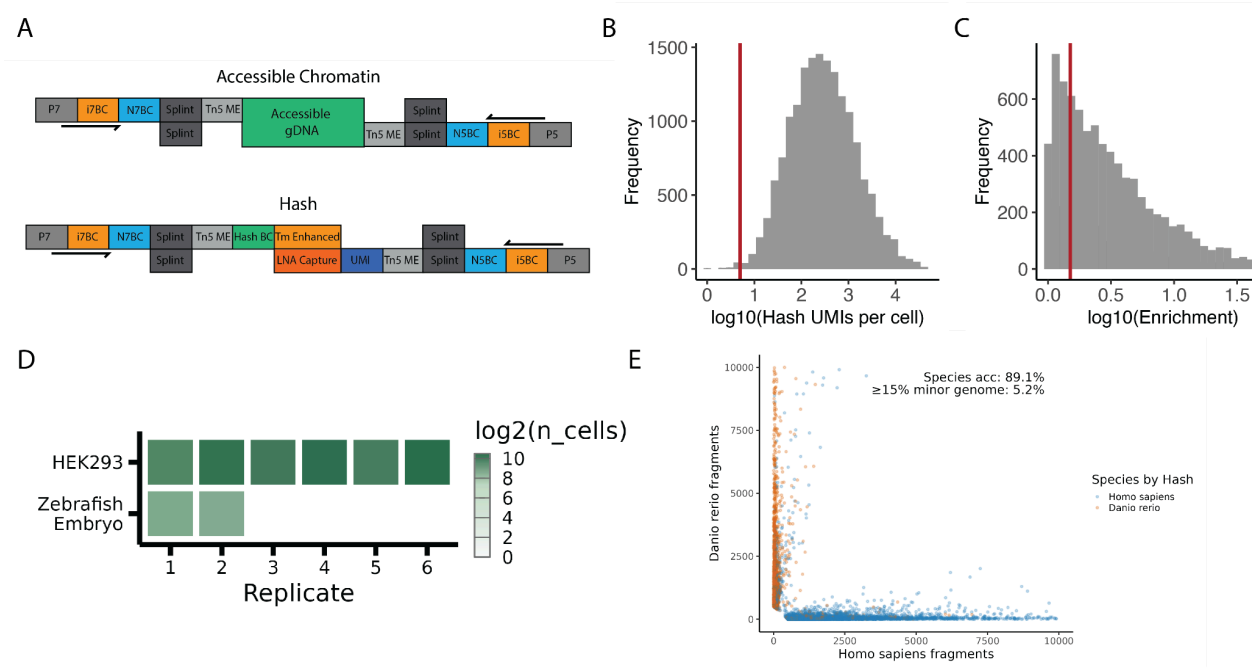


Figure 2.1 - sciPlex ATAC-seq oligonucleotides can be recovered from cell-culture and zebrafish embryos to individually index replicates at single-cell resolution. **A)** Barcode architecture of 3-lvl sciATAC-seq (top) and design of labeling hash barcode oligonucleotides capture sequence to make it compatible with 3-lvl sciATAC-seq (bottom). **B)** Hash UMIs captured per cell, cells with fewer than 5 hash UMIs (red line) were excluded from further analysis. **C)** Enrichment ratios calculated as UMI count ratio of the most abundance vs. the second most abundant hash oligo. An enrichment ratio of 1.5 was used to distinguish high quality assignments **D)** Layout of well assignment and color indicating number of cells recovered per hash label for fishbowl experiment. **E)** Fishbowl plot showing ground truth genome from aligned fragments and the assignment from hashing condition.

Expanding on that pilot study, we used sciPlex ATAC-seq to generate a scATAC-seq atlas of zebrafish embryonic development spanning gastrulation (6hpf) through organogenesis (72hpf) (Fig. 2A-B). Post quality control, we collected between 16 and 37 replicate embryos per

timepoint and amassed 7,213 to 223,013 high-quality single-nucleus chromatin accessibility profiles by timepoint totalling 601,852 cells from 276 embryos across two experiments and 10 timepoints (Fig. 2.2C-D).

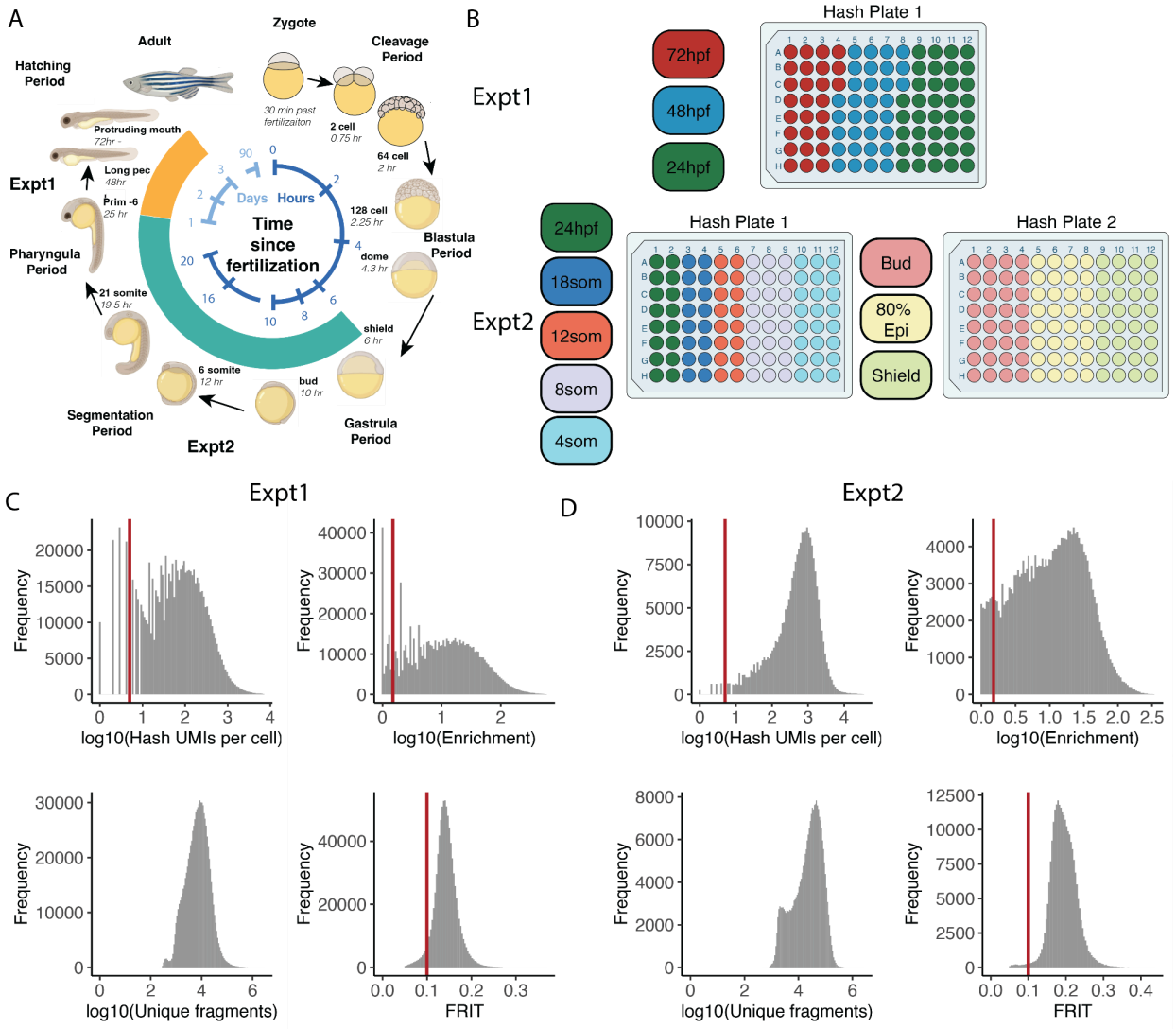


Figure 2.2 - sciPlex ATAC-seq hashing enables multiplexed single-cell profiling across hundreds of zebrafish embryos profiling ten developmental timepoints. A) Schematic of developmental stages sampled. B) Experimental design showing timepoint distribution across hash plates (colors denote distinct conditions, wells denote individually indexed embryos) C) & D) Experiment 1 & Experiment 2 quality-control distributions: Hash UMIs per cell (upper left); enrichment ratios calculated as UMI count ratio of the most abundance vs. the second most abundant hash oligo; unique fragments measured per cell (lower left); and FRIT (fraction of reads in TSS) (lower right).

2.2.2. Embryo-resolved peak calling with hashing increases peak discovery while controlling noise

Hashing of scATAC-seq libraries provides a unique opportunity to use independent replicates during peak calling on scATAC-seq data. Irrepreducible discovery rate (IDR), the recommended protocol defined in the ENCODE project (*38*, *39*), works well with two replicates, but becomes computationally intractable with three or four replicates, let alone hundreds. To overcome this, we define a consensus index score (CIS) for each peak. We separate fragment reads from each embryo via their hash ID and cluster identified from Iterative TF-IDF on 5k bp genomic bins using ArchR (*40–42*). We call MACS3 independently on each of these fragment files (*43*). The consensus index score is the number of unique times a peak is called within a cluster, we keep peaks with a consensus index score above [n]. The consensus index score is saved as metadata and can be used in annotation, regression, or neural network training as a metric of confidence in the peak.

Compared with performing a single MACS3 call per cluster on scATAC-seq, our approach to peak calling in an embryo and cell-type aware fashion increased the number of identified nonoverlapping 500bp peaks from 586,077 to 973,454 peaks using a consensus index score of three, meaning that for a peak to be added to the peak set, we had to observe the peak in three uniquely hashed embryos of a single cell-type cluster (Fig. 3A). This increases the sensitivity of peak identification while reducing noise and false positive peak calls. When we compared our data to bona-fide time-matched bulk zebrafish enhancers validated with GFP expression studies or embryonic H3K27Ac or H3K4Me1 ChIP-Seq data, 95% of the GFP validated enhancers, 98.3% H3K27Ac, and 97.4% H3K4Me1 of ChIP-seq derived enhancers were in the hash-aware peak set (Fig. 2.3B) (*6*, *44*). This comparison demonstrates that peaks

identified with the consensus index from hashing strongly overlap both the non-hashed call set and other large scale scATAC-seq datasets while also capturing a substantial fraction of known enhancer and promoter annotations. With respect to the whole zebrafish genome, the hash-aware peak set identified 29% of the zebrafish genome as containing a *cis*-regulatory element.

Compared with the scATAC-seq atlas of zebrafish development from (Sun et al., 2024) our hash-aware peaks showed high concordance: 78% of Sun et al.'s peaks overlapped ours, and 80% of ours overlapped theirs. We believe the discrepancy, in part, may be due to them profiling 4hpf and 5hpf blastula embryos while we do not. Additionally, because (Sun et al., 2024) relied on dozens of custom loaded Tn5 enzymes as part of their barcoding strategy, there is greater opportunity for heterogeneity due to transposition effect differences as there are known biases in Tn5 insertion when nucleus number varies (*45*). This could be one source of explanation for why our peak set had disagreement from theirs.

2.2.3. Matched sciPlex RNA-seq atlas for scATAC integration

Using sciPlex RNA-seq and building off previous atlas scale generation of single-cell data, we improved the scRNA-seq hashing protocol to use primary amine modified oligonucleotide hashes, NHS-Ester/Ethanol fixation, and trypsin free lysis, increasing nuclear yield, transcriptome quality, while maintaining hashing quality (*7*, *9*). We collected between 8 and 48 replicate embryos and amassed 32,124 to 178,901 high-quality single nucleus transcriptome profiles by timepoint totaling 1,032,110 cells from 307 embryos across two experiments. When combined with the sciPlex ATAC-seq data, after quality control, our dataset included approximately 1.63 million cells from 583 individually barcoded embryos that passed QC.

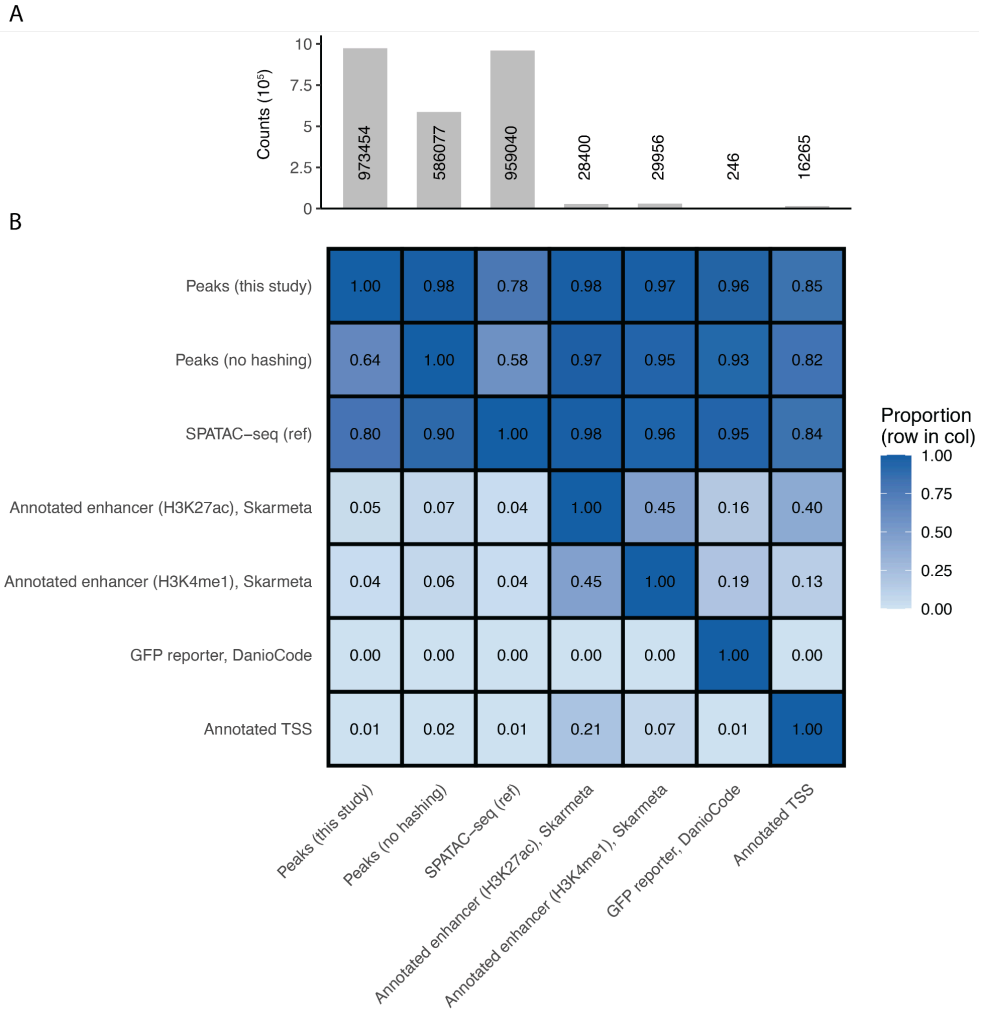


Fig. 2.3 - Peak set composition and literature overlap with bona fide regulatory annotations. A) Total unique peaks observed across the dataset. B) Heatmap showing proportion of our sciPlex ATAC-seq peak calling strategy with hashing and without hashing, an external reference scATAC-seq peak set (Sun et al.), and multiple annotated regulatory element collections (H3K27Ac/H3K4me1 enhancers (Skarmeta), GFP reporter elements (DanioCode), and annotated TSS). Each cell reports the proportions of the row feature set that overlaps the column feature set.

2.2.4. Computational Methods for Integrating Unpaired scRNA-seq and scATAC-seq Data

Recent studies have shown that the deep neural network approach GLUE performs well for co-embedding large (>500k) unpaired scRNA-seq and scATAC-seq datasets (*46*). We

benchmarked its performance on our data versus CCA from Seurat (*31*), and NNLS alignment on cell-type level RNA-seq data (*12*, *23*). Using Adjusted Mutual Information, AMI, a metric that quantifies “biologic structure conservation" across clustering in latent spaces (*47*), we showed that GLUE has an AMI of 0.73, CCA embedding has an AMI of 0.61, and NNLS alignment has an AMI of 0.52 indicating the highest level of biologic structure conservation occurred when using GLUE to generate a multiomic latent space (Fig. 2.4). For context, published benchmarks when using a small amount of PBMC truth co-assayed scATAC-seq and scRNA-seq, providing an empirical ceiling on cross-modality agreement even under highly favorable conditions, AMI topped out at 0.89, indicating an upper bound on information transfer between the scRNA-seq and scATAC-seq layers (*46*). An AMI of 1 would correspond to perfect cluster assignment between the RNA and the Multiomic latent space. This analysis demonstrates that, in the multiomic latent space, GLUE more faithfully recapitulates the biologic clustering structure observed in the scRNA-seq layer than either CCA or NNLS, two of the most commonly used methods for this task.

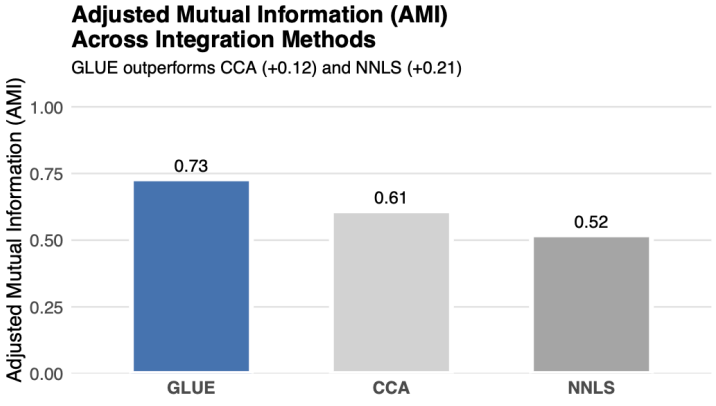


Fig. 2.4 - Performance of unpaired integration strategies for scRNA-seq and scATAC-seq integration. Barchart comparing the accuracy of embeddings learned between native scRNA-seq and learned co-embedding.

2.2.5. Cell-Type Annotation & Hierarchical Label Transfer

To assign cell-type labels to the scATAC-seq cells, we constructed a k-nearest neighbors (kNN) graph between scATAC-seq and scRNA-seq cells using the learned GLUE manifold and transferred the most frequent tissue level annotation among the neighboring scRNA-seq cells to scATAC-seq cells. This allowed us to hierarchically classify cells into six major partitions based on embryonic origin (Fig. 2.6). Within these partitions a second KNN graph was constructed per

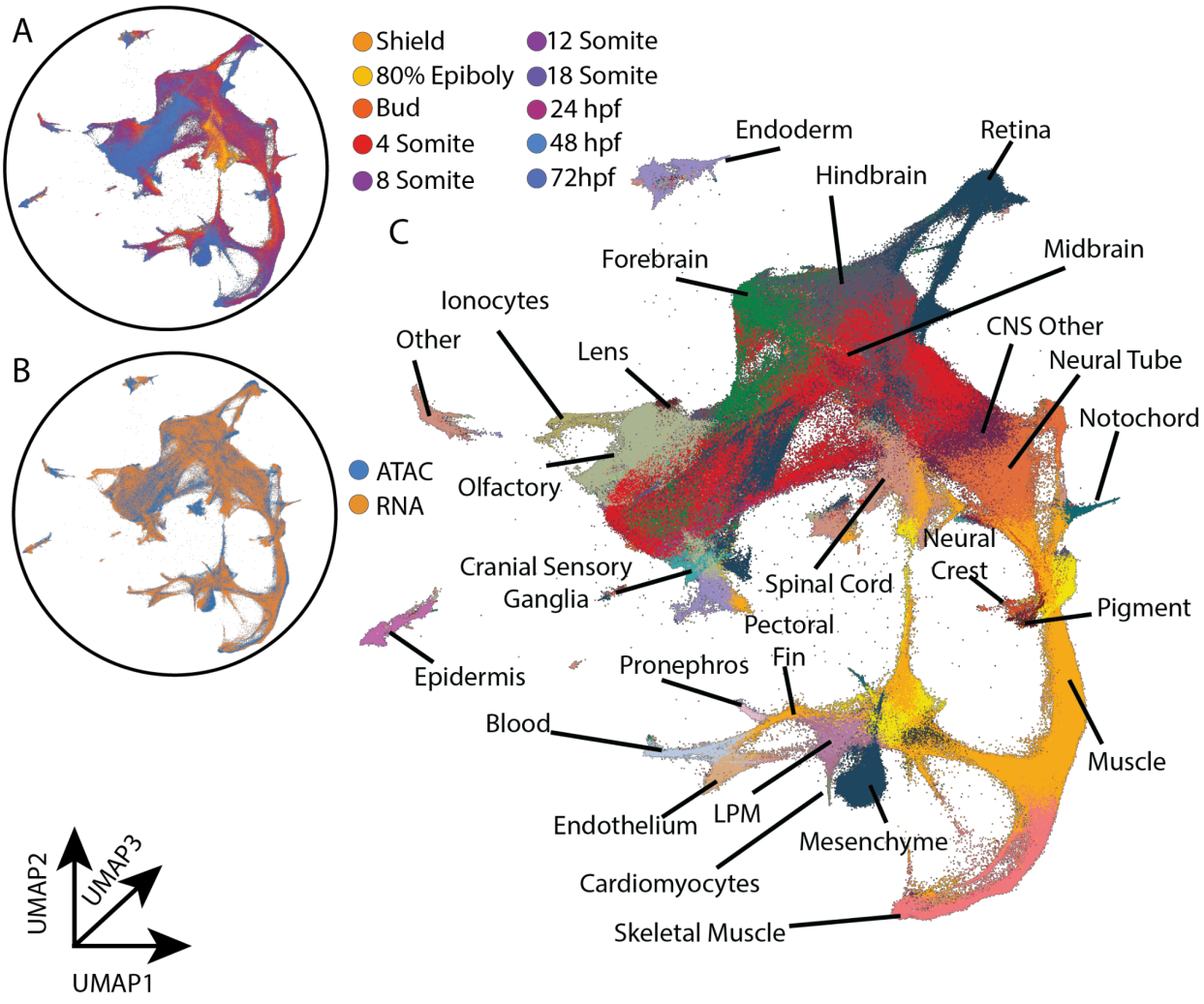


Fig. 2.5 - Integrated 1,638,000 scRNA/scATAC embryonic zebrafish single-cell atlas uniform manifold approximation and projection (UMAP) embedded in three dimensions. A) Inset colored by developmental time from hash annotation. B) Inset colored by assay. C) Colored by major tissue annotations, derived from a kNN graph of scRNA-seq cell-type labels.

partition. Restricting the transfer to locally coherent subspaces increased specificity and reduced misassignment, yielding 219 broad cell-types within the scRNA-seq/scATAC-seq dataset (*7*, *21*). These cell type labels, along with their sampled developmental timepoint (Fig. 2.5A) and individually indexed embryo hash were merged into the global atlas data structure (Fig. 2.5C). Future chapters of this thesis use both the global embedding and the partitioned sub-UMAPs for downstream analysis including top-marker discovery, differential gene expression and accessibility during perturbations, imputation, predictions of TF-TF interactions (*10*, *21*, *48*, *49*).

These sub-UMAP partitions provide a reduced, lineage-focused view that is easier to compute over and manually interpret, allowing us to verify label-transfer assignments via marker gene curation and to refine boundaries between closely related cell types (*21*). Some manual curation of clusters and cell-types resulted in a more cohesive paraxial mesoderm population, largely dominated by the muscle partition. The resulting embedding (Fig. 2.7A) follows scRNA-seq and scATAC-seq cells from gastrula undifferentiated mesodermal tissue through lineage commitment towards terminal differentiation into multiple muscle types and other terminal somitic tissues. To increase the interpretability of the sub-UMAPs and enable lineage restricted DEG comparisons, we constructed a cell-state transition diagram by using the algorithm PLATT and strong priors grounded in the biological literature (Fig. 2.7B). Notably, it is able to differentiate between two presomitic mesoderm (PSM) fates, an anterior somitic fate marked by pax3/pax7 expression population which remains more stem-like supplying much of the myogenic compartment that populates the myotome and a posterior somitic fate which more rapidly forms medial muscle adjacent to the notochord. This anteiror/posterior somitic cell fate lineage will be the focus for much of the analysis in Chapter 4.

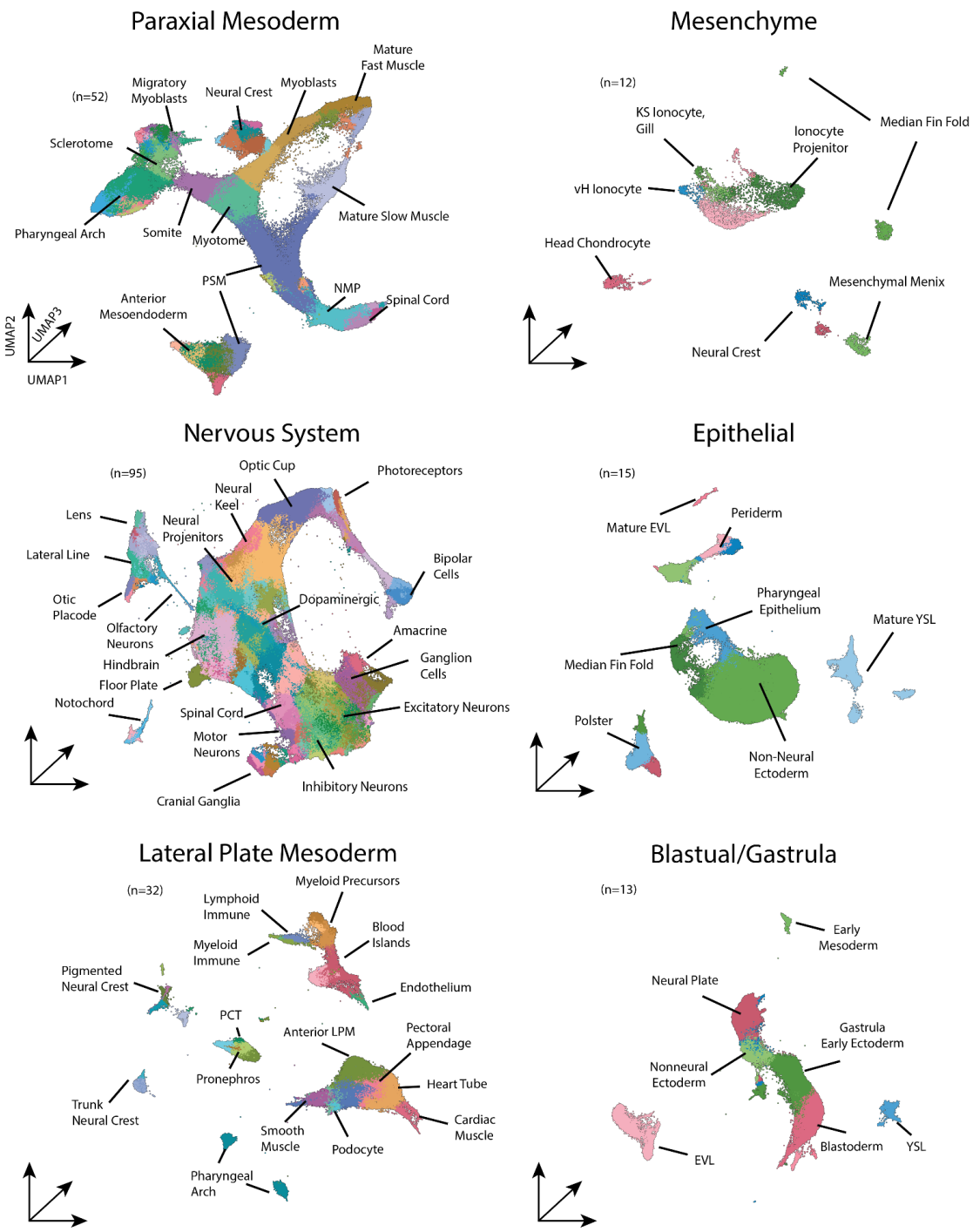


Fig. 2.6 - Sub-UMAP partitions of a global atlas. The global integrated UMAP was divided into six major partitions to enable higher-resolution reclustering and annotation within a more locally accurate embedding. Each partition corresponds to a broad embryonic compartment to facilitate cell-type label transfer. Partitions contain variable numbers of cell-types (upper-left) reflecting differences in lineage diversity.

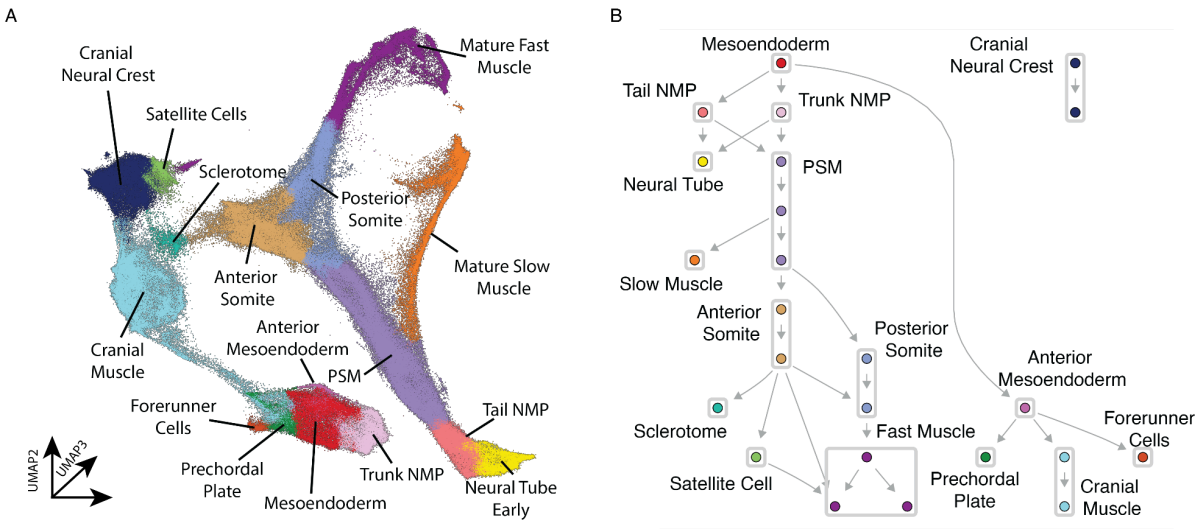


Fig. 2.7 - State Transition Graph - A) Muscle UMAP subspace colored by relevant cell-states for B) PLATT cell lineage state transition graph.

2.3 DISCUSSION

In this chapter, I demonstrate the feasibility of sciPlex ATAC-seq as a scalable strategy for building cell-type resolved chromatin accessibility atlases from individually indexed embryos. In addition to being lower cost and easier than alternative methods of scATAC-seq, it has the additional benefit of using a hash aware peak calling strategy that increases the number of identifiable peaks while providing a principled way to quantify peak quality. By using independent hashed embryos as replicates, *cis*-regulatory element discovery can be expanded without sacrificing quality. Peaks consistent across embryos can be prioritized as high-confidence, while unreplicated peaks provide an empirical handle on noise and batch effects. Together, these features improve the sensitivity and reliability of peak identification in a setting where traditional bulk-style peak calling can be underpowered for rare cell-types.

Building on this method, I constructed a combined reference atlas integrating scRNA-seq and scATAC-seq using state-of-the-art multiomic alignment methods, enabling coherent

annotation and downstream comparisons across assays. This neural network integrated reference supports cell-type annotation at high-resolution and provides a shared embedding for relating chromatin accessibility patterns to transcriptional programs during development.

There are some limitations with the developed methods. These data are not true co-captured scRNA-seq/scATAC-seq multiomic cells, and therefore the atlas should be interpreted primarily as a robust framework for high-resolution cell-type identity transfer rather than a definitive readout of temporal "order of operations" between accessibility and transcription. Although the assays mix well in the integrated space, the resolution and accuracy of alignment along continuous trajectories remain difficult to quantify without ground truth co-captured multiomic data. Learned local biases may persist even when global integration quality appears strong. A further constraint is that the graph prior used to initialize cross-modality structure relies heavily on transcription start site proximity, which can systematically bias *cis*-regulatory elements to gene assignments, especially for regulatory elements in gene-dense loci or complex long-range interactions. This challenge is not unique to the GLUE method used here; it is a general limitation across current methods for integrating scATAC-seq with scRNA-seq because direct physical links are unavailable at single-cell resolution. Possible improvements would be to incorporate a modest amount of co-capture multiomic data as a calibration set, enabling quantitative evaluation of alignment accuracy, particularly along trajectories like the paraxial mesoderm differentiation and providing an empirical benchmark for modality-specific timing claims.

With this atlas in place, the next chapter, Chapter 3, uses the resulting cell-type labeled scATAC-seq data to learn sequence features that encode regulatory logic to predict TF-TF interactions at single-cell resolution. Additionally, we were able to construct a cell-state

transition graph and show that the PSM tissue divides into two fates, one of the anterior and the other of the posterior somite before forming mature fast muscle which we will examine more deeply in Chapter 4.

2.4 METHODS

Animal rearing, staging, and stocks

Staging followed (Kimmel et al., 1995) and fish were maintained at 28.5°C under 14:10 light:dark cycles. Fish stocks used were wild-type AB. Fish were anesthetized prior to imaging or dissociation with MS222 and euthanized by overdose of MS222. All procedures involving live animals followed federal, state and local guidelines for humane treatment and 33 protocols approved by Institutional Animal Care and Use Committees (protocol #4405-02) of the University of Washington.

Preparation of scATAC-seq Barcoded Nuclei

Individual zebrafish embryos were manually dechorionated with forceps and transferred to a 10cm petri dish containing clean zebrafish embryo media. Transfer embryo from petri dish directly onto douncing ball-bearing bead in 30uL embryo media using a wide bore tip. Add 200uL of OMNI ATAC Lysis Buffer to each well. Cover the plate with silicon mat. Place in well plate homogenizer with conditions - 4.5Hz for 7min (24hpf - 72hpf embryos) or 4.5Hz for 5min (early - 24hpf embryos), 3.5Hz for 3 mins (shield - 12 somite). Centrifuge at 100g for 20s at 4C. Centrifuge through 20um filter placed over a deep well catch plate. Add additional OMNI ATAC Lysis Buffer to the beads and then pull up with minimal pipetting and put through 20um filter. Centrifuge filter at 100g for 20sec 4C. Add 7.5 uL of 1uM hash to each well and incubate on ice for 5 minutes. Pipette with wide bore tips. Add 510uL 1.5% (final conc) formaldehyde to wells and mix with widebore tip. Let fix on ice for 15 minutes. Pool samples and spin down pellet 780g 15minutes. Resuspend pellet in 1mL ATAC-RSB + 0.1% Tween and Count. Spin down pellet for 780g 15minutes. Resuspend in Freezing buffer targeting 5e6 cells per 1mL.

Preparation of scRNA-seq Barcoded Nuclei

Individual zebrafish embryos were manually dechorionated with forceps and transferred to a 10cm petri dish containing clean zebrafish embryo media. Transfer embryo from petri dish directly onto douncing ball-bearing bead in 30uL embryo media using a wide bore tip. Add 200uL of Hypotonic Lysis Buffer B. Cover the plate with silicon mat. Place in well plate homogenizer with conditions - 4.5Hz for 7min (24hpf - 72hpf embryos) or 4.5Hz for 5min (early - 24hpf embryos), 3.5Hz for 3 mins (shield - 12 somite). Cell lysis and fixation followed the protocol described in Martin et al., 2023, with an additional 5µl of C12 primary amine-modified hash DNA oligo (10uM, IDT, 5′-/5AmMC12/GTCTCGTGGGCTCGGAGATGTGTATAAGAGA CAG[10bp barcode] BAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -3’) mixed into the Hypotonic Lysis Buffer Solution B. [INSERT FIXATION DETAILS Fixation occurred using the water-soluble NHS ester - BS3 mixed into methanol solution.]

sci-ATAC-seq3 library construction

The fixed and hashed nuclei were processed according to the following protocol: https://www.protocols.io/view/sci-atac-seq3-ewov18xn7gr2/v1 (Dominick et al.)

sci-RNA-seq3 library construction

The fixed and hashed nuclei were processed according to the following protocol https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9839601/pdf/nihms-1846803.pdf (Martin et al., 2023)

Sequencing and read processing

sciPlex ATAC-seq libraries were first sequenced using Nextseq 2000 P3 100 cycle kits and after quality had been assessed, they were more deeply sequenced across two NovaSeqX 10billion 100 cycle kits. Custom primers used R1 -

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG, IDX1 - CTCCGAGCCCACGAGACGACAAGTC, IDX2 - ACACATCTGACGCTGCCGACGACTG ATTAC, R2 - GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG, using the read structure 34-35-35-34. Demultiplexing and basic quality control is linked at this github: demux- https://github.com/bbi-lab/bbi-sciatac-demux & basic quality control- https://github.com/bbi-lab/bbi-sciatac-analyze

sciPlex RNA-seq library

scRNA-seq Analysis were first sequenced using Nextseq 2000 P3 100 cycle kits and after quality had been assessed, they were more deeply sequenced across two NovaSeqX 10billion 100 cycle kits, using the read structure 34-10-10-84 and sequencing kit native primers. Demultiplexing and basic quality control is linked at this github: demux - https://github.com/bbi-lab/bbi-dmux & basic quality control - https://github.com/bbi-lab/bbi-sci/

Peak Calling

Fragment files were loaded into ArchR. Cell by genomic tile matrix was constructed using danRer11 genome reference and 5,000bp windows. Iterative TF-IDF LSI was run with the following clustering parameters: iterations=2, sampleCellsFinal=100000, varFeatures=80000, clusterParams = list(resolution = 0.2, sampleCells = 100000, maxClusters = 10, n.start= 10)). Using the Iterative TF-IDF LSI reduced feature set, clusters were called and pseudobulk fragment files of unique hash ID & clusters were made. To generate a peak set from these psueodbulked fragment files, I ran MACS3 with the parameters: --nomodel --nolambda --keep-dup all --call-summits --shift -75 --extsize 150 –gsize 1,373,454,788, with q-value thresholding (-q 0.1). Peaks were kept if they occurred in three or more replicates within a cell-cluster. Creating the union of peak sets comes from using the ArchR groupScoreQuantile

derived metric from the summit files. The peak with the higher summit score was taken as the true summit.

Multiomic Embedding for Label Transfer

Using the GLUE package, calculate highly variable genes n=4000 in the scRNA-seq using SeruatV3 implementation. Graph prior was initialized using genomic proximity between peaks and genes using the danRer11 genome and gtf. Only highly variable genes and associated peaks were used in GLUE embedding fit for both the data encoder and the graph encoder as well as representing the count data with the negative binomial distribution. The first layer of the data encoder used PCA and LSI representations for scRNA-seq and scATAC-seq respectively. The hidden-dimensionality layer for the encoder and discriminator 512, latent dimension=50, and hidden depth=2, dropout=0.2. Model training included a burn-in. After training, the latent coordinate space was saved and transferred into a monocle3 cds object. KNN were computed over five projection group labels. These projections were then made into subUMAPs for hierarchical cell label transfer.

Cell State Transition Graphs During Somitogenesis

For the muscle subUMAP partition, I constructed a cell-state transition graph using curated annotations of cluster labels as well as the software package PLATT.

2.5 CODE AVAILABILITY

Analysis was performed in R and python. Custom scripts can be found on github: https://github.com/acmullen-med/sciPlexTFx

Analyses of single-cell data were performed using GLUE, Seurat, scBasest, Monocle3, Cicero, Hooke, and PLATT. General tutorials can be found at: https://scglue.readthedocs.io/en/latest/

https://satijalab.org/seurat/

https://github.com/calico/scBasset

https://cole-trapnell-lab.github.io/monocle3/,

https://cole-trapnell-lab.github.io/cicero-release/

https://cole-trapnell-lab.github.io/hooke/

https://cole-trapnell-lab.github.io/platt/

Chapter 3: Prediction and Validation of Transcription Factor Interactions with a Sequence to Function Model Trained on scATAC-seq

3.1 INTRODUCTION

The core challenge in reverse genetics is choosing targets in the genome for perturbation. Surprisingly, it has been difficult to find transcription factors (TFs) that, when perturbed, produce a phenotype (*50*). Although TFs sit at the center of developmental gene regulatory networks, perturbing individual TFs often yields weak or ambiguous outcomes. This is not because TFs are unimportant, but because embryonic regulatory programs are robust by design (*51*, *52*). Many TFs are redundant paralogs arising from genomic duplication during evolution (*53*, *54*). These duplicated genes sometimes have similar expression domains, due to related *cis*-regulatory elements and similar DNA-binding motifs. As a result, loss of a single TF can be buffered by compensatory activity from paralogs or related family members, producing no overt morphological phenotype, modest transcriptional changes, and only subtle shifts in cell-type abundance (*7*). These features complicate the use of single gene perturbation reverse-genetic knockouts as a route to mechanistic insight. A natural question, and motivation for this chapter, is whether stronger and more interpretable phenotypes can be elicited by perturbing multiple TFs simultaneously within the same embryo. By challenging regulatory buffering capacity directly, combinatorial perturbations may expose dependencies and reveal interactions that are invisible to single transcription factor loss (*55*).

Identifying TF pairs to knock out to prevent differentiation is a difficult task (*56*, *57*). The zebrafish genome contains roughly 2,700 TFs, meaning that there are over 7 million possible TF-TF pairs. Even with high-throughput perturbation methods, brute force exploration is not

experimentally tractable (*7*, *26–28*). We need to develop heuristics in order to prioritize TF pairs with higher likelihood of producing phenotypes.

Expanding the typical tasks performed by sequence to function neural networks offers one such heuristic (*36*). These tools are a class of deep neural networks which model underlying DNA sequence, typically noncoding DNA, to predict some genomic property like transcription, accessibility, or histone state (*33*). When trained on ground truth co-assay scRNA-seq/scATAC-seq data, these models have shown that the insertion of a motif into random DNA sequence and prediction of the cell type specific chromatin accessibility correlates with TF transcriptional abundance in that cell type (*36*). We extend this logic to pairs of motifs and consequently pairs of TFs. We compute both the magnitude of effect from two motifs on accessibility as well as the degree of epistasis, i.e., the synergy when the combined effects are greater than simple additivity. We hypothesize that if we perturb TF-TF combinations with high predicted epistasis, we will observe dramatic and interpretable phenotypes when compared to their single TF knockout partners.

In this chapter, I describe how I used cell type annotated scATAC-seq data from Chapter 2 to train a neural network to make predictions about TF-TF interactions. We then microinjected CRISPR-Cas9 RNP into single-cell staged zebrafish embryos, targeting these TFs as individuals and in combinations (*7*). Although we make predictions across the whole embryo, we focus our attention to the paraxial mesoderm on discover that the TF combinations *myod1/six1a/six1b*, *myod1/tcf12*, and *myod1/myf5* produce a more dramatic phenotype of somitically derived fast muscle loss than any of their single gene knockouts alone.

3.2 RESULTS

TFs bind to accessible chromatin, so we reasoned that DNA sequence-based modeling of scATAC-seq could reveal TF motif co-occurrence patterns in *cis*-regulatory noncoding DNA. However, there are a number of statistical challenges that arise from attempting to find motif co-occurrence in noncoding DNA. First and foremost, using a purely frequentist, a genome wide search for motif-motif co-occurence across the hundreds of thousands of motif-motif pairs and hundreds of cell types would be severely underpowered to find cell type specific motif-motif interactions due to multiple hypothesis correction, especially given the sparsity of single-cell chromatin data. Moreover, developmental trajectories form continuous trajectories rather than discrete clusters, making defining the boundaries of cell type specific states challenging. Further complicating conventional statistics, motifs can tolerate degenerate nucleotide sequences and have complex orientations and spacings. These challenges make convolutional neural networks (CNNs) well studied for identifying co-occurring motif combinations, because they use shared parameters across DNA sequences. Some network architectures, such as scBasset, allow for a single model to be trained on the entire dataset, removing some of the challenges imposed by continuous developmental trajectories.

scBasset is a sequence to function neural network that uses one-hot encoded DNA sequence of peaks to predict scATAC-seq data (Fig. 3.1A) (*35*, *36*). There are a number of CNNs designed to predict ATAC-seq from DNA sequence; however, we believed that scBasset was best suited for our whole embryo developmental data generated with sciPlex ATAC-seq because it had three attributes: i) generated a single comprehensive model for all cell types while retaining prediction results at single-cell resolution (in contrast to making a model per cell type), ii) if needed, it had the ability to regress out batch effects by sequencing experiment or gene

perturbation, and iii) had a relatively small number of parameters so that millions of inferences could assess all motif-motif co-occurrence pairs using a few GPUs.

Armed with single-cell resolution motif-motif co-occurrence predictions, we would be able to project the results onto the multiomic GLUE (scRNA-seq/scATAC-seq) latent space. This projection, when paired with a lookup table of TF-motif pairs, would transform our motif-motif *in silico* predictions into experimentally testable TF-TF interactions for a specific cell type.

Restricting our analysis to scATAC-seq peaks with active regulatory histone marks (H3K27ac, H3K4me1, and H3K4me3) would likely improve our ability to identify interacting TFs when compared to using all ATAC-seq accessible peaks. However, the data for cell type specific histone profiles via single-cell Cut&Tag or ChIP-seq data for common histone modifications has not been produced for whole zebrafish embryos. Instead we used the consensus index score (CIS), generated via hashing in Chapter 2, to retain only high-quality reproducible peaks from sciPlex ATAC-seq data.

3.2.1 Sequence to function neural networks can predict transcription factor importance at single-cell resolution

We trained a scBasset model on the cell type annotated sciPlex ATAC-seq data from Chapter 2. We performed hyperparameter tuning to determine bottleneck size and level of consensus index (peak set size) using adjusted mutual information (AMI) between the learned scBasset DNA sequence latent space and the cell type annotated multiomic GLUE latent space shown in Fig. 2.5 as our performance readout. (Supp. Fig. 3.5.1).

To predict the importance of individual transcription factors at single-cell resolution, we used the CIS-BP database of transcription factor DNA binding motifs for zebrafish. We generated a background distribution of 1000 random DNA sequences that were GC nucleotide

balanced to the chromatin accessible peak set. We predicted the accessibility in each cell using these background sequences. We then injected each motif from the CIS-BP database into the center of the DNA sequence and recomputed accessibility for each cell (Fig. 3.1B). Rather than inserting a single consensus motif per PWM (i.e., the maximum-weight nucleotide at each PWM position), we generated sequences by drawing from the probability of the nucleotide in the

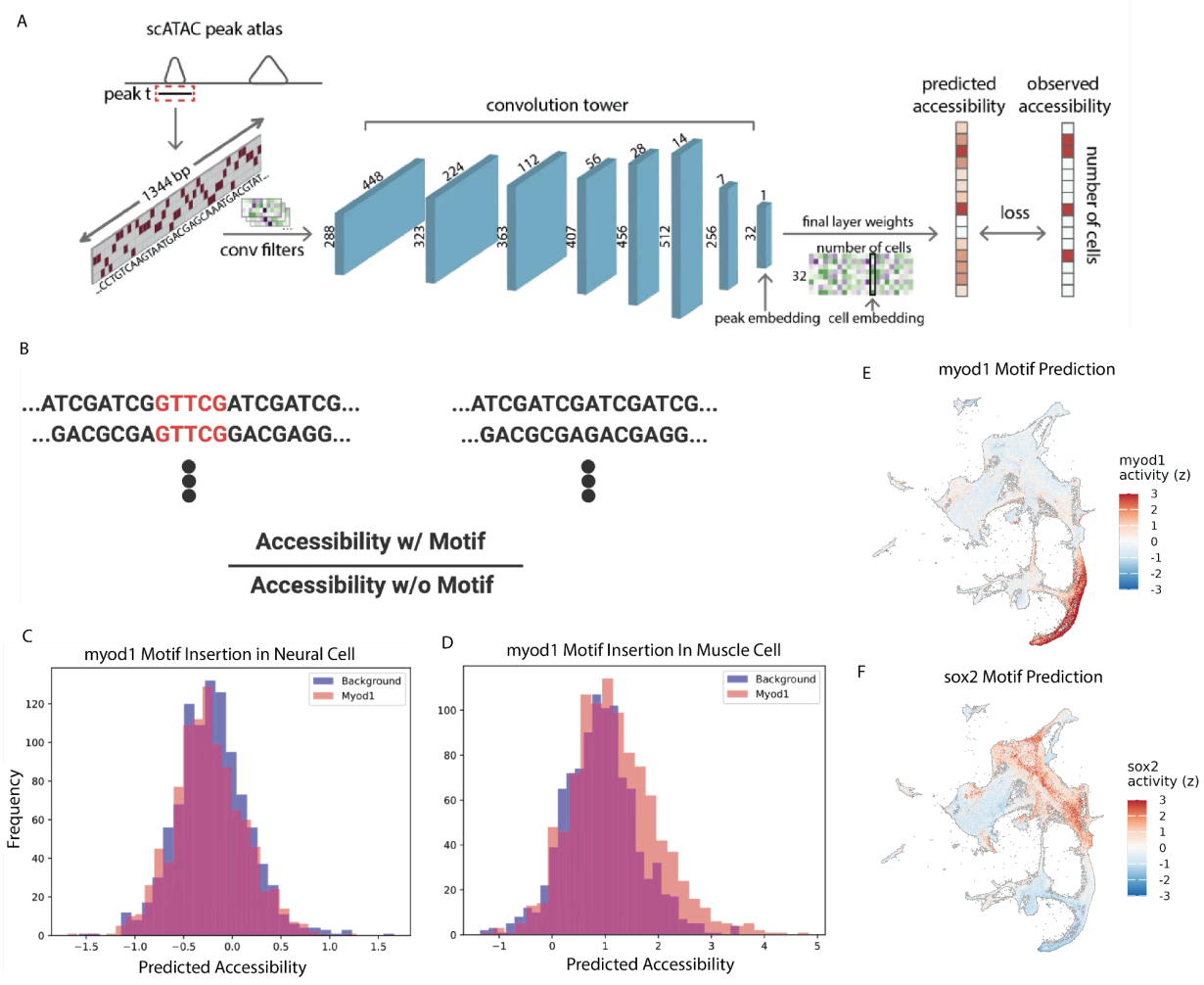


Figure 3.1 Strategy for TF influence prediction at single-cell resolution. A) scBasset neural network architecture. B) Motif injection task performed on random 1344bp nucleotide sequences. C) Histogram of predicted accessibility in a single neuron for 1000 sequences (blue) and 1000 sequences with myod1 motif inserted in the center (red). D) Histogram of predicted accessibility in a single muscle cell for sequences (blue) and sequences with myod1 motif inserted in the center (red). E) Z-score of predictions plotted on global multiomic embedding for *myod1* and F) predictions for *sox2*

PWM. For instance, if a degenerate position had $p(A) = p(T) = 0.5$, the inserted sequences contained A or T with equal probability at the position. We did this for 1000 samples of the PWM and inserted them into the background distribution sequences. We performed forward passes of the model using these DNA sequences generating two distributions for each cell a background and a motif inserted predictions for each cell, representing the influence of the CIS-BP motif on that single-cell (Fig. 3.1C-D). When plotted on the GLUE latent space, the initial motif predictions make intuitive sense: insertion of a *myod1* motif causes an increase in accessibility for muscle cells but not in neurons, as expected (Fig. 3.1E). The motif for *sox2* shows a similar behavior but for neurons (Fig. 3.1F). We performed this analysis for all of the motifs in the CIS-BP database (>500).

To find pairs of motifs that may cooperate, we averaged predictions within each cell type, and we obtained a cell type x motif matrix. For each motif, we took its cell-type prediction vector and correlated it with all other motif vectors to produce a heatmap (Fig. 3.2A) which showed correlations between motif predictions at an embryo wide scale. However, these correlations are confounded by the PWM motif similarity because two PWMs with similar information content should have similar neural network predictions. To extract meaningful biologic correlations from the heatmap that cannot be explained by DNA sequence similarity alone, we performed a linear regression to remove the effects of the PWM motif similarity. A heatmap of the KL divergence, an information theory measure of dissimilarity, between each PWM in the CIS-BP database was computed (Fig. 3.2B). Fitting a linear regression allowed us to remove the correlation of TF influence that can be explained by PWM similarity.

$$cor_i = 1 + \beta\ KL_i + \varepsilon_i$$

Plotting the residuals of this regression (Fig. 3.2C) shows correlations of chromatin accessibility from the motif insertion task that cannot be explained by DNA sequence similarity. Using this technique of finding motif-motif interactions, we showed that 2277 motif pairs had correlations above 0.5 and a residual from the linear model greater than 0.1. Notable motif pairs that meet this criteria are Mef2/Myod1, one of the classic transcription factor interactions involved in myogenesis with extensive biochemical support (*58*). Additional literature supported interactions include Smad2/Smad4 and Pouf5/Sox2 dimers (*59*, *60*). Beyond these canonical examples, we identified hundreds to thousands of predicted motif-motif and cognate TF-TF pairs that are not yet described in the literature, providing a rich set of candidates for further hypothesis generation, experimental perturbation, and validation.

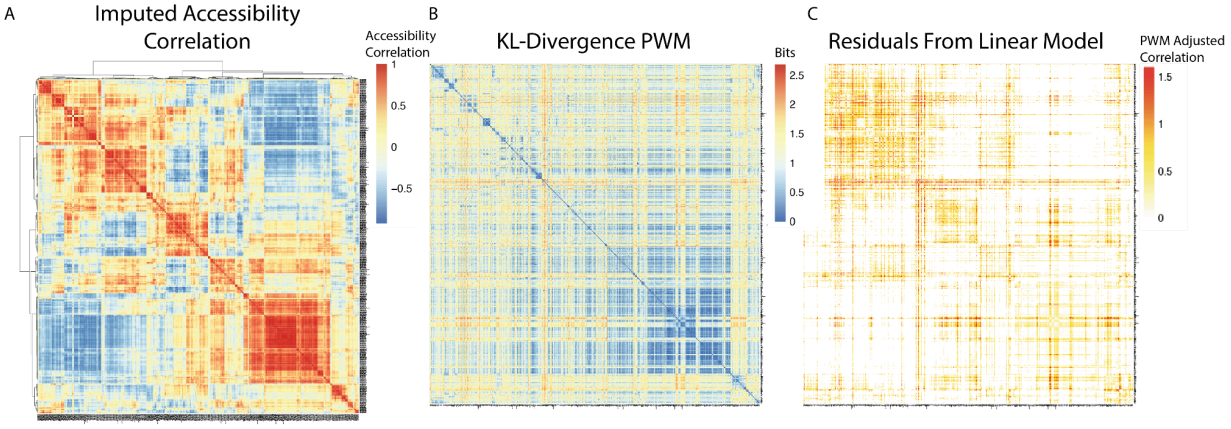


Figure 3.2 Predicted accessibility reveals correlations of cell-type accessibility beyond PWM similarity. A) Pairwise correlation matrix of scBasset predicted accessibility profiles across cell-types for CIS-BP motifs, hierarchically clustered. B) KL-divergence heatmap between CIS-BP binding motifs (displayed in cluster motif order from Panel A) C) Residual after regressing predicted correlation on PWM KL divergence relevating motifs pairs with higher than expected concordance of accessibility prediction (displayed in cluster motif order from Panel A).

3.2.2 Motif-motif synergistic predictions at single-cell resolution suggest cell-type specific interactions.

While the previously described method using state of the art methods generated targets for further validation, substantial evidence from the literature suggests that transcription factor behavior is cell type specific, likely mediated through TF-TF interactions (*61*). The previously described method in *3.2.1* made predictions about interactions that were not at single-cell resolution; rather, they were at embryo resolution. To generate cell type resolved motif-motif predictions, we extended the single motif insertion task to pairs of motifs separated by a fixed 20bp spacer drawn from the background sequence. For each pair, we performed forward passes through the neural network to obtain four predicted accessibility distributions: background alone, motif 1 alone, motif 2 alone, and the motif pair. In addition to quantifying the accessibility shift induced by the motif pair, we calculated synergy as the combined effect divided by the sum of the individual effects, testing whether the joint effect exceeded that expected from the two motifs acting independently.

Across the CIS-BP TF DNA binding motif collection we evaluated approximately 270,000 motif pairs, producing a dense cell by motif pair dataset of predicted motif-motif interactions across all embryonic cell states (Fig. 3.3). When summarized at the cell type level, these predictions showed strong cell type and lineage restriction. 51,331 motif pairs exhibited a positive accessibility change with a z-score greater than 3. These were often sharply lineage restricted accessibility effects, producing coherent blocks of activity across related lineages. Of those 51,331 cell type specific motif pairs, 6,096 exhibited synergy greater than 25%, indicating that TFs using those motifs may cooperate epistatically at *cis*-regulatory elements to drive cell type specific gene expression. In addition to expanding the number of predicted motif-motif interactions, this approach resolved pairs at single-cell and cell type resolution, providing

high-resolution predictions about transcription factor combinations for reverse-genetic perturbation that may induce cell type loss.

3.2.3 Focused examination of synergistic motif and TF interactions in paraxial mesoderm

Restricting our focus from embryo-wide predictions to the paraxial mesoderm lineage enabled a more direct translation from sequence-level motif-motif interactions to protein level TF-TF predictions that were experimentally testable. To organize our analysis, we used the paraxial mesoderm cell-state transition graph generated using the software tool PLATT (Fig. 2.7) (*21*), spanning early-stage gastrula mesoendoderm, presomitic mesoderm, somitic intermediates, and terminally differentiated fast muscle.

We first mapped motif pairs to candidate TF pairs; complicating this, DNA binding motifs often map to TF families rather than individual transcription factors (*62*). To map motif interactions to candidate TF interactions, we assigned each motif to its most likely TF or TF family using the databases CIS-BP, JASPAR, and HOCOMOCO (*63*, *64*). For motifs assigned to broader TF families, we used our multiomic GLUE (scRNA-seq/scATAC-seq) embedding to identify family members present in the paraxial mesoderm lineage, reducing number of the plausible TF-TF candidates. We prioritized candidate interactions by combining three criteria: strong predicted positive synergy, a positive change in motif-motif accessibility, and specificity of cognate TF-TFs to the trajectory segment of interest. Together, these criteria produced a ranked list of TF-TF interaction hypotheses for targeted perturbation experiments designed to test whether paired perturbations produce a stronger phenotype than either perturbation alone.

Figure 3.3 Motif-motif activity predictions at cell-type resolution show structured cell type specific interactions throughout the embryo. Heatmap of double motif accessibility predictions by cell type across the atlas. Columns represent motif-motif pairs and rows represent annotated cell types; both axes are hierarchically clustered via dendrogram to group similar prediction profiles and cell type response patterns. Values are z-scored within each motif-motif pair across all cells in the atlas (positive = higher predicted activity than expected relative to other cells; negative = lower), highlighting clusters of motif pairs with shared cell type specificity and cell type modules with similar predicted motif-motif regulatory interactions.

In the paraxial mesoderm, the motif-motif pairs with the highest ranking using these criteria (upper right of Fig. 3.4B) were dominated by the motif pairs Mef2/Myod1, a canonical example of cooperative transcription factor regulation in myogenesis (*65*, *66*). Unfortunately, the paraxial mesoderm of zebrafish express six unique *mef2* family member transcripts whose proteins have high amino acid sequence similarity. This feature, in addition to their overlapping expression, makes reverse-genetic interrogation challenging due to redundancy. While we have successfully used the CRISPR F0 RNP microinjection system to target up to four genes simultaneously, increasing the number of targets may elevate the risk of mosaicism, off-target effects, and reduced phenotypic penetrance (*7*, *21*). For this reason, we focused our analysis on TF pairs in which both TFs belong to a small paralog family. An additional consideration is that some TFs may be maternally deposited in the egg, making F0 RNP CRISPR Cas9 injection less effective. Prior to experimental perturbation, we used reference tables to confirm that each TF's paralogs were not maternally expressed or were maternally expressed only at minimal levels (*67*, *68*). We therefore focused validation on high-ranking TF pairs that were highly ranked in our predictions and experimentally tractable.

Among the tractable candidate pairs, the Myod1/Six1a motif pair showed increasing synergy and accessibility from presomitic mesoderm through muscle differentiation relative to earlier states (Fig. 3.4A), consistent with established roles of myogenic bHLH factors and Six-family TFs in muscle gene regulation and differentiation programs (*69*). To our knowledge, the literature has not reported testing this pairing using a paired loss of function experiment *in vitro* or *in vivo*. In addition, we selected Myod1/Tcf12 as an additional novel candidate pair. Although Tcf12 is a bHLH transcription factor, the same superfamily as Myod1, bHLH proteins frequently form heterodimers. However, direct evidence of Myod1 and Tcf12 cooperation during

muscle differentiation has not been reported, making this pair an additional novel prediction to test experimentally, albeit lower risk. We also chose Myod1/Myf5 as a positive control interaction with extensive literature surrounding its cooperation (*4*, *70*, *71*), providing a benchmark to calibrate whether our network inference recapitulates well established myogenic regulation.

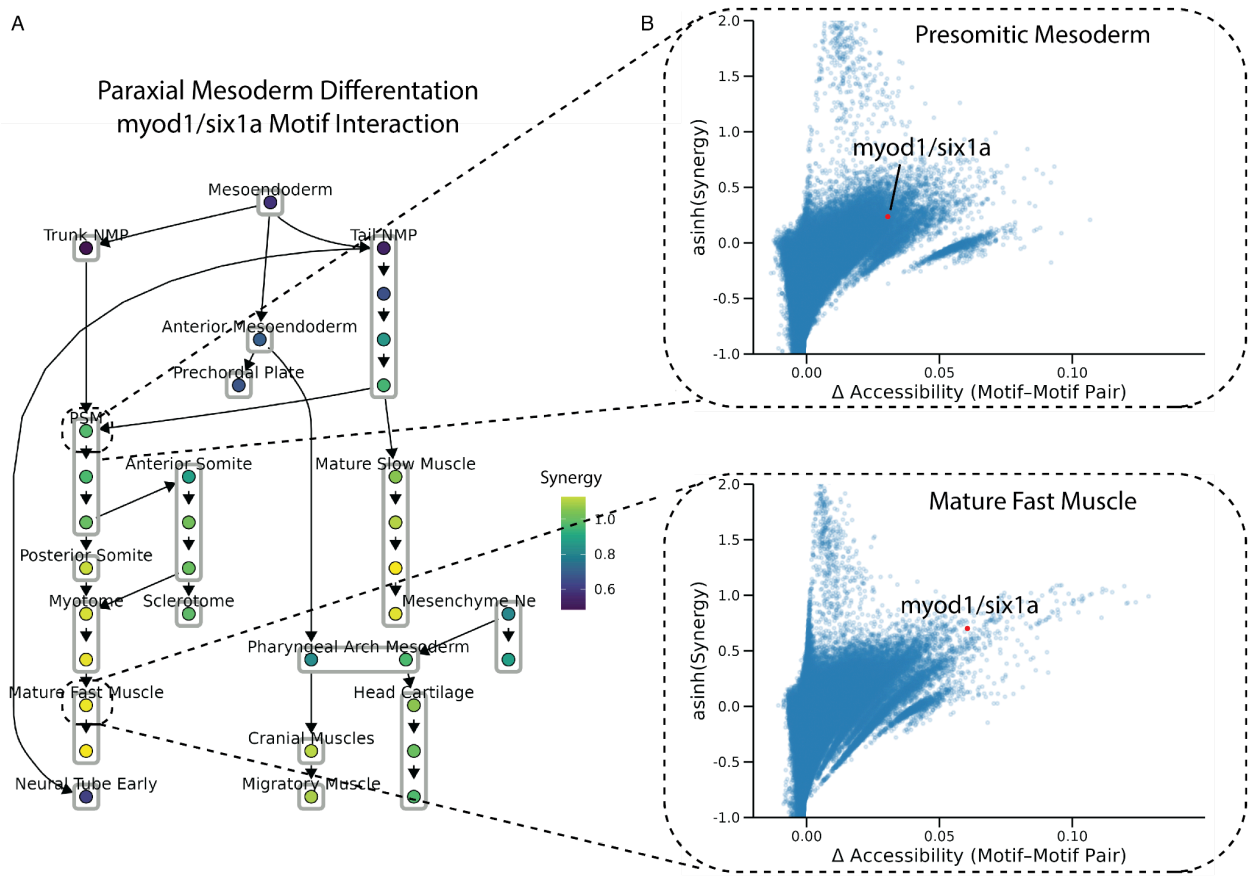


Figure 3.4 Paraxial mesoderm cell-state transition graph shows an increase in synergistic activity in posterior somite cells along differentiation trajectory to terminal fast muscle. A) PLATT lineage graph constructed from the paraxial mesoderm subUMAP from gastrula stage mesoendoderm to differentiated terminal muscle colored by synergy score of *myod1/six1a* predictions. B) Scatterplot of all motif-motif interactions in the presomitic mesoderm (top) and mature fast muscle (bottom) showing the change in accessibility relative to random background for all 270k motf-motif pairs and a sine log transformed (asinh) synergy score.

3.2.4 Characterizing myod1 cooperative knockouts for fast muscle

To evaluate the accuracy of our synergistic predictions and to further investigate transcription factors that cooperate with Myod1 we generated F0 CRISPR-Cas9 knockouts (crispants) in zebrafish for *myod1, six1a/six1b, tcf12,* and *myf5*, as well as double/triple knockout *myod1/six1a/six1b, myod1/tcf12,* and *myod1/myf5*. We designed gRNAs against these genes using CHOPCHOP, a bioinformatic tool well-suited for gRNA design for vertebrate genomes (*72*). We confirmed there were no alternative transcription start states or isoforms of the transcript that would bypass our CRISPR cuts. For each gene, we designed three gRNAs, prioritizing high scoring sites in the first exon or the DNA binding domain annotated on UniProt if it had superior predicted activity. We validated each gRNA individually by injecting preannealed Cas9 RNP into one cell stage zebrafish embryo followed by genomic DNA extraction at 24hpf. Target loci were Sanger sequenced, and editing efficiencies were quantified using TIDE (*7*, *73*). After validation, we injected three gRNAs per gene.

To assess phenotype, we performed immunohistochemistry to evaluate fast muscle development by staining with the antibody F310 (Fig. 3.5) (*74*). Across conditions, single gene crispants showed no appreciable phenotype at 48hpf, whereas double knockouts produced a clearer disruption of somitic fast muscle formation, consistent with cooperative or redundant regulation. For each condition we injected approximately 50 embryos and processed 20 embryos for immunohistochemistry.

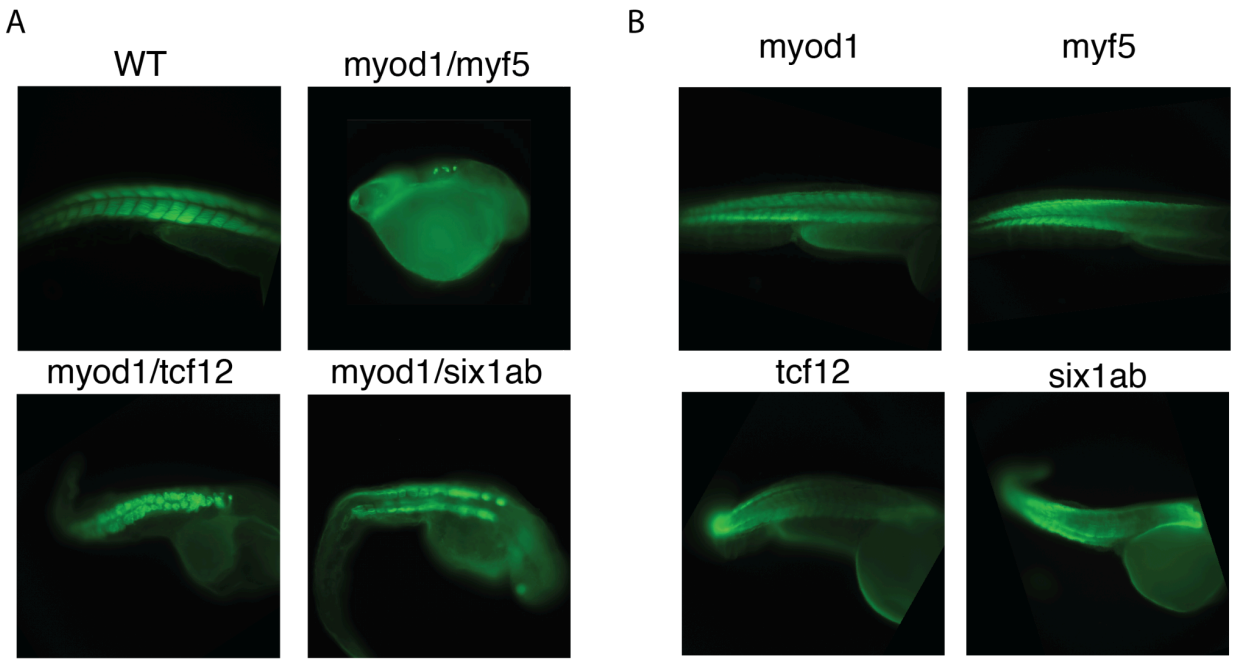


Figure 3.5 - Double knockouts of predicted myod1 patterns cause pronounced fast-muscle defects in 48hpf zebrafish. Representative whole-mount immunofluorescence images of zebrafish embryos stained for a fast-muscle marker F310 (green). (A) Wild-type (WT) embryos show the expected segmental pattern of somitic fast muscle. In contrast, double crispants targeting *myod1* together with *myf5*, *tcf12* or *six1a/six1b* display markedly disrupted fast-muscle formation and abnormal somitic patterning, particularly after the 15th somite. (B) Single-gene crispants (*myod1, myf5, tcf12, six1ab*) show comparatively mild or no overt disruption at the same stage. Images are representative of embryos analyzed per condition (≈50 injected; ~20 processed for immunohistochemistry).

3.3 DISCUSSION

In Chapter 3, we trained a sequence to function neural network and extended an inference task for prioritizing transcription factor combinations for reverse-genetic experiments. We hypothesized that pairs of TF activity were important for establishing cell type specific chromatin accessibility gene regulatory programs during embryogenesis. To develop heuristics to triage candidate TF-TF pairs for experimental evaluation in the multimillion member gene pair

space, we trained scBasset on cell-type annotated sci-Plex ATAC-seq data from Chapter 2. We then performed a large *in silico* motif insertion experiment by running forward passes of the neural network with random DNA sequence and DNA sequence with a TF DNA binding motif, or motif pair, inserted, allowing us to predict the motif's influence on defining regulatory potential in a single-cell or cell-type.

The key observation from the *in silico* analysis of single motifs was that model activity is not just cell-type specific, but also structured across motif pairs in a way that reflects shared regulatory logic. Importantly, these correlations could not be explained by position weight matrix (PWM) similarity alone. While many motifs with similar predicted activity patterns were often close neighbors in PWM space, some were not, suggesting intersections of regulatory activity. This suggests that the sequence to function model was capturing context-dependent regulatory information beyond simple motif identity, potentially integrating features such as motif arrangement, local sequence context, or other higher order DNA sequence patterns that are not well summarized by PWM similarity.

We extended the standard motif insertion paradigm from single motifs to paired motifs, testing all motif-motif combinations to identify potential TF-TF interactions at cell-type specific resolution. This required an extensive amount of compute and forward passes in the neural network model. It was motivated by the belief that TF-TF cooperation likely happens in a cell-type specific context, something that was obscured by the previous analysis and traditional state of the art methods. By measuring whether the predicted effect of inserting two motifs exceeded the sum of their individual effects, we estimated synergy between the two TFs. From this *in silico* screen we identified over 6,000 motif pairs with predicted strong cell-type specificity and high synergy.

We next shifted our focus towards testable biological hypotheses and focused on the paraxial mesoderm lineage, because of its abundance in the zebrafish and easy visualization under a dissecting microscope. Within the lineage, our motif-motif prediction task nominated several motif pairs whose joint activity was enriched in the paraxial mesoderm derived cell-types and whose combined insertions had a large effect size and was synergistic. Notably, we prioritized the combinations *myod1/six1a*, *myod1/tcf12*, and *myod1/myf5* as candidates to validate the neural network predictions *in vivo*. Using CRISPR RNP injection into F0 zebrafish embryos, we saw dramatic morphological differences highlighted with immunohistochemistry using F310, an antibody specific to mature fast muscle.

Despite the promising performance for these three gene pairs, several limitations should temper expectations and provide avenues for improvement for this work. First and foremost, the model is trained to predict accessibility from DNA sequence, not to predict transcription factor or histone modification ChIP-seq data. While single-cell TF specific ChIP-seq data is not yet available, future application for histone modifications could provide useful insight into the repression vs activation at these noncoding loci. Second, the PWM-based motif insertion *in silico* experiment is only as robust as the underlying motifs from the CIS-BP database. There are varying levels of accuracy, some motifs coming from SELEX, others coming from *in vitro* ChIP-seq experiments, and the strongest still coming from FACS sorted *in vivo* bulk ChIP-seq. PWMs change based on the cellular context, and as a result the number of false negative predictions from this technique is likely quite high. Hence, the absence of an interaction does not mean that the two TFs do not interact in a particular cellular context. Finally, the experimental validation with CRISPR Cas9 RNP injection into F0 embryos introduces mosaicism and variable editing efficiencies between embryos and target loci; without well powered and rigorous

quantitative phenotyping, results can be ambiguous and positive effects can be difficult to attribute to specific mechanisms. Locating where the cell-type blockade along the differential trajectory within the paraxial mesoderm occurs remains difficult. We will attempt to dive deeper into the possible mechanism of these TFs interacting in Chapter 4.

3.4 METHODS

scBasset Model Training and Hyperparameter Fitting

Training used binary cross-entropy loss with early stopping if validation auROC improved by less than $1 \times 10{-}6$ in 50 epochs, otherwise 1000 epochs were used. Holdout set was peaks, not cells. Bottleneck size was tuned as was the consensus index score (hash replication rate) for included data. In addition to auROC, because the primary use of the model was multi-task, we compared the adjusted mutual information (AMI) as a metric of biologic conservation between clusters in the learned sequence latent space and the cell-type labels in the GLUE multiomic latent space from Chapter 2 (Supp Fig. 3.6.1). When training a model on the atlas of all of zebrafish embryonic development, these hyperparameter analyses indicated that the 128 neuron bottleneck performed well via AMI. This bottleneck size was four times larger than the native scBasset architecture. The final layer weights in the bottleneck act as cell embeddings. With this larger bottleneck, we found the best performance when using a consensus index score (hash replicate) of 3. We updated model parameters using stochastic gradient descent using the Adam update algorithm. The best performance was achieved with a batch size of 128, intial learning rate of 0.01, $\beta 1$ of 0.95 and $\beta 2$ of 0.9995.

Shuffled Sequence and Motif Insertion

We performed motif insertion on scBasset to compute a motif activity score for each PWM in every cell. Specifically, we generated 1,000 genomic background sequences by performing dinucleotide shuffling of 1,000 randomly sampled peaks from the sciPlex ATAC-seq peak set using fasta ushuffle. For each TF in the CIS-BP Zebrafish2.0 database, we sampled 1000 motif sequences from the PWM and inserted it into the center of the 1344 bp genomic background sequences. We ran forward passes through the model for both the motif-inserted sequences and

background sequences to predict normalized accessibility across all cells. We took the difference in predicted accessibility between the motif-inserted sequences and background sequences as the motif influence for each sequence. We averaged this influence score across all 1,000 sequences for each cell to generate a single-cell prediction of raw motif activity. Finally, we z-score normalized the raw activities to generate the final motif activity prediction. All 250,000 pairwise motif sequences were also generated and used as forward passes in the model. A 20bp nucleotide space was put between each motif pair from the original background sequence.

CRISPR-Cas9 Mutagenesis in Zebrafish Embryos

gRNAs were designed using CHOP-CHOP online tools. gRNA and RNP preparation closely follow a recently published protocol for efficient CRISPR–Cas9 mutagenesis in zebrafish [saunders]. Briefly, gRNAs were synthesized as crispr RNAs (crRNAs, IDT), and a 50 μmol crRNA:trans-activating crispr RNA (tracrRNA) duplex was generated by mixing equal parts of 100 μmol stocks. Cas9 protein (Alt-R S.p. Cas9 nuclease, v.3, IDT) was diluted to a 25 μmol stock solution in 20 nmol HEPES-NaOH (pH 7.5), 350 mmol KCl, 20% glycerol. The RNP complex mixture was prepared fresh for each injection by combining 1μl 25 μmol crRNA:tracrRNA duplex (with equal parts each gRNA per gene target), 1μl of 25 μmol Cas9 Protein and 3 μl nuclease-free water. Before injection, the RNP complex solution was incubated for 5 min at 37°C and then kept at room temperature. Approximately 1–2 nl was injected into the cytoplasm of one-cell-stage embryos. Guide efficiency was validated with genomic DNA extraction, PCR amplification and Sanger Sequencing and evaluation of cutting efficiency per guide with TIDE.

Whole Mount Immunohistochemistry

We used the following antibodies: anti-F310 (mouse monoclonal antibody, DHSB, catalogue no. 16A11, 1:100), Goat anti-Mouse IgG Alexa Fluor 488 (Thermo Fisher, catalogue no. A21236, 1:400). For all immunohistochemistry, embryos were collected at reported stages, embryos were raised in 1-phenyl-2-thiourea (MilliporeSigma, catalogue no. P7629) to suppress pigment formation, anaesthetized with MS222 (10 mg ml−1 in buffered embryo medium; Sigma-Aldrich) and fixed in 4% paraformaldehyde overnight at 4 °C. Antibody staining was performed as previously described. After staining, the embryos were moved into 70% glycerol. Embryos were imaged on a Nikon AZ100 microscope. Images were opened in Napari as .nic files.

3.5 CODE AVAILABILITY:

Analysis was performed in R and python. Custom scripts can be found on github:

https://github.com/acmullen-med/sciPlexTFx

Analyses of single-cell data were performed using GLUE, Seurat, scBasest, Monocle3, Cicero, Hooke, and PLATT. General tutorials can be found at:

https://scglue.readthedocs.io/en/latest/

https://satijalab.org/seurat/

https://github.com/calico/scBasset

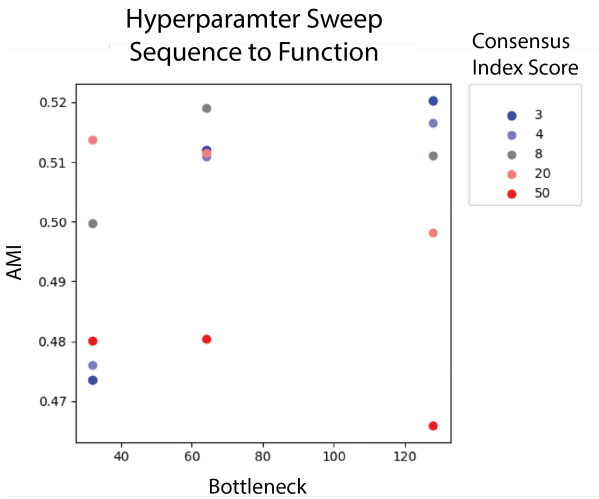
https://cole-trapnell-lab.github.io/monocle3/,

https://cole-trapnell-lab.github.io/cicero-release/

https://cole-trapnell-lab.github.io/hooke/

https://cole-trapnell-lab.github.io/platt/

3.6 SUPPLEMENTARY FIGURES



Supp Figure 3.5.1 - Adjusted mutual information as a function of latent bottleneck layer size and filter on consensus index score (hash abundance per peak).

Chapter 4 - Mechanisms of Fast-Muscle Loss and *cis*-Regulatory Control in Somitogenesis

4.1 INTRODUCTION

In Chapter 3, we used F0 CRISPR perturbations to test predicted transcription factor interactions and observed reproducible developmental phenotypes exclusively in the double knockouts in the paraxial mesoderm lineage. While these experiments established that specific TF-TF perturbations can disrupt paraxial mesoderm differentiation, they do not explain how the phenotype arises. A central goal of this chapter is to move from phenotype description towards mechanism of action, identifying the cell states that are lost or diverted, the gene programs that fail to turn on or become dysregulated, and the *cis*-regulatory elements (CREs) whose accessibility changes during these transcriptional defects.

To address these questions, we used sciPlex RNA-seq to molecularly phenotype crispant embryos of all single and combined gene perturbations (*myod1*, *tcf12*, *six1a/six1b*, *myf5*, *myod1/six1a/six1b*, *myod1/tcf12*, *myod1/myf5*). This approach enabled unbiased transcriptome-wide readouts across hundreds of embryos and dozens of conditions in a single experiment, allowing us to distinguish between selective loss of cell lineages and disruption of specific gene program cascades. In parallel, we used sciPlex ATAC-seq on hundreds of perturbed embryos to connect altered gene expression to changes in the *cis*-regulatory landscape.

This data provided us with the ability to answer four questions. First, which cell types or transitional states are reduced or increased in cellular abundance along the paraxial mesoderm lineage during TF–TF perturbation? Second, within affected cell-types, which genes show the largest and most specific expression changes during double TF knockout but not either single alone? Third, do putative regulatory elements linked to those genes show concordant changes in

accessibility consistent with a causal *cis*-regulatory function? Finally, among candidate CREs implicated by these analyses, which are necessary and/or sufficient to drive cell type specific transcription in vivo? Together, these analyses and experiments aim to connect observed morphological defects observed after genetic perturbations to their underlying molecular basis at single-cell and *cis*-regulatory element resolution yielding a mechanistic explanation for the developmental phenotypes observed in Chapter 3.

4.2 RESULTS

4.2.1 sciPlex RNA-seq analysis of double gene perturbations

Using sciPlex RNA-seq we profiled 344 individual zebrafish embryos to measure 506,538 cells across three timepoints: 12-somite, 24hpf, and 48hpf and nine conditions: F0 injected CRISPR Cas9 against *myod1, myf5, tcf12, six1a/six1b, myod1/myf5, myod1/tcf12, myod1/six1a/six1b*, as well as non-targeted control injections and WT embryos for normalization (Fig. 4.1A). These embryos were distributed across two experiments. We projected the sciPlex RNA-seq data into the multiomic GLUE embedding (Methods) (Fig. 4.1B) and subset the data to the paraxial mesoderm projection.

Using the projected scRNA-seq data, we computed the cell-type abundance shifts along the paraxial mesoderm lineage graph. The three multigene perturbations predicted from our synergy calculations in Chapter 3 all showed increases in the anterior somitic lineage and decreases in the mature fast muscle (Fig. 4.2). Relative to WT, *myod1/myf5* produced a broad redistribution of cell types, including an increase in the anterior somite and loss of terminal fast muscle loss (20x increase and 54x decrease respectively). Relative to the *myod1/myf5* perturbation, *myod1/tcf12* and *myod1/six1ab* showed more modest shifts in composition with with an enrichment of the anterior somitic compartment (12× increase and 5× increase

respectively) and depletion of fast muscle (4× decrease and 25× decrease respectively). These findings are consistent with the immunohistochemistry showing the most dramatic phenotypic

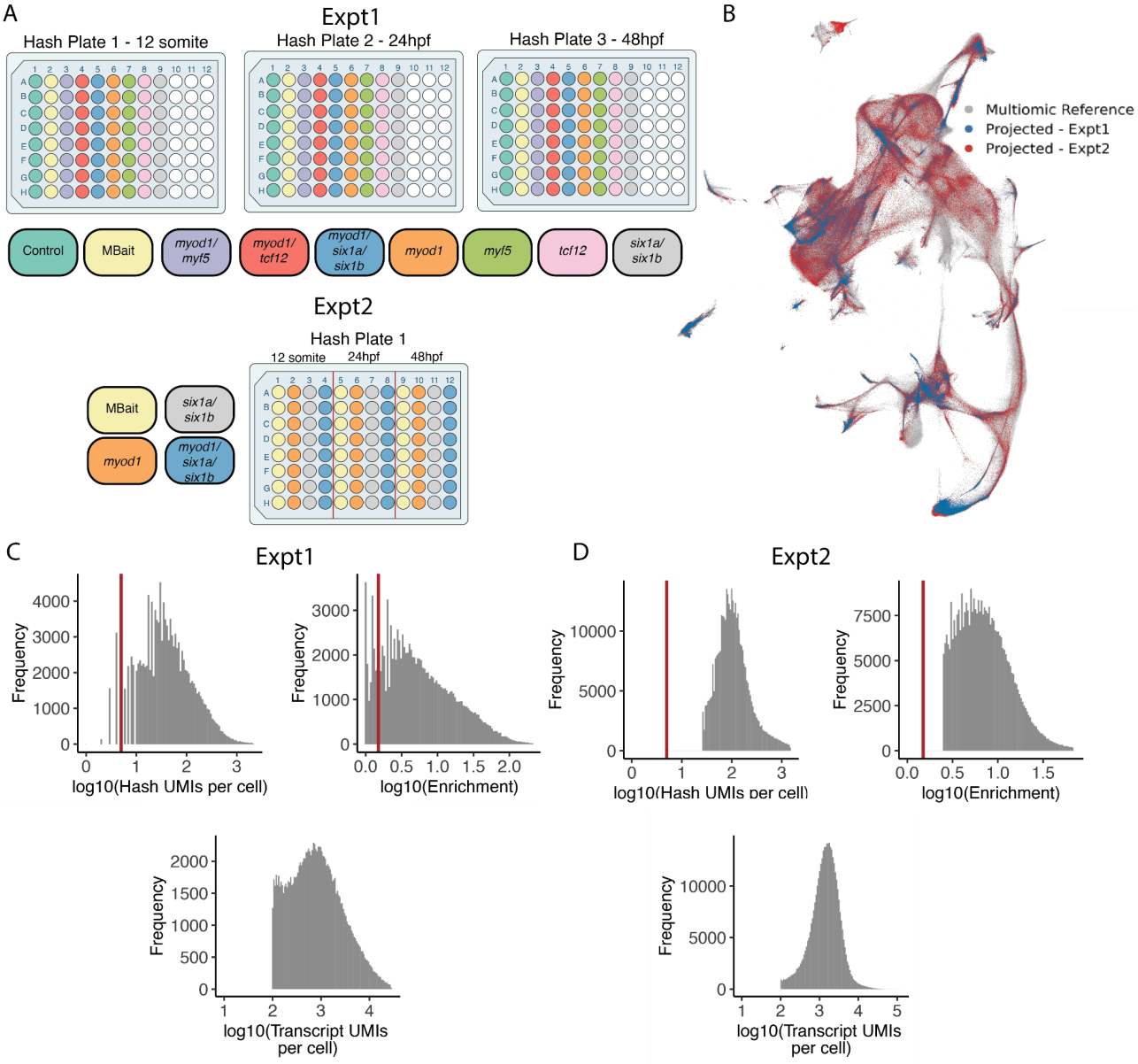
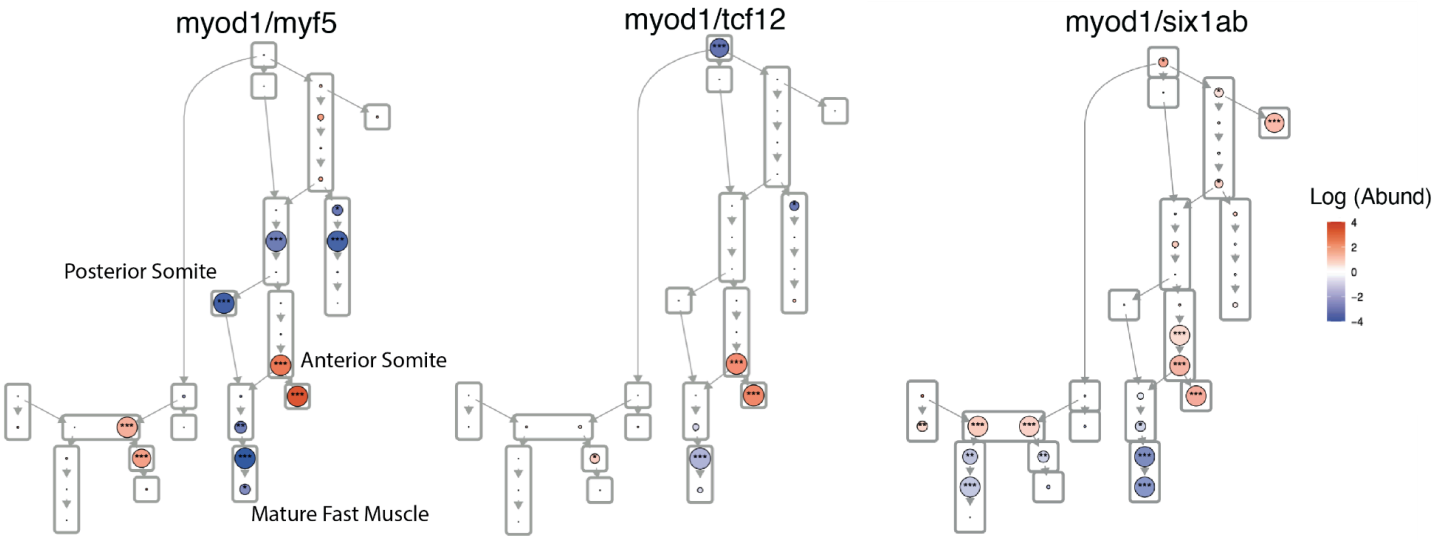


Figure 4.1 sciPlex RNA-seq hashing enables multiplexed profiling of CRISPR Cas9 injected embryos projected onto an annotated global multiomic reference. A) Experimental design showing embryo distribution across hash plates (colors denote distinct perturbations, wells denote individually indexed embryos, hash plate or red line boundary denote timepoint separation) B) Projected sciPlex RNA-seq data into global multiomic GLUE UMAP embedding C) & D) Experiment 1 & Experiment 2 quality-control distributions: Hash UMIs per cell (upper left); enrichment ratios calculated as UMI count ratio of the most abundance vs. the second most abundant hash oligo; unique transcripts measured per cell (lower middle)

muscle loss in the *myod1/myf5* knockout (Fig. 3.5). Additionally, the *myod1/myf5* double perturbation had a loss in the PSM and posterior somite, indicating that disruption to the myogenic cascade may be due to a related but different mechanism of action.

Consistent with the clear phenotype of fast muscle loss in the *myod1/six1ab* double perturbation, differential gene expression showed a large number of genes specific to the double knockout. Analysis in the late presomitic mesoderm, prior to anterior/posterior somite fate commitment and prior to detectable shifts in lineage cell-type abundance, identified hundreds of DEGs with a broad signature distinct from single gene perturbation (Fig. 4.3A). In the late PSM, 848 DEGs were unique to the *myod1/six1a/six1b*, highlighted by the venn diagram, indicating disruption of the late PSM beyond the observations from single knockouts. Because the cell-state



Figures 4.2 Differential abundance of cell-types along paraxial mesoderm lineage reproducible redistribution of cell-states across the lineage-transition graph. Comparing wild-type cell-type abundances in embryos versus three double-gene perturbations show a reproducible redistribution of cell states. Across all three perturbations, lineage proportions shift toward anterior somitic fate and a corresponding depletion of downstream/more differentiated lineages.

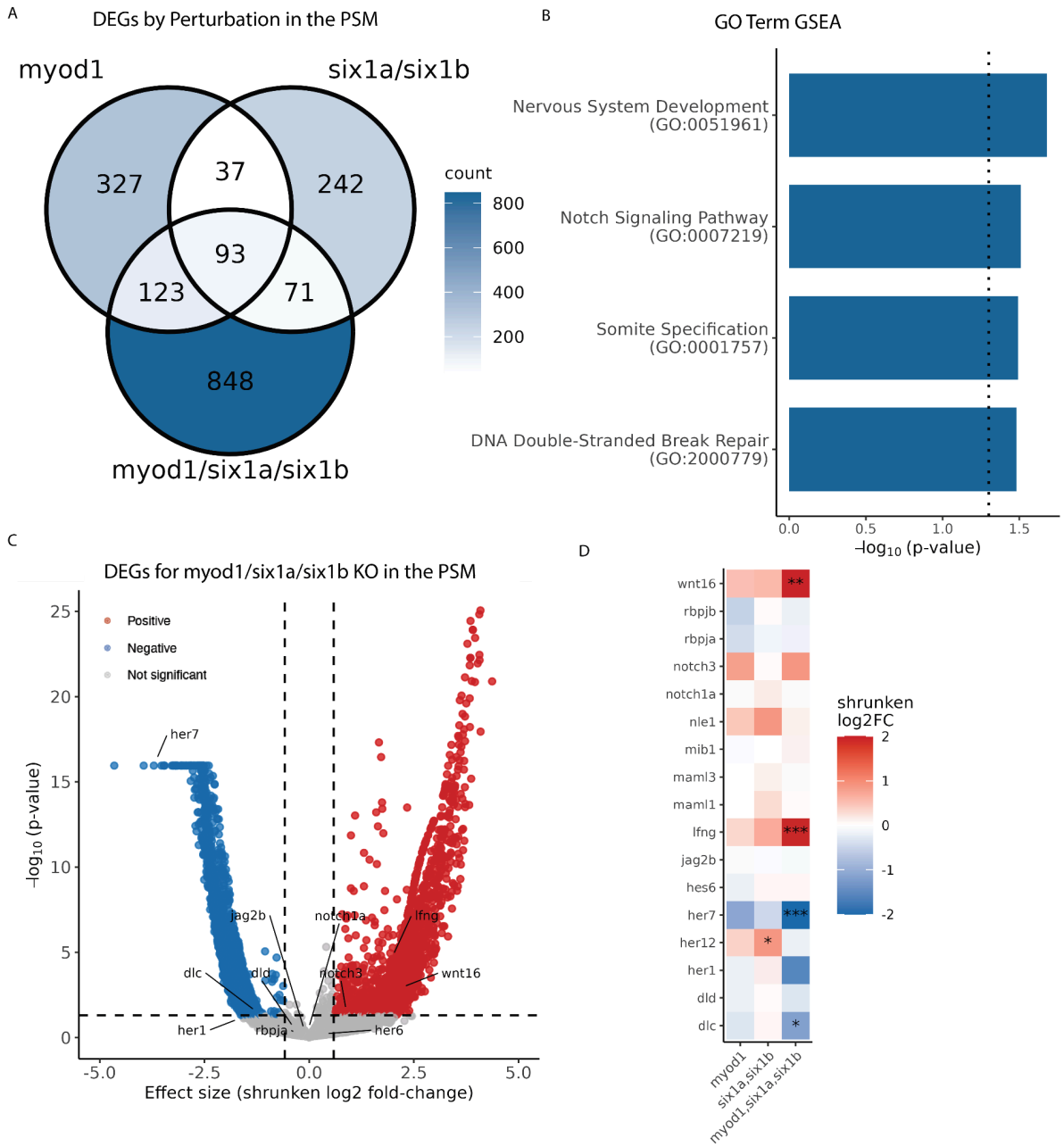


Figure 4.3 Perturbation-Specific Differential Gene Expression in Distal PSM. (A) Venn diagram summarizing differentially expressed genes (DEGs) identified in the late presomitic mesoderm (PSM) across single and double gene perturbation conditions. *myod1/six1ab* exhibits a large perturbation specific signature. **(B)** Gene-set enrichment analysis showing changes in pathways involved in somite specification and Notch signaling using biological process gene ontology (GO) terms. **(C)** Volcano plot of the *myod1/six1ab* specific DEGs, showing effect size (log2 fold change) versus statistical significance (–log10 adjusted p), highlighting the magnitude and direction of transcriptional changes in the late PSM. Text labels are on components of the Delta-Notch signaling pathway during somitogenesis **(D)** Heatmap of gene expression changes during genetic perturbation relative WT in the presomitic mesoderm (*) shows statistical significance using shrunken log2FC (*=0.05, ** =0.01, ***=0.001).

was positioned immediately before the cell type abundance shifts in the cell-lineage, we suspected that the DEGs specific to the double perturbation contained the disrupted mechanism leading to the phenotype and morphologic defects.

We performed gene set enrichment analysis using gene scores derived from the -log10(p-values) comparing *myod1/six1a/six1b* to WT gene expression within PSM cells. Using a *Danio rerio* specific GO:Biological Processes library (*75*), we detected several enriched gene sets, including expected sets like somite specification, negative regulation of nervous system development, and DNA double stranded break repair (Fig. 4.3B). Among these gene sets, Notch signaling emerged as disrupted in the gene knockout condition, prompting closer examination given its established role in regulating intercellular communication during somitogenesis.

The volcano plot shows that the *myod1/six1ab* DEGs have a wide range of effect sizes for differential expressed genes in the PSM. Only some of the genes in the Notch signaling enrichment analysis (GSEA) genes showed statistically significant changes during perturbation (Fig 4.3C). Specifically, the DEG analysis revealed a decrease in the transcripts of *dlc*, *her1*, *her7*, and an increase of the transcript *lfng* (Fig. 4.3C/D).

The Delta-Notch signaling pathway is critical to somitogenesis because it synchronizes activity in the PSM as cells bud off to create new somites. Disruption of this oscillatory gene program could explain the failure of somitogenesis observed in the double knockout zebrafish. Delta proteins are transmembrane proteins that act as ligands to the Notch receptor of adjacent cells. When Delta binds Notch, the intracellular domain of Notch is cleaved at the cell membrane and the Notch intracellular domain (NICD) translocates into the nucleus with transcriptional cofactors like RBPJ to activate transcription of *her1* and *her7* transcription factors, initiating the somitogenesis cascade as well as the negative feedback loops required to generate clean

boundaries of the nascent somite. Lunatic fringe (Lnfg), a glycotransferase protein, modifies the Notch receptor, making it more or less permissive to Delta ligand binding depending on cellular context (*76–80*). The transcription factors *her7* and *her1*, core components to the somite oscillator have their levels downregulated. This pattern of gene expression changes suggests that the *dlc* ligand of the Delta-Notch signaling intercellular signaling pathway is downregulated while *jag2b* stays intact during paraxial mesoderm specification during gene knockout of *myod1/six1a/six1b*. The results suggest that the cells in the PSM may be trying to compensate for gene-regulatory changes to Delta-Notch by increasing *lnfg* and noncannonical Notch signaling through an increase of *wnt16* expression. When we plot the patterns of differential gene expression on the state-transition graph using PLATT it shows that *lnfg* increases its expression

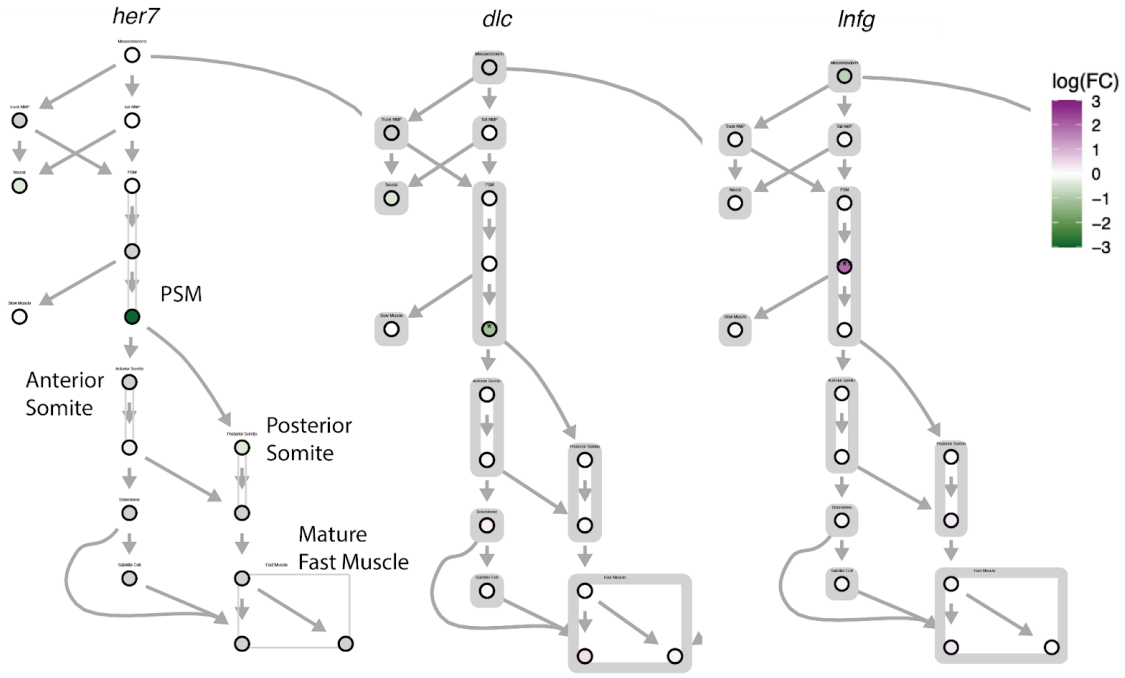


Figure 4.4 DEGs on Paraxial Mesoderm Lineage Graph comparing *myod1/six1ab* vs WT embryos. Colored by log fold change for *her7, dlc,* and *lnfg*, all members of the Delta-Notch signaling cascade.

earlier in the cascade than a drop in *her7* or *dlc* expression, suggesting that its alteration precedes the changes, consistent with observations in chick and mouse (Fig 4.4) (*76*, *80*).

4.2.2 Imaging validation via in situ hybridization analysis of double gene perturbations

To validate these DEGs, we performed whole mount *in situ* hybridization of 4-somite and 12-somite staged embryos (Fig 4.5). The observed pattern for *her7* and *her1* suggests that coherence between cells has been disturbed, such that the traveling wavefront of segmentation fails. The *dlc* transcript, which produces DeltaC protein, is completely ablated.

Together, the transcriptomic and whole-mount *in situ* hybridization data show that these CRISPR perturbations disrupt the oscillatory expression of *her1* and *her7* in the presomitic mesoderm (PSM), indicating a loss of intercellular synchronization within the segmentation clock. This breakdown in coherent wave propagation, along with loss of *dlc* expression, supports a model in which perturbation of the oscillatory network directly impairs clock function thereby causing defective somitogenesis.

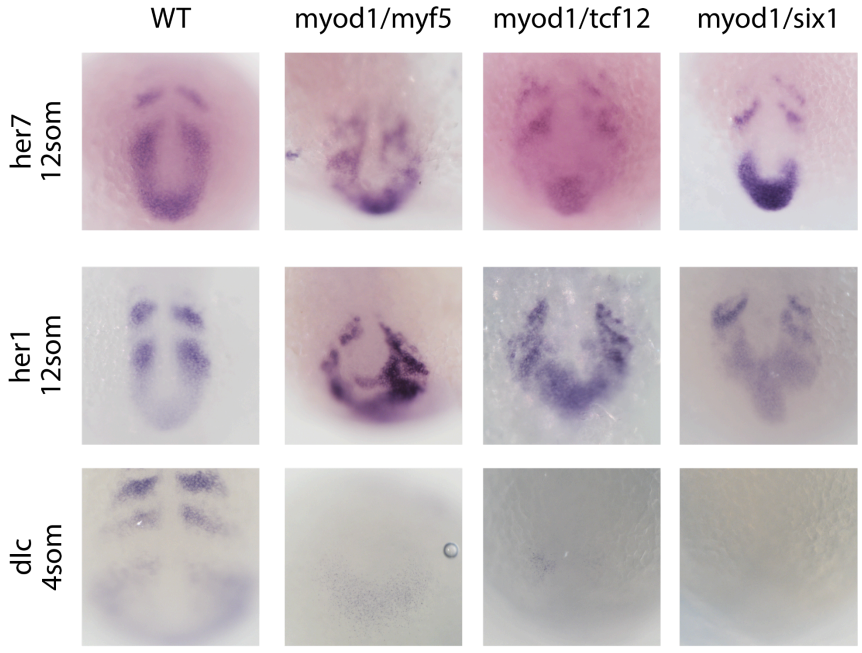


Figure 4.5 *her1/her7/dlc* whole mount *in situ* hybridizations at 4 and 12 somite stages.

4.2.3 sciPlex ATAC-seq analysis of double gene perturbations

To move beyond transcript level changes and identify the *cis*-regulatory elements directly embedded within this oscillatory circuit, we next profiled chromatin accessibility in whole CRISPR-perturbed embryos using sciPlex ATAC-seq. This approach allowed us to measure an additional 272,848 individual cells across 10 perturbations and 5 timepoints providing a high-resolution view of how disruption of key segmentation clock components reshape the regulatory landscape (Fig 4.6). We projected these data into our reference multiomic GLUE atlas to attach cell type labels and interpret the results within our previously established reference frames.

Using the software tool Cicero, we linked enhancer and promoters of differential expressed genes from our CRISPR perturbed sciPlex RNA-seq experiments (Fig 4.7a). Doing this for the genes *her7* and *her1* it became quickly apparent that they share a 14kb long 5' regulatory element. One such enhancer-promoter contact in this region was a 500bp fragment -10kb upstream of *her7* and -2.8kb upstream of *her1*. We performed an energy based motif scan and identified an enrichment of Myod1 and Six1a binding motifs at this locus (*81*) (Fig 4.7b). Additionally, by plotting the fragment reads attributable to each perturbation condition in the PSM, we identified that the noncoding regulatory element -10kb upstream of *her7* closed in response to the combined knockout of *myod1/six1a/six1b* but neither *myod1* or *six1a/six1b* alone (Fig 4.7b). In this way, sciPlex ATAC-seq helped us identify a possible regulatory element involved in mediating the changes in *her1* and *her7* and this genetic circuit.

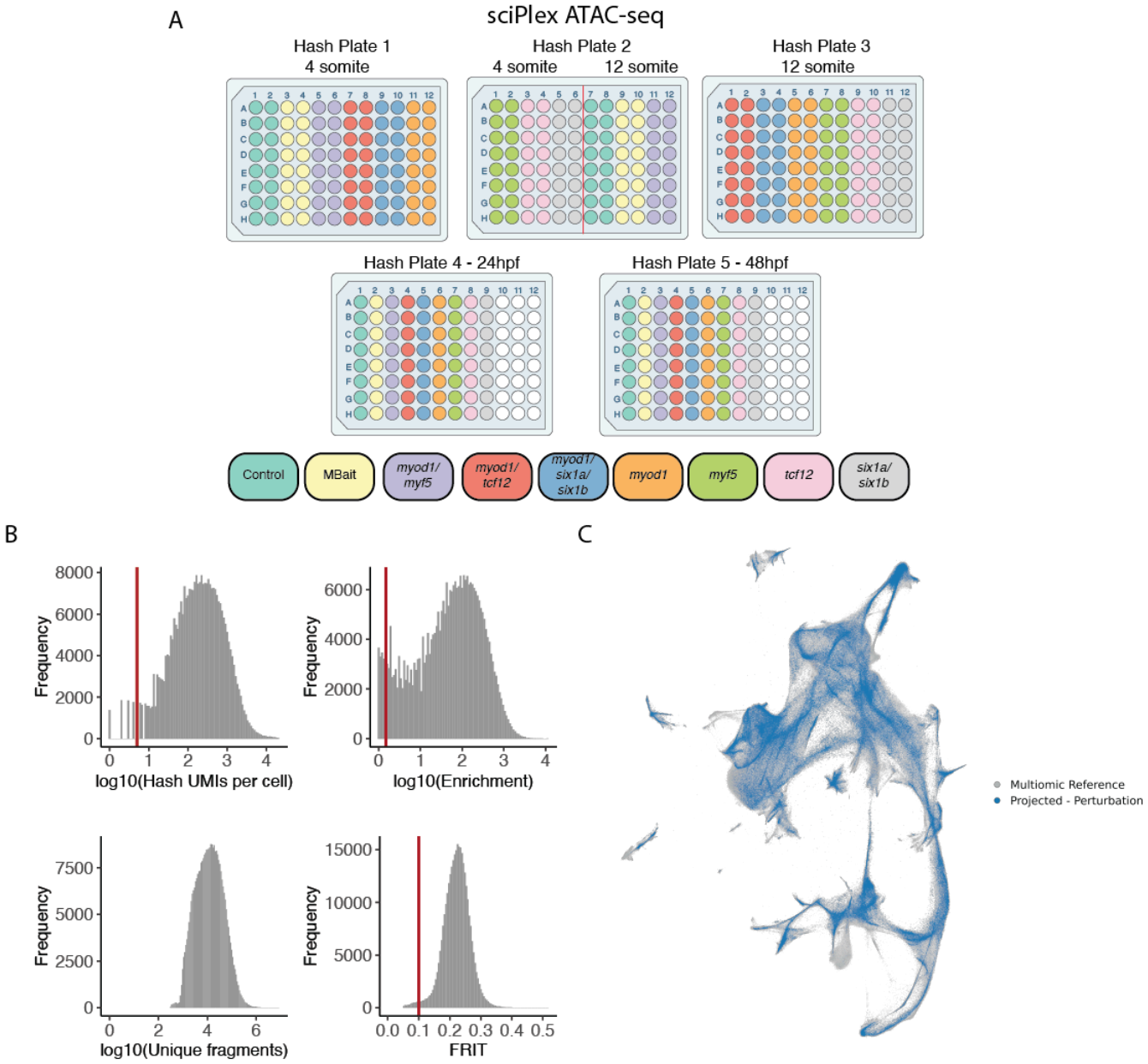


Figure 4.6 sciPlex ATAC-seq hashing enables multiplexed profiling of CRISPR Cas9 injected embryos projected onto an annotated global multiomic reference. A) Experimental design showing embryo distribution across hash plates (colors denote distinct perturbations, wells denote individually indexed embryos, hash plate or red line boundary denote timepoint separation) B) Quality-control distributions: Hash UMIs per cell (upper left); enrichment ratios calculated as UMI count ratio of the most abundance vs. the second most abundant hash oligo; unique fragments measured per cell (lower left); fraction of reads in TSS (lower right) C) Projected sciPlex ATAC-seq data into global multiomic GLUE UMAP embedding

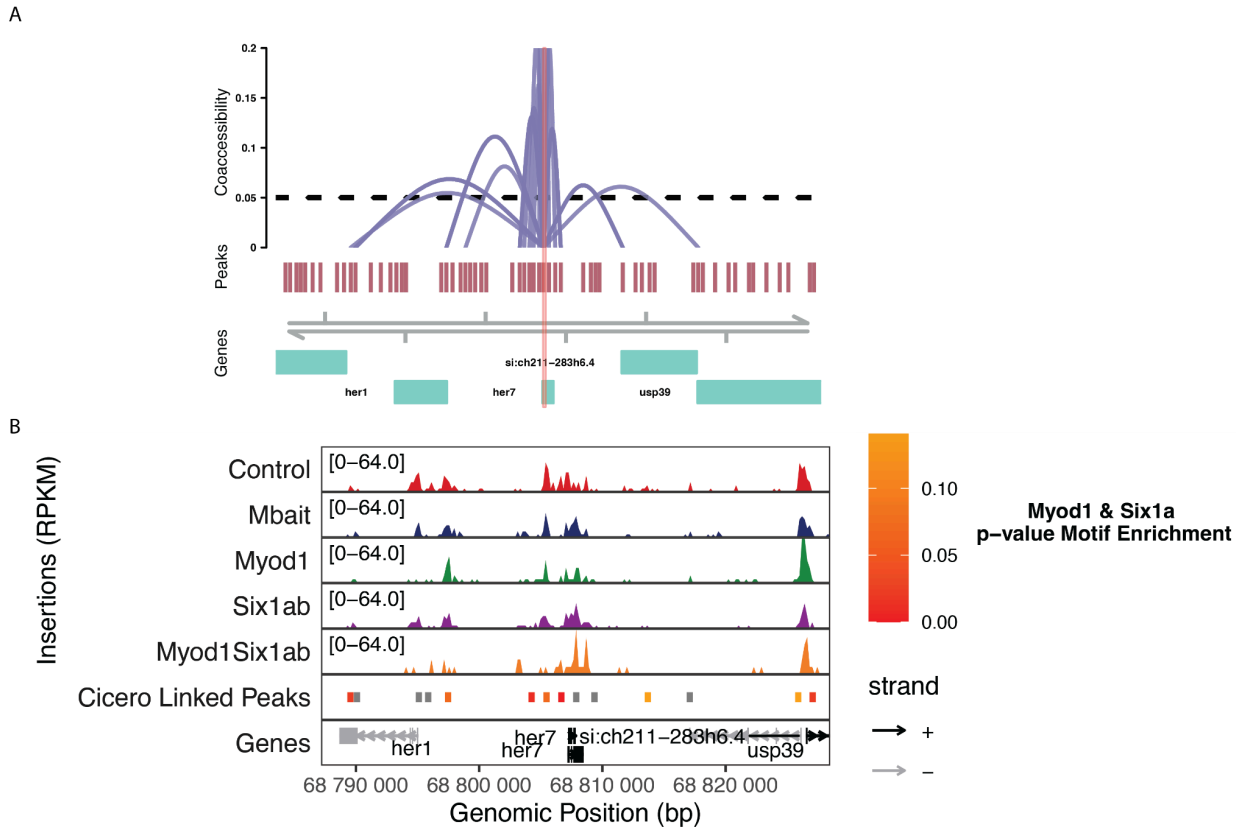


Figure 4.7 Cicero plots and fragment pileup plots of the presomitic mesoderm cells at the *her1* and *her7* locus during genetic perturbation of *myod1*/*six1ab*. A) Shows the coaccessibility of the promoter of *her7* and *cis*-regulatory elements in the surrounding +/- 50kb. One such element was in the intergenic space shared between *her*1 and *her7*. B) Shows fragment pileups under different genetic perturbation. It also shows statistical likelihood of observing *myod1* and *six1a* binding motifs in the peaks linked by Cicero.

4.2.4 Functional validation of cis-regulatory element

To provide additional support that this noncoding *cis*-regulatory element is directly regulated by Myod1 and Six1a/Six1b, we leveraged complementary computational approaches that interrogated both accessibility via DNA sequence and predicted protein-DNA interaction. Using scBasset in silico mutagenesis, we quantified the contribution of each nucleotide within the element to predicted chromatin accessibility by systematically testing all possible

single-nucleotide substitutions and measuring their effects on accessibility across cells (*36*). This analysis identified a 50bp region within the identified peak -10kb upstream of *her7* as having highly important bases in the trunk neuromesodermal progenitors and the PSM. The bases that had the most attributable importance had the Myod1, Six1a, and Rbpj binding motifs. Rbpj interacts with the Notch intracellular domain (NICD) and translocates into the nucleus to drive transcription (Fig. 4.8A). To further assess whether these factors could physically engage the sequence, we used AlphaFoldV3 to generate structural predictions with Myod1, Six1a, and Rbpja together with the DNA sequence of the -10 kb *her7* cis-regulatory element (Fig. 4.8B). The structure appeared coherent and plausible. Additionally, coloring by DNA binding motif (pink) and transcription factor DNA binding protein domain (green) revealed close spatial proximity. Taken together, these analyses provide orthogonal support that this CRE contains sequence-encoded regulatory information consistent with direct control by Myod1 and Six1a/b.

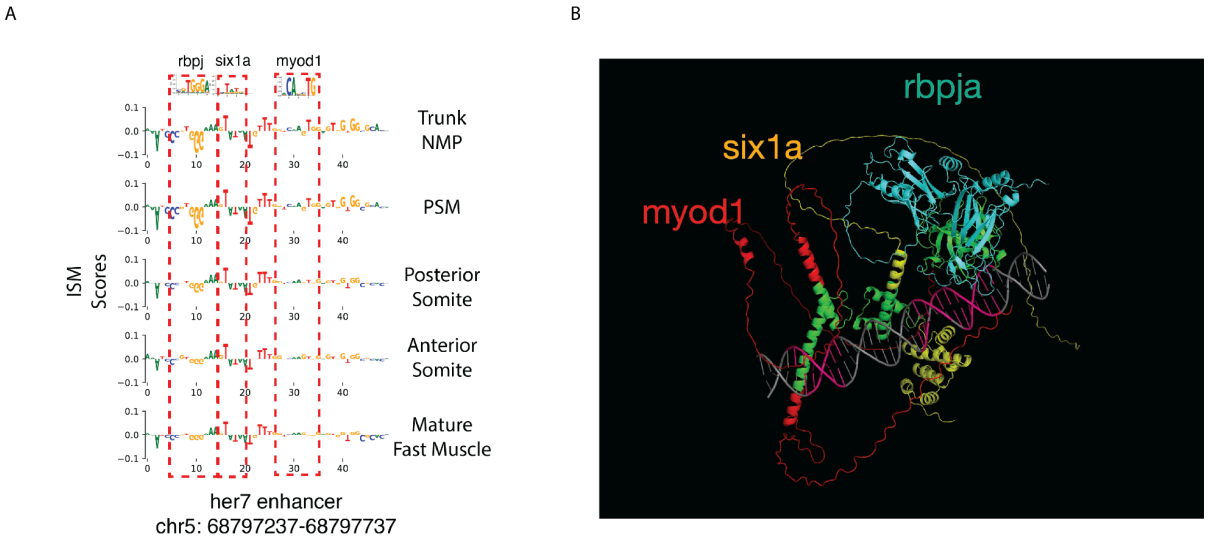


Figure 4.8 Further Computational Support for a Direct Link Between Myod1, Six1a, and the Delta-Notch Oscillator A) Sequence to functional neural network *in silico* mutagenesis & B) AlphaFoldV3 structures suggests protein interaction with specific DNA loci

To test the function of this enhancer *in vivo*, we generated 12 independent founder zebrafish carrying CRISPR-mediated deletions of the element and confirmed germline transmission of the edited alleles by genotyping their offspring. Across founders, germline transmission was generally modest, with edited alleles comprising only ~10–20% of the germline, which substantially limited the number of homozygous mutant embryos recovered in any given incross. This low transmission created a practical challenge for establishing a clear phenotype–genotype relationship by simple counting alone. To address this, I modeled the segregation data using a beta-binomial framework, which allowed me to account for excess variance across clutches and quantify the uncertainty in transmission more appropriately than a standard binomial model. Across four independent incrosses, this analysis revealed a statistically significant enrichment, p=1.95e-5, linking the PCR-defined enhancer deletion genotype with the observed phenotype, providing genetic support that loss of this *cis*-regulatory element contributes to the developmental defect (Fig. 4.9).

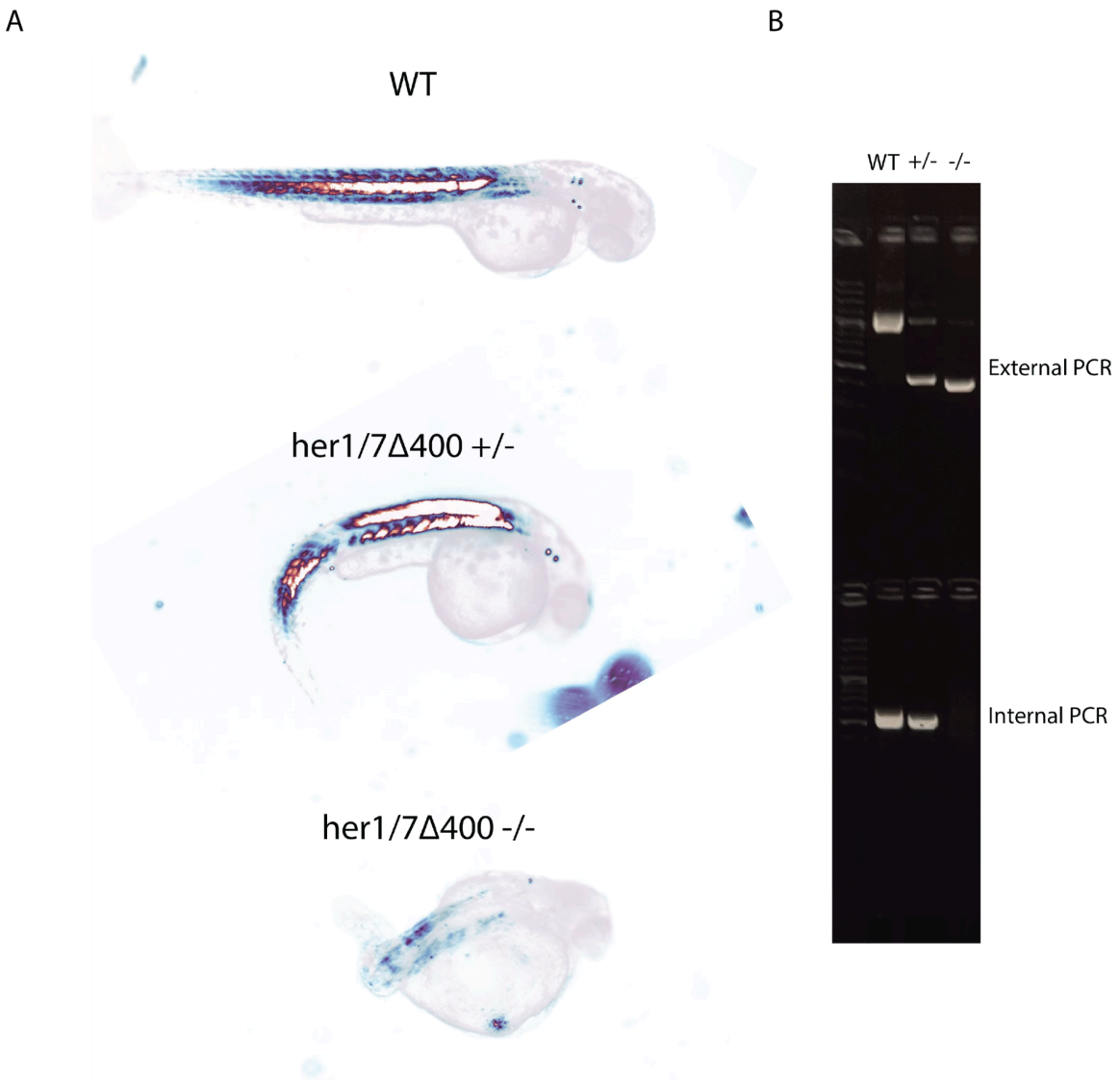


Figure 4.9 F0 Incross of Germline Transmitting Parents Suggest CRE is Necessary for Somite Segmentation via Polarized Light & PCR Genotyping. A) Shows brightfield overlaid with polarized birefringence. Correctly developed mature fast muscle cells B) PCR genotyping by gel electrophoresis of the individual fish shown in panel A

4.3 DISCUSSION

Using sciPlex RNA-seq and sciPlex ATAC-seq we generated libraries of individually indexed RNP F0 injected zebrafish embryos. We also performed reverse-genetic perturbations of single TFs or pairs of TFs suspected of interacting in the paraxial mesoderm cell lineage from predictions generated from sequence to function neural networks. Building cell type abundance graphs, we demonstrate that in *myod1/six1ab* and *myod1/tcf12* double KO embryos, there is a shift in the cellular abundance from the posterior somitic compartment toward the anterior somitic compartment.

We identified thousands of differentially expressed genes in the paraxial mesoderm when using lineage aware comparisons enabled by PLATT. Hundreds of these differentially expressed genes occurred in the distal presomitic mesoderm (PSM) before cell type loss meaningfully occurred, reducing the likelihood of observing and quantifying knockon effects, measured transcription effects due to the loss of the cell type rather than the direct mechanism of perturbation. We confirmed these DEGs using whole mount *in situ* hybridization showing that the Delta-Notch signaling required for proper segmentation during somitogenesis appeared both noisy at the TF level, specifically *her1* and *her7* gene expression, and *dlc*,the ligand to the notch receptor and activator of the pathway, was nearly absent.

In the sciPlex ATAC-seq data, we observed similar cell type abundance losses as sciPlex RNA-seq. Building a Lyon promoter-enhancer contact map, we attributed a noncoding regulatory element -2.8kb upstream of *her1* and -10kb upstream of *her7* as being linked to the promoters of *her7* and *her1*. enriched for *six1a*

4.4 METHODS

scBasset in silico Mutagenesis

To perform the cell type specific *in silico* mutagenesis, we computed the importance scores of all single nucleotides in the sequence of interest. For each position, we ran four forward passes of the model, exploring all single nucleotide changes for the whole sequence. We compared the alternative accessibility prediction to that of the true reference nucleotide to compute the change in accessibility for each cell. We normalized the scores such that they summed to zero. We then took the normalized score at the reference nucleotide as the importance score for that position.

AlphaFoldV3 Predictions

Structural predictions were generated using the online AlphaFoldV3 interface (https://alphafoldserver.com/). As protein input, provided the danRer11 amino acid sequences for the three transcription factors of interest. As nucleic acid input, we provided the forward and reverse danRer11 DNA sequences corresponding to the -10kb *her7 cis*-regualtory element (CRE). Predictions were run using the default AlphaFoldV3 settings and the resulting structural prediction files were downloaded for visualization and analysis.

Projecting scRNA-seq into WT GLUE Embedding

To minimize batch effects during projection into the GLUE embedding, we first transformed the query expression matrix into the reference PCA basis accounting for eigenvalues.

$Z_{projected} = X_{query} V_{ref} \Sigma_{ref}^{-1}$ We then perform forward passes of the GLUE encoder. After generating the latent transformed coordinates for each cell. We ran the monocle3 function load_transform_models() using the reference UMAP embedding and the $Z_{projected}$ values and computed the UMAP coordinates using the reference embedding.

Projecting scATAC-seq into WT GLUE Embedding

To minimize batch effects during projection into the GLUE embedding, we first transformed the fragments file into a peak-by-cell matrix using the overlapping peak set defined from the WT scATAC-seq data. We performed TF-IDF, normalized counts in each cell, log(1 + x*1000), and then followed the same projection procedure used for mapping scRNA-seq data into the WT GLUE embedding.

Whole Mount In situ Hybridization

Alkaline phosphatase ISH was performed using standard conditions (*7*). We used the following riboprobes: *her1*, *her7*, *dlc*, *dld*, and anti-Dig POD (sheep antibody, Roche, catalogue number 11207733910 1:5000).

4.5 CODE AVAILABILITY

Analysis was performed in R and python. Custom scripts can be found on github:

https://github.com/acmullen-med/sciPlexTFx

Analyses of single-cell data were performed using GLUE, Seurat, scBasest, Monocle3, Cicero, Hooke, and PLATT. General tutorials can be found at:

https://scglue.readthedocs.io/en/latest/

https://satijalab.org/seurat/

https://github.com/calico/scBasset

https://cole-trapnell-lab.github.io/monocle3/,

https://cole-trapnell-lab.github.io/cicero-release/

https://cole-trapnell-lab.github.io/hooke/

https://cole-trapnell-lab.github.io/platt/

Chapter 5 - Future Directions and Summary of Findings

5.1 FUTURE DIRECTIONS

I discuss several near term experiments and a few longer-term molecular and computational directions that will make it easier to functionally annotate noncoding DNA and interpret the vast regulatory landscape of the genome. These include incremental experimental improvements and extensions of existing software that aim to produce more reliable, mechanistically grounded annotations of regulatory elements. Although this work was carried out in zebrafish, the methods and principles can be applied to more directly clinically relevant models. Over time, more descriptive functional annotations of the noncoding genome should make genetic information more useful for clinical prognostication and treatment.

5.1.1 Stable Lines

I have nearly completed the generation of several stable zebrafish transgenic and mutant lines needed to confirm phenotypes from our study of paraxial mesoderm noncoding regulatory elements. The immediate next steps are incrossing F1 heterozygote her1/7Δ400 enhancer KO fish and confirming genotype phenotype relationships by PCR and Sanger sequencing. There is a second set of F0 reporter lines that need to be screened for germline transmission of a *her1-Venus* fusion protein to increase the brightness of the live imaging data presented in Fig. 4.8 and Fig. 4.9.

5.1.1.1 her1/7 Enhancer KO

The highest priority line is her1/7Δ400 enhancer KO line (Fig. 5.10). We have generated 15 founder fish across 3 major deletion classes, Δ260, Δ400, Δ780 located -2.8kb upstream of *her1* and -10kb upstream of *her7*. From current germline transmission data, we were able to

estimate a 2-4% frequency of homozygous embryos from F0 incrosses. However, this is comparable to the background rate of nonspecific embryo loss observed during routine zebrafish rearing, complicating phenotype-genotype attribution. Functionally, this is within the “noise floor” on correctly linking phenotype to genotype to demonstrate a correlation.

Although Chapter 4 demonstrated a statistical enrichment for the genotype through a beta-binomial model, obtaining higher and more predictable homozygous frequencies will improve interpretability and allow for whole mount *in situ* hybridization experiments to show reductions in *her1, her7,* and other genes involved in the mechanism of action resulting in segmentation defects. This will be accomplished with fin clips on F1s in the coming weeks, followed by genotype confirmed F1 incrosses to generate 25% homozygous enhancer deletion fish. This will increase our confidence in the observed findings on the necessity of a 500bp section of noncoding DNA -10kb upstream of *her7* and -2.8kb of *her1* to form the anterior/posterior body axis.

5.1.1.2 Safe Harbor Reporter her1:her1-Venus and Excision Reporter

Reporter lines were generated by PhiC31 targeted integrase rather than Tol2 transposase system. PhiC31 restricts insertions to a single neutral predefined landing pad on chromosome 24. In addition to being in a neutral locus, it limits the number of integrations to one within the genome. Random integration can confound interpretation because reporter activity may reflect local chromatin dynamics rather than the regulatory sequence being tested.

Using the PhiC31 integrase, we used *her1*:her1-Venus reporter construct that includes a lens specific fluorescent marker (*cryaa*:EGFP), enabling rapid visual confirmatory screening without conventional PCR genotyping. In parallel, we generated a matched *her1*[Δ800]:her1-Venus reporter with an excision at the candidate noncoding element located

-2.8kb upstream of her1. Comparing these two lines at the same genomic landing site will provide a controlled test of whether this element is necessary for oscillatory presomitic mesoderm reporter activity. While these experiments were performed in F0 embryos, they were confounded by ectopic expression in the yolk syncytial layer, where we suspect ectopic plasmid is able to make copies of her1-Venus. That confounding activity will be minimized in the F1 embryos increasing our confidence in the results suggesting that the CRE is necessary for activity of the reporter.

5.1.1.3 Sufficiency Reporter for -10kb CRE

To test sufficiency, we used PhiC31 to integrate a reporter consisting of a 500bp piece of DNA -10kb upstream of *her7* in front of a minimal promoter and EGFP making a reporter at the landing pad on Chromosome 24. Reporter signal was dim in F0 embryos, and establishing stable germline lines may improve signal to noise by reducing noise and enabling consistent quantification across embryos. If fluorescence remains weak, we could evaluate alternative promoter architectures, including replacing the minimal promoter with *hsp70l*:EGFP, which may increase reporter brightness even in the absence of heat-shock.

5.1.2 Improved Perturbation Strategies for CREs

Generating stable mutant lines for each differentially accessible enhancer is not scalable. The number of putative functional CREs is in the hundreds of thousands, even with modern genome editing technologies the time required to establish and validate stable lines makes functional evaluation experimentally impractical. Compounding this, perturbation of individual CREs has historically produced subtle phenotypes or modest transcriptional changes, reflecting redundancy in the gene regulatory architecture. Given that clear phenotypes in our TF experiments often required combinatorial perturbation, it is likely that robust functional

dissection of noncoding regulation will similarly require simultaneous perturbation of multiple CREs. This motivates the need to develop a screening strategy that is not prohibitively expensive in time and effort.

5.1.2.1 CRISPRi

CRISPR interference (CRISPRi) uses a catalytically inactive Cas9 (dCas9) fused to histone modifying transcriptional repressor domain (KRAB) that silences regulatory activity without modifying DNA sequence. In the context of perturbing CREs, relative to conventional Cas9 cut, CRISPRi offers practical advantages. CRISPRi can be deployed rapidly without hundreds of hours of work to generate a stable deletion line and it supports multiplex targeting of enhancer motifs in the same embryo. Recently, it was demonstrated that injection of mRNA encoding CRISPRi and gRNAs targeting the promoter of neural crest genes into a single-cell staged zebrafish embryo was able to disrupt the neural crest lineage, specifically melanophores until 72hpf (*82*).

However, there are a few challenges and downsides to using CRISPRi as it stands today. You can not order CRISPRi the same way you can order active CRISPR Cas9 from molecular biology vendors, thus requiring you to either make recombinant protein yourself or inject mRNA into the embryo. If you rely on translation of mRNA it increases the likelihood of mosaicism across the embryo and replicates, especially problematic when performing loss of function tests. Codon optimization of the protein sequence to aid translation may be helpful to speed *in vivo* translation. There is a goldilocks period where the embryo has translated the mRNA, but not divided enough to become mosaic. Additionally, it is not known under what contexts the CRISPRi loses its effectiveness of silencing in the context of normal cellular development in embryogenesis. A major wave of histone reprogramming happens around the maternal to zygotic

transition at around 3hpf. Transcription factors may overcome the histone silencing performed at specific loci. That behavior could vary based on the cell type, stage, and specific silencing modifications and will likely need to be empirically defined in each cellular context and stage. An alternative strategy would be to constitutively express CRISPRi as well as gRNAs, but it has historically proved challenging to retain gRNA expression throughout embryonic development.

5.1.2.2 CRISPRoff

CRISPRoff extends the CRISPRi concept by including two fused methyltransferase DNMT3 enzymes onto the inactivated dCas9. These are able to methylate cytosines at a locus causing more durable silencing. In principle, this enables heritable and longer lasting repression across cell divisions and developmental time. The DNMT1 naturally occurring in cells allows CpG methylation from a hemi-methylated strand of DNA.

Because of these seeming benefits, we performed pilot experiments injecting mRNA of CRISPRoff with gRNAs. However, we found that the BFP, and our reconfigured GFP versions, failed to show consistent and robust fluorescence when we injected mRNA into single-cell zebrafish embryos. Additionally, in the absence of injection with gRNAs, there were morphological defects in the embryos. This could be due to the large amount of mRNA being injected into the embryos, as the CRISPRoff fusion protein was 13kb long, or it could be due to partial translation occurring making excess methyltransferase proteins wreaking havoc in the epigenome. Future research directions could optimize expression of fluorescence, codon optimize CRISPRoff for zebrafish embryogenesis, and perform western blots to confirm translation of dCas9 and DNMT3.

5.1.3 Integration of Evolutionary Information Into CRE Annotation

An underexplored axis in single-cell analysis is using evolutionary constraint as a way of prioritizing candidate CREs for examination. UCSC provides a fish specific five species conservation track computed with phastCons for danRer10, the previous assembly genome, called “fishcons”. This track is generated from a hidden markov model that provides a score from 0 to 1 for the degree of conservation among five related fish species. Having a danRer11, the most updated zebrafish assembly, fishcons track would be helpful for triaging opening/closing noncoding elements identified from our scATAC-seq. This assumes that noncoding sequences preserved across evolutionary time are enriched for regulatory function.

5.1.4 Software For Long Range Interactions

Lyon, the successor software tool to Cicero currently under development, performs well at predicting promoter-enhancer pairs but systematically penalizes distal pairs, even when long range regulation is biologically plausible. This bias is problematic because the most canonical developmental enhancers act over very long distances, for example, the ZRS limb enhancer regulates SHH from ~1Mb away. Incorporating a structured 3D genome prior, such as Hi-C contact frequency or TAD boundaries, could provide more biologically grounded restrictions on relevant promoter-enhancer pairing and improve recovery of true distal interactions.

In addition, Lyon does not currently incorporate our consensus index score / hash replicate reproducibility score when fitting promoter-enhancer pairs. Including it as a covariate in the network model fit could be a sensible pursuit. Relatedly, Lyon does not penalize promoters of other genes, causing variability to be explained by promoter-promoter contacts rather than the motivated promoter-enhancer contacts. Applying a simple mask of adjacent promoters may go a long way in improving the regression and interpretability of the results. While

promoter-promoter links may suggest coexpression, that signal is more easily captured by scRNA-seq. Overall, adding biologically grounded priors and explicit penalties where sensible could sharpen enhancer-promoter connectivity and improve interpretability of CRE maps.

5.1.5. Emerging Sequencing Technologies

As we interrogate the noncoding genome at higher level resolution, long read sequencing will become increasingly important. In our own work, we have repeatedly encountered noncoding regions of our laboratory AB strain that have not aligned perfectly to the UCSC danRer11 reference genome. This is not entirely surprising as noncoding DNA tends to harbor more natural variation and zebrafish laboratory strains can drift from reference assemblies over time. Although a telomere-to-telomere zebrafish assembly was recently released as a preprint, it would benefit our own work to generate our own assembly for two reasons. The first is that when we design guides in CHOP-CHOP we want to confirm that the PAM recognition site by Cas9 is correct. The second is if we want to ascribe importance to individual bases, as we did in Fig. 4.8, having the ground truth genomic sequence for our model to fit and predict is critically important.

5.2 SUMMARY OF FINDINGS

We optimized and deployed the assay, sciPlex ATAC-seq to profiling individually indexed zebrafish embryos from gastrulation through organogenesis. We profiled hundreds of embryos across ten timepoints and perturbations. Using the replicate IDs generated from hashing, we improved the strategy for peak calling to increase the number of peaks called while controlling peaks originating from noise. We aligned all of this data into a global multiomic GLUE (scRNA-seq/scATAC-seq) embedding to aid in its interpretation. Through this we were able to attach over 200 high-resolution cell type labels to scATAC-seq. This constitutes the largest to date single-cell dataset of zebrafish scRNA-seq and scATAC-seq.

Using this data we trained sequence to function neural networks to predict TF-TF synergistic interactions. We identified thousands of possible interactions that are cell type and lineage restricted. We selected a small set of interactions to validate experimentally and observed that knocking out pairs of predicted TFs produced a more dramatic phenotype than either alone. Specifically, we observed that *myod1/six1ab*, *myod1/tcf12*, and *myod1/myf5* produced severe alterations to the formation of mature fast muscle derived from somites.

To learn a mechanism that could be mediating that loss of muscle, we performed sciPlex RNA-seq and sciPlex ATAC-seq on single and double perturbed embryos. Using a state-transition lineage tree created in the software tool PLATT, we localized the differentiation blockade to a cell fate decision in presomitic mesoderm (PSM), specifically at the step of choosing posterior or anterior somitic states. Evaluating differentially expressed genes in the presomitic mesoderm, the cell-population before the failed cell fate commitment, we observed a change in *her1, her7,* and *dlc*, as well as other components in the Delta-Notch signaling cascade. We confirmed these changes with whole mount *in situ* hybridization of zebrafish embryos at 4-somite and 12-somite stage.

We fit a Lyon model, a successor to the software tool Cicero, on the sciPlex ATAC-seq data of perturbed zebrafish embryos to the loci of differentially expressed genes to make promoter-enhancer coaccessibility linkages. We identified that a CRE element -10kb upstream of *her7* and -2.8kb upstream of *her1* closed in response to knockout of *myod1/six1ab*, but not in either knockout alone. We validated the sufficiency and necessity of this noncoding element via reporter assays and knocking out the endogenous locus.

This demonstrates the ability of using multiplexed single-cell methods to build whole embryo atlas, learn novel biology with advanced computational tools, and perform mechanistic

experiments to learn a new element that is necessary in mediating Delta-Notch signaling in zebrafish somitogenesis.

BILBIOGRAPHY

1. R. J. F. Loos, 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).

2. D. Thanos, T. Maniatis, Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

3. M. Halpern, R. Ho, C. Walker, C. Kimmel, Induction of muscle pioneers and floor plate is distinguished by the zebrafish no tail mutation. *Cell* **75**, 99–111 (1993).

4. Y. Hinits, D. P. S. Osborn, S. M. Hughes, Differential requirements for myogenic regulatory factors distinguish medial and lateral somitic, cranial and fin muscle fibre populations. *Development* **136**, 403–414 (2009).

5. G. E. Hollway, R. J. Bryson-Richardson, S. Berger, N. J. Cole, T. E. Hall, P. D. Currie, Whole-somite rotation generates muscle progenitor cell compartments in the developing zebrafish embryo. *Dev. Cell* **12**, 207–219 (2007).

6. D. Baranasic, M. Hörtenhuber, P. J. Balwierz, T. Zehnder, A. K. Mukarram, C. Nepal, C. Várnai, Y. Hadzhiev, A. Jimenez-Gonzalez, N. Li, J. Wragg, F. M. D'Orazio, D. Relic, M. Pachkov, N. Díaz, B. Hernández-Rodríguez, Z. Chen, M. Stoiber, M. Dong, I. Stevens, S. E. Ross, A. Eagle, R. Martin, O. Obasaju, S. Rastegar, A. C. McGarvey, W. Kopp, E. Chambers, D. Wang, H. R. Kim, R. D. Acemel, S. Naranjo, M. Łapiński, V. Chong, S. Mathavan, B. Peers, T. Sauka-Spengler, M. Vingron, P. Carninci, U. Ohler, S. A. Lacadie, S. M. Burgess, C. Winata, F. van Eeden, J. M. Vaquerizas, J. L. Gómez-Skarmeta, D. Onichtchouk, B. J. Brown, O. Bogdanovic, E. van Nimwegen, M. Westerfield, F. C. Wardle, C. O. Daub, B. Lenhard, F. Müller, Multiomic atlas with functional stratification and developmental dynamics of zebrafish cis-regulatory elements. *Nat. Genet.* **54**, 1037–1050 (2022).

7. L. M. Saunders, S. R. Srivatsan, M. Duran, M. W. Dorrity, B. Ewing, T. H. Linbo, J. Shendure, D. W. Raible, C. B. Moens, D. Kimelman, C. Trapnell, Embryo-scale reverse genetics at single-cell resolution. *Nature*, doi: 10.1038/s41586-023-06720-2 (2023).

8. D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, A. J. Shendure, Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 906–910 (2015).

9. B. K. Martin, C. Qiu, E. Nichols, M. Phung, R. Green-Gladden, S. Srivatsan, R. Blecher-Gonen, B. J. Beliveau, C. Trapnell, J. Cao, J. Shendure, Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nat. Protoc.*, doi: 10.1038/s41596-022-00752-0 (2022).

10. J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, J. Shendure, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

11. J. Cao, J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers, A. Adey, R. H. Waterston, C. Trapnell, J. Shendure, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

12. J. Cao, D. R. O'Day, H. A. Pliner, P. D. Kingsley, M. Deng, R. M. Daza, M. A. Zager, K. A. Aldinger, R. Blecher-Gonen, F. Zhang, M. Spielmann, J. Palis, D. Doherty, F. J. Steemers, I. A. Glass, C. Trapnell, J. Shendure, A human cell atlas of fetal gene expression. *Science* **370** (2020).

13. J. Shendure, G. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. Rosenbaum, M. D. Wang, K. Zhang, R. Mitra, G. M. Church, Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).

14. K. D. Barbee, X. Huang, Magnetic assembly of high-density DNA arrays for genomic analyses. *Anal. Chem.* **80**, 2149–2154 (2008).

15. F. V. De Rop, G. Hulselmans, C. Flerin, P. Soler-Vila, A. Rafels, V. Christiaens, C. B. González-Blas, D. Marchese, G. Caratù, S. Poovathingal, O. Rozenblatt-Rosen, M. Slyper, W. Luo, C. Muus, F. Duarte, R. Shrestha, S. T. Bagdatli, M. R. Corces, L. Mamanova, A. Knights, K. B. Meyer, R. Mulqueen, A. Taherinasab, P. Maschmeyer, J. Pezoldt, C. L. G. Lambert, M. Iglesias, S. R. Najle, Z. Y. Dossani, L. G. Martelotto, Z. Burkett, R. Lebofsky, J. I. Martin-Subero, S. Pillai, A. Sebé-Pedrós, B. Deplancke, S. A. Teichmann, L. S. Ludwig, T. P. Braun, A. C. Adey, W. J. Greenleaf, J. D. Buenrostro, A. Regev, S. Aerts, H. Heyn, Systematic benchmarking of single-cell ATAC-sequencing protocols. *Nat. Biotechnol.* **42**, 916–926 (2024).

16. J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, L. T. Nguyen, J. Y. H. Kwon, B. Barak, W. Ge, A. J. Kedaigle, S. Carroll, S. Li, N. Hacohen, O. Rozenblatt-Rosen, A. K. Shalek, A.-C. Villani, A. Regev, J. Z. Levin, Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).

17. C. Trapnell, Revealing gene function with statistical inference at single-cell resolution. *Nat. Rev. Genet.* **25**, 623–638 (2024).

18. G. T. Booth, R. M. Daza, S. R. Srivatsan, J. L. McFaline-Figueroa, R. G. Gladden, A. C. Mullen, S. N. Furlan, J. Shendure, C. Trapnell, High-capacity sample multiplexing for single cell chromatin accessibility profiling. *BMC Genomics* **24**, 737 (2023).

19. S. R. Srivatsan, J. L. McFaline-Figueroa, V. Ramani, L. Saunders, J. Cao, J. Packer, H. A. Pliner, D. L. Jackson, R. M. Daza, L. Christiansen, F. Zhang, F. Steemers, J. Shendure, C. Trapnell, Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020).

20. E. Barkan, M. Duran, N. Lammers, A. Tresenrider, D. Jackson, H. Lee, B. Haagen, L. Saunders, P. Abitua, D. Kimelman, C. Trapnell, Embryo-scale single-cell chemical

transcriptomics reveals dependencies between cell types and signaling pathways, *bioRxivorg* (2025). https://doi.org/10.1101/2025.04.03.646423.

21. M. Duran, E. Barkan, A. Tresenrider, H. Lee, R. Z. Friedman, N. Lammers, M. Colón, J. Franks, B. Ewing, D. Kimelman, C. Trapnell, A statistical framework for inferring genetic requirements from embryo-scale single-cell sequencing experiments, *bioRxivorg* (2025). https://doi.org/10.1101/2025.04.03.646654.

22. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

23. S. Domcke, A. J. Hill, R. M. Daza, J. Cao, D. R. O'Day, H. A. Pliner, K. A. Aldinger, D. Pokholok, F. Zhang, J. H. Milbank, M. A. Zager, I. A. Glass, F. J. Steemers, D. Doherty, C. Trapnell, D. A. Cusanovich, J. Shendure, A human cell atlas of fetal chromatin accessibility. *Science* **370** (2020).

24. J. Cao, D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, F. J. Steemers, A. C. Adey, C. Trapnell, J. Shendure, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).

25. S. Ma, B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, T. Law, C. Lareau, Y.-C. Hsu, A. Regev, J. D. Buenrostro, Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020).

26. A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, A. Regev, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).

27. P. Datlinger, A. F. Rendeiro, C. Schmidl, T. Krausgruber, P. Traxler, J. Klughammer, L. C. Schuster, A. Kuchler, D. Alpar, C. Bock, Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).

28. A. J. Hill, J. L. McFaline-Figueroa, L. M. Starita, M. J. Gasperini, K. A. Matreyek, J. Packer, D. Jackson, J. Shendure, C. Trapnell, On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).

29. R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S.-I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, J. Shendure, Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

30. A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, G. E.

Crawford, High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

31. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

32. Z.-J. Cao, G. Gao, Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).

33. K. Sokolova, K. M. Chen, Y. Hao, J. Zhou, O. G. Troyanskaya, Deep learning sequence models for transcriptional regulation. *Annu. Rev. Genomics Hum. Genet.* **25**, 105–122 (2024).

34. J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

35. D. R. Kelley, J. Snoek, J. L. Rinn, Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).

36. H. Yuan, D. R. Kelley, scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods*, doi: 10.1038/s41592-022-01562-8 (2022).

37. Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, D. R. Kelley, Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

38. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

39. Q. Li, J. B. Brown, H. Huang, P. J. Bickel, Measuring Reproducibility of High-Throughput Experiments. *The Annals of Applied Statistics* **5**, 1752–1779 (2011).

40. J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, W. J. Greenleaf, ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

41. B. Parks, W. Greenleaf, Scalable high-performance single cell data analysis with BPCells, *bioRxivorg* (2025). https://doi.org/10.1101/2025.03.27.645853.

42. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv [stat.ML]* (2018). http://arxiv.org/abs/1802.03426.

43. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

44. O. Bogdanovic, A. Fernandez-Miñán, J. J. Tena, E. de la Calle-Mustienes, C. Hidalgo, I.

van Kruysbergen, S. J. van Heeringen, G. J. C. Veenstra, J. L. Gómez-Skarmeta, Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22**, 2043–2053 (2012).

45. D. N. Conrad, K. T. Phong, E. Korotkevich, C. S. McGinnis, Q. Zhu, E. D. Chow, Z. J. Gartner, Reducing batch effects in single cell chromatin accessibility measurements by pooled transposition with MULTI-ATAC, *bioRxivorg* (2025). https://doi.org/10.1101/2025.02.14.638353.

46. S. Fu, S. Wang, D. Si, G. Li, Y. Gao, Q. Liu, Benchmarking single-cell multi-modal data integrations. *Nat. Methods*, doi: 10.1038/s41592-025-02737-9 (2025).

47. X. Nguyen, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).

48. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

49. X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, C. Trapnell, Reversed Graph Embedding Resolves Complex Single-Cell Trajectories. *Nat. Methods* **14**, 979 (2017).

50. Z. Gu, M. L. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, W.-H. Li, Role of Duplicate Genes in Genetic Robustness Against Null Mutations. *Nature* **421**, 60–63 (2003).

51. M. A. El-Brolosy, D. Y. R. Stainier, Genetic compensation: A phenomenon in search of mechanisms. *PLoS Genet.* **13**, e1006780 (2017).

52. A. Rossi, Z. Kontarakis, C. Gerri, H. Nolte, S. Hölper, M. Krüger, D. Y. R. Stainier, Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature* **524**, 230–233 (2015).

53. J. C. Opazo, G. T. Butts, M. F. Nery, J. F. Storz, F. G. Hoffmann, Whole-genome duplication and the functional diversification of teleost fish hemoglobins. *Mol. Biol. Evol.* **30**, 140–153 (2013).

54. M. Tasnim, P. Wahlquist, J. T. Hill, Zebrafish: unraveling genetic complexity through duplicated genes. *Dev. Genes Evol.* **234**, 99–116 (2024).

55. P. C. Phillips, Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867 (2008).

56. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

57. F. Spitz, E. E. M. Furlong, Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

58. J. Molkentin, B. Black, J. F. Martin, E. Olson, Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins. *Cell* **83**, 1125–1136 (1995).

59. J. Massagué, J. Seoane, D. Wotton, Smad transcription factors. *Genes Dev.* **19**, 2783–2810 (2005).

60. A. Reményi, K. Lins, L. J. Nissen, R. Reinbold, H. R. Schöler, M. Wilmanns, Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* **17**, 2048–2059 (2003).

61. D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, R. A. Young, Control of Pancreas and Liver Gene Expression by HNF Transcritpion Factors. *Science* **303**, 1374–1378 (2004).

62. E. Wingender, T. Schoeps, M. Haubrock, M. Krull, J. Dönitz, TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* **46**, D343–D347 (2018).

63. M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker, T. R. Hughes, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

64. I. V. Kulakovskiy, Y. A. Medvedeva, U. Schaefer, A. S. Kasianov, I. E. Vorontsov, V. B. Bajic, V. J. Makeev, HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* **41**, D195–202 (2013).

65. Sunjay Kaushal, Jay W. Schneider, Bernardo Nadal-Ginard, Vijak Mahdavi, Activation of the Myogenic Lineage by MEF2A, a Factor That Induces and Cooperates with MyoD. [Preprint] (1994).

66. Y. Xu, J. He, H. L. Tian, C. H. Chan, J. Liao, T. Yan, T. J. Lam, Z. Gong, Fast skeletal muscle-specific expression of a zebrafish myosin light chain 2 gene and characterization of its promoter by direct injection into skeletal muscle. *DNA Cell Biol.* **18**, 85–95 (1999).

67. M. T. Lee, A. R. Bonneau, C. M. Takacs, A. A. Bazzini, K. R. DiVito, E. S. Fleming, A. J. Giraldez, Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature* **503**, 360–364 (2013).

68. L. Fishman, A. Modak, G. Nechooshtan, T. Razin, F. Erhard, A. Regev, J. A. Farrell, M. Rabani, Cell-type-specific mRNA transcription and degradation kinetics in zebrafish

embryogenesis from metabolically labeled single-cell RNA-seq. *Nat. Commun.* **15**, 3104 (2024).

69. J. C. Talbot, E. M. Teets, D. Ratnayake, P. Q. Duy, P. D. Currie, S. L. Amacher, Muscle precursor cell movements in zebrafish are dynamic and require Six family genes. *Development* **146** (2019).

70. M. A. Rudnicki, P. N. Schnegelsberg, R. H. Stead, T. Braun, H. H. Arnold, R. Jaenisch, MyoD or Myf-5 is required for the formation of skeletal muscle. *Cell* **75**, 1351–1359 (1993).

71. Y. Hinits, V. C. Williams, D. Sweetman, T. M. Donn, T. P. Ma, C. B. Moens, S. M. Hughes, Defective cranial skeletal development, larval lethality and haploinsufficiency in Myod mutant zebrafish. *Dev. Biol.* **358**, 102–112 (2011).

72. K. Labun, T. G. Montague, M. Krause, Y. N. Torres Cleuren, H. Tjeldnes, E. Valen, CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).

73. E. K. Brinkman, T. Chen, M. Amendola, B. van Steensel, Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014).

74. L. Maves, A. J. Waskiewicz, B. Paul, Y. Cao, A. Tyler, C. B. Moens, S. J. Tapscott, Pbx homeodomain proteins direct Myod activity to promote fast-muscle differentiation. *Development* **134**, 3371–3382 (2007).

75. M. V. Kuleshov, J. E. L. Diaz, Z. N. Flamholz, A. B. Keenan, A. Lachmann, M. L. Wojciechowicz, R. L. Cagan, A. Ma'ayan, modEnrichr: a suite of gene set enrichment analysis tools for model organisms. *Nucleic Acids Res.* **47**, W183–W190 (2019).

76. N. Zhang, T. Gridley, Defects in somite formation in lunatic fringe-deficient mice. *Nature* **394**, 374–377 (1998).

77. L.-T. Yang, J. T. Nichols, C. Yao, J. O. Manilay, E. A. Robey, G. Weinmaster, Fringe glycosyltransferases differentially modulate Notch1 proteolysis induced by Delta1 and Jagged1. *Mol. Biol. Cell* **16**, 927–942 (2005).

78. R. Kopan, M. X. G. Ilagan, The canonical Notch signaling pathway: unfolding the activation mechanism. *Cell* **137**, 216–233 (2009).

79. C. Hicks, S. H. Johnston, G. diSibio, A. Collazo, T. F. Vogt, G. Weinmaster, Fringe differentially modulates Jagged1 and Delta1 signalling through Notch1 and Notch2. *Nat. Cell Biol.* **2**, 515–520 (2000).

80. J. K. Dale, M. Maroto, M.-L. Dequeant, P. Malapert, M. McGrew & O. Pourquie, Periodic Notch inhibition by Lunatic Fringe underlies the chick segmentation clock. *Nature* **421**, 272–275 (2003).

81. Y. Zhao, D. Granas, G. D. Stormo, Inferring binding energies from selected binding sites. *PLoS Comput. Biol.* **5**, e1000590 (2009).

82. N. B. Barrientos, E. A. Shoppell, R. J. Boyd, V. C. Culotta, A. S. McCallion, Optimized CRISPR inhibition and activation opens key avenues for systematic biological exploration in zebrafish, *bioRxivorg* (2024). https://doi.org/10.1101/2024.09.16.613289.